L. Hernández Acosta. A. Dehning. W. Gero Oed and D. Reinhardt. No Ears but Aware: Fingerprinting Attacks on Smart Speakers. Proceedings of the 22nd International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Work-in-Progress), 2024.

No Ears but Aware: Fingerprinting Attacks on Smart Speakers

Luca Hernández Acosta*, Adrian Dehning†, Winfried Gero Oed‡, Delphine Reinhardt§
Computer Security and Privacy (CSP), University of Göttingen, Göttingen, Germany
*hernandez@cs.uni-goettingen.de, †a.dehning99@web.de, ‡winnus@posteo.de, §reinhardt@cs.uni-goettingen.de

***Abstract*—Smart speakers are gaining popularity in smart homes, enhancing convenience in managing other devices and providing entertainment. Despite these benefits, they introduce privacy concerns that need to be addressed. In this study, we examine the susceptibility of smart speakers to fingerprinting attacks, employing Jaccard index and *Deep Neural Networks* (DNNs), including *Convolutional Neural Networks* (CNNs), *Long Short-Term Memory networks* (LSTMs), and *Stacked Autoencoders* (SAEs). We've implemented a new tool for collecting traffic traces, notably for multi-interaction skills. Jaccard index achieves a minimum of 97% accuracy in predicting the category of a voice command, while the CNN model reaches at least 90%. Both methods struggle with specific command identification that goes beyond the category. Future work will focus on expanding our dataset, as the DNN approach might significantly benefit from more extensive data for improved training. Together, these approaches underscore the profound privacy threats posed by fingerprinting attacks on smart speakers and highlight the urgency for enhanced security measures to safeguard user privacy in the expanding domain of IoT devices.**

***Index Terms*—Smart Speaker, Privacy, Voice Command Fingerprinting, Security, IoT**

I. Introduction

In the era of IoT, smart home speaker systems like Amazon Echo and Google Home are ubiquitous. These devices are activated by a wake word and can perform a variety of functions based on user's voice commands. These commands are then processed on remote cloud servers, maintained by the manufacturers. While this technology offers considerable benefits in terms of convenience, it also introduces significant challenges concerning user privacy and security [1]. One of them is the voice command fingerprinting attack shown in Fig. 1 and adopted in this paper. In this attack, attackers can infer specific users' voice commands by analyzing the characteristics of the encrypted network traffic without decrypting it between the smart speaker and the cloud server [2–4]. The practicality of this attack stems from its simplicity – sniffing network traffic is easier and less conspicuous than attempting decryption. While existing studies have explored voice command fingerprinting attacks, there remains a need to delve into more complex interactions like Alexa skills and associated smart home interactions. Additionally, there is a gap in comparing the efficacy of proven approaches like Jaccard index and DNNs in this context.

In our study, while we still collect and analyze standard voice commands, we place a particular emphasis on Amazon Alexa skills. This dual approach allows us to compare the more complex skill interactions with regular commands. Since no dataset exists, we have hence developed an automated tool for the collection of voice command traffic, including a variety of Alexa skills interactions. This tool systematically collects traffic traces, to analyze and fingerprint different voice commands. The collected dataset comprises traffic traces of voice interactions, including simple commands, multi-interaction skills, smart home devices (lights and sockets), Spotify usage, and alarms.

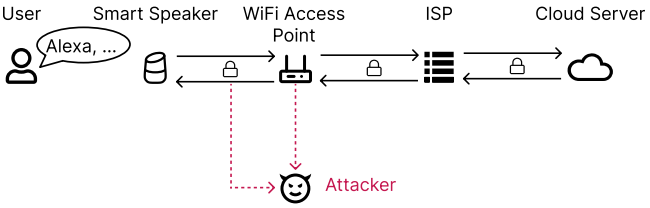

Fig. 1. Threat model

Based on this tool and the collected dataset, we have applied two different approaches. The first is based on a statistical analysis based on the Jaccard index, while the second is based on deep learning techniques, i.e., *Convolutional Neural Networks* (CNNs), *Long Short-Term Memory* (LSTM) networks, and *Stacked Autoencoders* (SAEs). Our choice for the former approach is motivated by its use in [5], while the latter is also applied in [2]. By doing so, we can hence benchmark our results against related work. Our results indicate that the Jaccard index fingerprinting method is effective in inferring categories, demonstrating accuracies between 97% to 100%. However, when it comes to identifying specific voice commands within these categories, there is a notable decrease in performance. For instance, accuracy drops to as low as 5% for certain smart home commands, whereas it reaches up to 93% for multi-interaction skill commands. A similar trend is observed with deep learning techniques involving CNN, LSTM, and SAE models. Here, the CNN model is the most proficient, achieving an accuracy of 97% in a three-category framework and 90% across six categories. Nonetheless, these models also struggle in accurately predicting specific voice commands, with success rates varying from 6% in the smart home category, specifically for light device commands, to 60% for multi-interaction skills.

The contributions of this paper are as follows:

- Development of a novel tool for capturing traffic traces from Amazon Echo Dot, enhanced for both simple single

interactions and complex, multi-interaction Alexa skills,
- Investigation of the effectiveness of the Jaccard index in categorizing voice commands, ranging from basic commands to multi-interaction skills,
- Analysis of the performance of CNN, LSTM, and SAE models in classifying voice command traffic, addressing both simple and complex command types, and
- Comparison of the the Jaccard index approach with deep learning techniques, assessing their performance in fingerprinting diverse types of voice command traffic.

II. Related Work

Automated tools for collecting voice traffic have enabled large-scale data analysis on devices like Amazon Echo and Google Home [5]. We've refined these tools to not only capture simple voice commands but also to efficiently gather Alexa skill interactions for our voice command fingerprinting analysis. Therefore, our new generated dataset not only includes standard voice commands but also integrates interactions with Alexa skills, thus broadening the analytical scope and providing new insights into voice command fingerprinting. Research in the area of voice command fingerprinting has adapted website fingerprinting methods, such as Jaccard index, for smart speakers like Amazon Echo, in order to identify voice commands [5]. Deep learning techniques, including CNNs, LSTMs, and SAEs, applied to these datasets, have shown high accuracies in inferring voice commands, further underscoring the privacy risks [2, 4]. Additionally, the fingerprinting of voice applications on Amazon Echo has been explored beyond simple voice commands by analyzing encrypted traffic of interactions with Amazon Echo Skills [3]. Our work extends the field of voice command fingerprinting by not only analyzing simple voice commands but also placing a focus on Alexa skills interactions. By incorporating advanced deep learning methodologies, our research progresses beyond the conventional machine learning algorithms like *Random Forest* and *Support Vector Classifiers*, as used in prior works [3], which took a first step in fingerprinting Alexa skills. Moreover, our study offers a unique comparative analysis between two distinct fingerprinting methodologies: the established Jaccard index approach and the sophisticated *Deep Neural Network* (DNN) based techniques. This comparison is instrumental in evaluating the effectiveness and intricacies of various strategies in fingerprinting voice command traffic, especially in the context of Alexa skills.

III. Methodology

To achieve our goal of investigating voice command fingerprinting, our methodology employs a Python-scripted laptop as a wireless access point, interfacing with an Amazon Echo Dot via a headset to gather a diverse dataset of simple voice commands and Alexa skill interactions. Moreover, we employ a dual fingerprinting approach, using the Jaccard index and DNNs, including CNNs, LSTMs, and SAEs, for detailed voice command assessment. More comprehensive details on our data collection and fingerprinting strategies are elaborated in the following subsections.

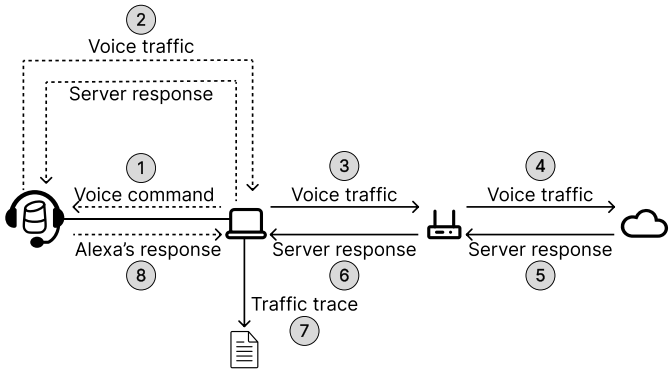

Fig. 2. Data collection

A. Dataset Generation

Since no dataset exists that includes or focuses on Alexa skill interactions, we have first generated a dedicated dataset. We have decided to include the following Alexa skills: "NASA Mars" (answers questions about Mars), "Mixologist" (suggests drink recipes), "Capital Cities Quiz" (multiple-choice quiz on capital cities), "Death Info for Westeros" (confirms deaths of "Game of Thrones" characters), "Magic 8-Ball" (answers yes/no questions), "Stopwatch" (time tracking), "Area Code" (provides area info for codes), "Spin the Wheel" (random selection from user-provided names), "Save Water by Colgate" (tips on water conservation), and "Wine Gal" (wine recommendations for meals). These skills are multi-interaction skills where the skill awaits further user input after a request. Additionally, we collected data on skills for smart home device interactions, like lights and power sockets, and the Spotify skill for music playback. We also included timer/alarm commands and basic simple commands (e.g. "What is the weather for tomorrow?"). Multi-interaction skills, that include more than one request-response pair, are especially challenging. These skills require advanced automation tools for data collection, as they involve sequences of interactions unique to each skill. Selection criteria for these skills included allowing multiple interactions, the feasibility of automation, not triggering real-world events (e.g. phone calls), no need for account linking, and a reasonable interaction duration. The complexity of capturing multi-interaction skills was a significant aspect, as it demanded a sophisticated approach to automate interactions without human input and ensure efficient data collection. This involved selecting multi-interaction skills that did not necessitate continuous, fluent conversation but rather consisted of sequential requests, as exemplified by the "NASA" skill, a skill where Alexa answers user-posed questions about Mars and prompts for more until the skill is stopped. This strategy allowed for more efficient data collection by focusing on skills where the conversation pattern was predictable and could be automated without relying on real-time speech recognition. Fig. 2 illustrates our collection tool setup. Data is collected

using a Python script running on a laptop configured as a wireless access point, capturing all network traffic in .pcap files. The script interfaced with Alexa through a connected headset, facilitating the transmission of audio data. Text-To-Speech technology, particularly the eSpeak speech synthesizer, was employed to generate audio for voice interactions in real-time. The dataset consists of approximately 3,200 traces across all categories, reformatted into a .csv file with each entry featuring a timestamp and the packet length, marked with '-' for incoming and '+' for outgoing traffic.

B. Fingerprinting Approaches

Having established a dataset featuring diverse Alexa Skill interactions and standard voice commands, we now explore the application of two distinct fingerprinting methods: the Jaccard index and DNNs.

1) Jaccard Index Approach: The Jaccard index, a measure of similarity between two sets, is calculated as the size of the intersection divided by the size of the union of the sets. In the context of voice command fingerprinting, consider two traffic traces from different voice commands, each represented as a set of packet sizes. The direction "from" the considered entity is coded as a positive value, while the direction "to" is coded as a negative value. For example, suppose Traffic Trace A is represented as {+150, -300, +450} and Traffic Trace B as {+150, +200, -300}. To compute the Jaccard index, we first find their intersection, i.e.,{+150, -300}, in this case. Next, we determine the union of the sets, which includes all unique elements from both sets: {+150, -300, +450, +200}. The Jaccard index is then calculated as the size of the intersection divided by the size of the union:

$$J(A,B) = \frac{|\{+150, -300\}|}{|\{+150, -300, +450, +200\}|} = \frac{2}{4} = 0.5 \quad (1)$$

This index of 0.5 suggests a moderate level of similarity between the two traffic traces. We hence use this index to compare the similarity of traffic traces in the fingerprinting process.

2) Deep Learning Approach: In this study, we further apply several DNN models for voice command fingerprinting, building upon modified versions of models used in prior research [2]. These included:

- CNNs for capturing spatial dependencies within data,
- LSTM networks, ideal for processing time-series data and capturing temporal dependencies, and
- SAEs for unsupervised learning and feature extraction.

We have tailored each model to the dataset, which featured six distinct classes of voice commands. The two-step training process first predicts the class of a voice command, followed by identifying the specific command within that class. Pre-processing and training of the models involve limiting voice command traces to the first 600 packets and normalizing the data using scikit-learn's min-max scaler. The dataset is then split, allocating 80% for training and 20% for testing, with a further 20% of the training set used for validation purposes. Each category of voice commands is adequately represented in all sets to ensure a thorough evaluation.

TABLE I
COMPARISON OF CATEGORY AND COMMAND PREDICTION ACCURACIES FOR JACCARD INDEX (MULTI: MULTI-INTERACTION SKILLS, SMART: SMART HOME)

	Simple	**Multi**	**Smart**	**Spotify**	**Alarms**
Category	100%	98%	99%	97%	99%
Command	55%	93%	5%	65%	96%

IV. EVALUATION AND RESULTS

A. Jaccard Index

Tab. I summarizes the accuracies achieved by our Jaccard index approach for fingerprinting different voice commands. Using 100 distinct traces per category, we found the Jaccard index reliably identified the correct category with at least 97% accuracy. However, accurately detecting the exact voice command varied widely, with accuracies ranging from 5% for smart home devices to 93% for multi-interaction skills and 96% for alarms. This variation underscores the approach's effectiveness in category detection but highlights challenges in pinpointing specific commands.

Overall, the Jaccard index demonstrated a strong capability to distinguish between categories of skill interactions, with accuracy reaching up to 100% in some instances. However, its performance in identifying specific voice commands was variable, particularly in the smart home category.

B. DNN Approach

We have further conducted a comprehensive evaluation of DNN models including eight tests. Each test is designed to assess the model's ability to accurately identify and classify categories and the specific voice commands in that category, its results can be seen in Tab. II. The tests are structured as follows:

Category Prediction Tests: We initiated our evaluation with a three-category test, involving traces from three categories (alarms, lights, and Spotify), aiming to predict the correct category. This was followed by a more extensive six-category test, which included all six categories: simple commands, multi-interaction skills, smart home devices (including sockets and lights), alarms, and Spotify. The initial three-category test, demonstrates the models' capabilities in broader category classification, with the CNN, LSTM, and SAE models achieving prediction accuracies of 97%, 87%, and 79% respectively. In the more expansive six-category test, the models demonstrated differing levels of accuracy. Specifically, the CNN, LSTM, and SAE models achieved accuracies of 90%, 81%, and 73%, respectively.

Specific Voice Command Evaluation: Subsequent tests delved into each specific category, focusing on the model's precision in predicting the exact voice command within that category. This granular approach allowed us to gauge the model's performance in more detailed scenarios. Our findings revealed varied results across different models and categories. In the voice command-specific tests, the CNN and LSTM models exhibited similar yet distinct performance patterns.

TABLE II
MODEL PREDICTION ACCURACIES ACROSS DIFFERENT CATEGORIES

Category	CNN	LSTM	SAE
Three-Categories	97%	87%	79%
Six-Categories	90%	81%	73%
Alarms (specific command)	67%	68%	60%
Lights (specific command)	6%	4%	4%
Simple (specific command)	11%	7%	3%
Skills (specific command)	60%	63%	38%
Sockets (specific command)	14%	36%	28%
Spotify (specific command)	39%	36%	18%

While the CNN showed a range of accuracies, achieving 67% in alarms, 60% in multi-interaction skills, 39% in Spotify, 14% in sockets, 11% in simple commands, and 6% in lights, the LSTM model's performance varied in different categories, indicating that each model had its strengths in predicting certain types of voice commands. These results illustrate the diverse challenges posed by different categories, with some such as lights and simple commands proving to be more complex for accurate prediction. These results are further displayed in Tab. II

In the comparison between Jaccard index and DNN methods for fingerprinting voice commands, Jaccard index slightly outperforms DNN in category prediction with a 97% accuracy, compared to DNN's 90% using CNN. However, both methods face challenges in accurately predicting specific voice commands. It's important to note that while the Jaccard index's performance is not expected to improve with more data due to its static methodology, the DNN approach has the potential for enhanced accuracy with additional training data.

Overall, our research highlights the intricate nature of voice command fingerprinting, with models showing differing levels of efficacy across categories. The results underscore the need for further optimization and training of these models, particularly in categories with a high diversity of commands or intricate classification requirements.

V. DISCUSSION AND FUTURE WORK

In future work on voice command fingerprinting, several issues should be addressed to enhance our contributions. Firstly, refining the automation of data collection is crucial, especially for complex multi-interaction skills. Advanced speech recognition is key in this context, enabling accurate automation and capture of nuanced conversations inherent to these skills. Secondly, expanding the dataset size is essential for the optimal training of deep learning models. A more extensive dataset, encompassing a broader range of voice commands and interactions, is critical for improving the models' accuracy and generalization capabilities. Finally, developing and rigorously testing effective countermeasures, such as packet padding, is important. This includes evaluating the impact of these countermeasures on network performance and user experience and assessing their effectiveness in mitigating privacy risks associated with voice command fingerprinting. Through these enhancements, future research can provide a more comprehensive and secure framework for voice command fingerprinting.

VI. CONCLUSION

In our ongoing research, we are delving into the nuances of voice command fingerprinting, focusing on a comprehensive dataset that includes traces from e.g. simple voice commands and interactions with Amazon Alexa skills. This investigation encompasses two distinct approaches for executing fingerprinting attacks: Firstly, the Jaccard index method, and secondly, DNN-based approaches (CNN, LSTM, and SAE). Both the Jaccard index method and the DNN-based approaches demonstrated comparable effectiveness in accurately classifying voice commands into categories. However, each method faced challenges in precisely predicting the specific voice command used within a given category. Despite the dataset's current limitations, our study successfully demonstrates the feasibility of conducting voice command fingerprinting attacks with a specialized focus on Alexa skills, although the overall performance can still be improved in the future. We thus aim to validate our hypothesis that enlarging the training dataset will substantially improve the effectiveness of fingerprinting attacks on skill interactions, thus contributing to better understanding of the security as well as privacy challenges and potential vulnerabilities within smart speaker ecosystems. This future work will be crucial in advancing our understanding and capabilities of voice command fingerprinting, particularly in more complex and varied interaction scenarios.

REFERENCES

[1] L. Hernández Acosta and D. Reinhardt, "A Survey on Privacy Issues and Solutions for Voice-Controlled Digital Assistants," *Pervasive and Mobile Computing (PMC)*, 2021.

[2] C. Wang, S. Kennedy, H. Li, K. Hudson, G. Atluri, X. Wei, W. Sun, and B. Wang, "Fingerprinting Encrypted Voice Traffic on Smart Speakers With Deep Learning," in *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2020.

[3] S. Naraparaju, "Fingerprinting Voice Applications on Smart Speakers Over Encrypted Traffic," in *Proc. of the 8th IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2020.

[4] J. Mao, C. Wang, Y. Guo, G. Xu, S. Cao, X. Zhang, and Z. Bi, "A Novel Model for Voice Command Fingerprinting Using Deep Learning," *Journal of Information Security and Applications*, 2022.

[5] S. Kennedy, H. Li, C. Wang, H. Liu, B. Wang, and W. Sun, "I Can Hear Your Alexa: Voice Command Fingerprinting on Smart Home Speakers," in *Proc. of the 7th IEEE Conference on Communications and Network Security (CNS)*, 2019.