University of Basel

A Testbed for Voice Assistant Traffic Fingerprinting

Master Thesis Presentation

Milan van Zanten

University of Basel

21.03.2024

Outline

1. Voice Assistants
2. Traffic Fingerprinting
3. Testbed
4. Results
5. Demo

Ask questions any time!

A Testbed for Voice Assistant Traffic Fingerprinting

Devices

Specifically, *Smart Speakers*

Echo Dot	HomePod Mini	Google Home Mini
Amazon Alexa	*Siri*	*Google Assistant*

Devices

Specifically, *Smart Speakers*

Echo Dot

Amazon Alexa

HomePod Mini

Siri

Google Home Mini

Google Assistant

Smart Speaker Privacy / Security

none (╯°□°）╯︵ ┻━┻

Smart Speaker Privacy / Security

There are concerns…

- Usually located where sensitive conversations take place
- Necessarily always listening
  - Misactivations
- Used to control smart home devices (e.g. door locks)
- No authentication*

* Voice recognition is still insecure.

Smart Speaker Privacy / Security

There are concerns…

- Usually located where sensitive conversations take place
- Necessarily always listening
  - Misactivations
- Used to control smart home devices (e.g. door locks)
- No authentication*

About 40% of households in the U.S. own a smart speaker.

* Voice recognition is still insecure.

Attacks on Smart Speakers

Active:

- Malicious activations
- Similar pronounciations, “skill squatting”
  - (e.g. “Boil an egg” $\rightarrow$ “Boyle an egg”)[1]

Passive:

- Traffic Fingerprinting

[1] D. Kumar et al., *“Skill Squatting Attacks on Amazon Alexa”*, August 2018, Available: https://www.usenix.org/conference/usenixsecurity18/presentation/kumar

Attacks on Smart Speakers

Active:

- Malicious activations
- Similar pronounciations, “skill squatting”
  - (e.g. “Boil an egg” → “Boyle an egg”)[2]

Passive:

- Traffic Fingerprinting

[2] D. Kumar et al., *“Skill Squatting Attacks on Amazon Alexa”*, August 2018, Available: https://www.usenix.org/conference/usenixsecurity18/presentation/kumar

A Testbed for Voice Assistant Traffic Fingerprinting

"[SSL] tra ffic analysis aims to recover confidential information about protection sessions by examining unencrypted packet fields and unprotected packet attributes. For example […] the volume of network traffic flow"

— Wagner and Schneier[3]

[3] D. Wagner and B. Schneier, *"Analysis of the SSL 3.0 Protocol"*, November 1996, Available: https://dl.acm.org/doi/10.5555/1267167.1267171

"[SSL] tra ffic analysis aims to recover confidential information about protection sessions by examining unencrypted packet fields and unprotected packet attributes. For example [...] the volume of network traffic flow"

— Wagner and Schneier[4]

[4]D. Wagner and B. Schneier, *"Analysis of the SSL 3.0 Protocol"*, November 1996, Available: https://dl.acm.org/doi/10.5555/1267167.1267171

"[SSL] tra ffic analysis aims to recover confidential information about protection sessions by examining unencrypted packet fields and unprotected packet attributes. For example [...] the volume of network traffic flow"

— Wagner and Schneier[5]

... packet direction, timing, and more

[5] D. Wagner and B. Schneier, *"Analysis of the SSL 3.0 Protocol"*, November 1996, Available: https://dl.acm.org/doi/10.5555/1267167.1267171

Timeline

1996 Wagner and Schneier[1], coined SSL traffic analysis

1998 Cheng and Avnur, website traffic analysis

website fingerprinting (WF)...

2016 Abe and Goto, deep learning WF

2019 Kennedy et al., apply WF techniques to voice assistants (VA)

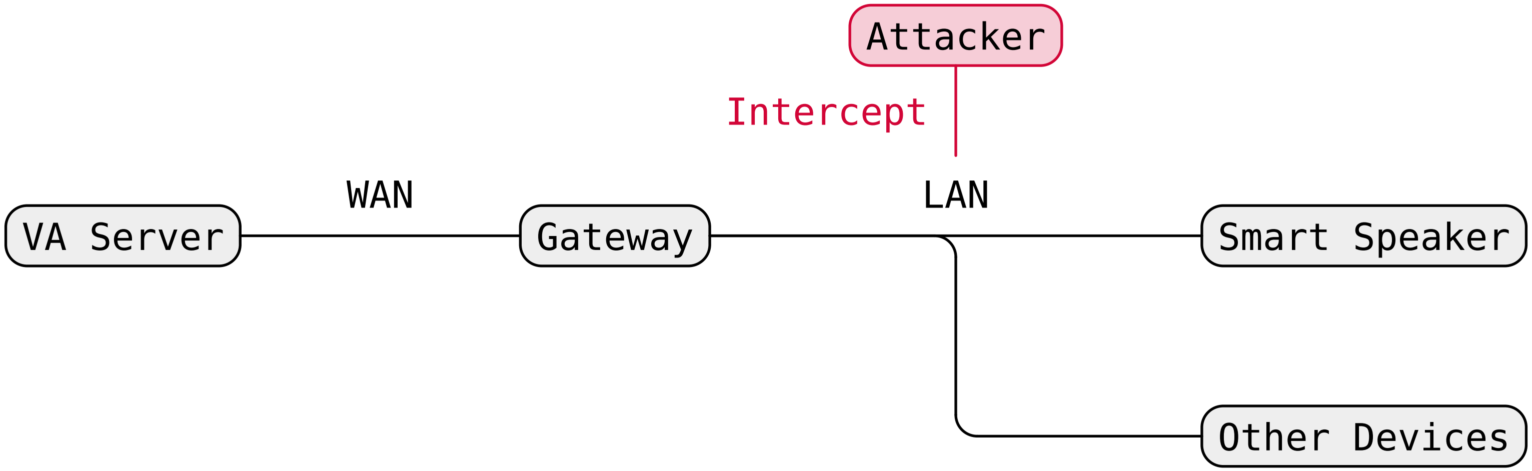
2020 Wang et al., deep learning VA fingerprinting

2022 Mao et al., temporal features

2023 Ahmed, Sabir and Das, invocation detection

[1]Timeline references can be found at the end of the presentation.

Threat Model



1. The attacker can intercept traffic from smart speaker
2. The attacker knows the smart speaker address
3. The attacker knows the type of smart speaker used
4. The attacker knows the beginning and end of an interaction

Closed-World

- Fixed list of monitored voice commands
- Traffic is considered to come from one of the monitored commands
- Multiclass classification

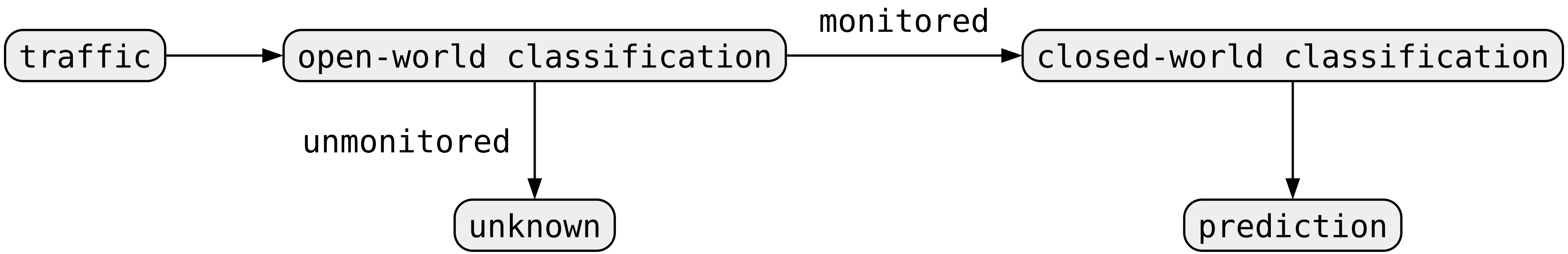
Predicts which command was used.

Open-World

- Traffic can also come from new, unmonitored commands
- Binary-classification

Predicts whether traffic is from monitored or unmonitored command.

Combining Both Models



A **Testbed** for Voice Assistant Traffic Fingerprinting

Comparison

Website Fingerprinting:

- Requires a large amount of data
- Data collection usually via program making requests
- Only dependent on network environment
- Fast

Voice Command Fingerprinting:

Comparison

Website Fingerprinting:

- Requires a large amount of data
- Data collection usually via program making requests
- Only dependent on network environment
- Fast

Voice Command Fingerprinting:

- Requires a large amount of data

Comparison

Website Fingerprinting:

- Requires a large amount of data
- Data collection usually via program making requests
- Only dependent on network environment
- Fast

Voice Command Fingerprinting:

- Requires a large amount of data
- Interaction by speaking

Comparison

Website Fingerprinting:

- Requires a large amount of data
- Data collection usually via program making requests
- Only dependent on network environment
- Fast

Voice Command Fingerprinting:

- Requires a large amount of data
- Interaction by speaking
- Hampered by environment noise

Comparison

Website Fingerprinting:

- Requires a large amount of data
- Data collection usually via program making requests
- Only dependent on network environment
- Fast

Voice Command Fingerprinting:

- Requires a large amount of data
- Interaction by speaking
- Hampered by environment noise
- Slow and inefficient

Comparison

Website Fingerprinting:

- Requires a large amount of data
- Data collection usually via program making requests
- Only dependent on network environment
- Fast

Voice Command Fingerprinting:

- Requires a large amount of data
- Interaction by speaking
- Hampered by environment noise
- Slow and inefficient

→ Sophisticated testbed

> "The content of voice commands may vary from date to date; therefore, more efficient data collection tools need to be developed."
>
> — Mao et al.[1]

[1] Jianghan Mao et al., *"A novel model for voice command fingerprinting using deep learning"*, March 2022, Available: https://doi.org/10.1016/j.jisa.2021.103085

Requirements

- Sound isolation
  - Isolated box
  - Separate speaker/microphone
- Efficiency
  - Every second saved per interaction means hours saved when collecting dozens of thousands interactions
  - Dynamic interaction length by listening for silence
- Robustness
  - Autonomously reset VA if error occurs
  - Monitoring system

System

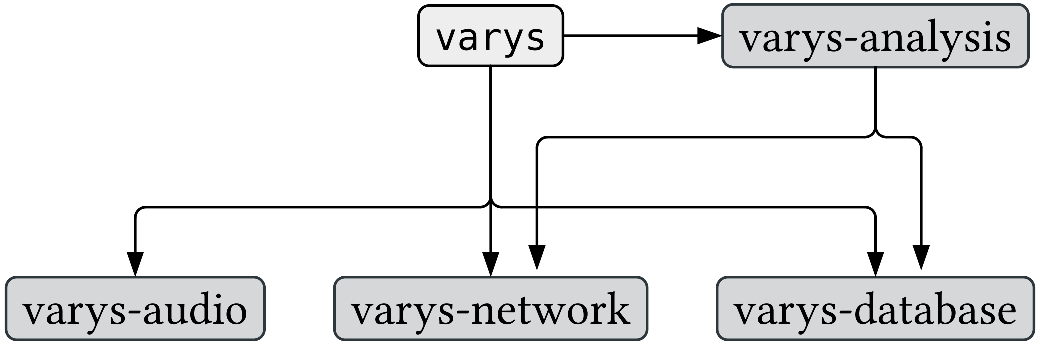
`varys` The main executable combining all modules into the final system.

`varys-analysis` Analysis of data collected by varys.

`varys-audio` Recording audio and the TTS and STT systems.

`varys-database` Abstraction of the database system where interactions are stored.

`varys-network` Collection of network traffic, writing and parsing of `.pcap` files.



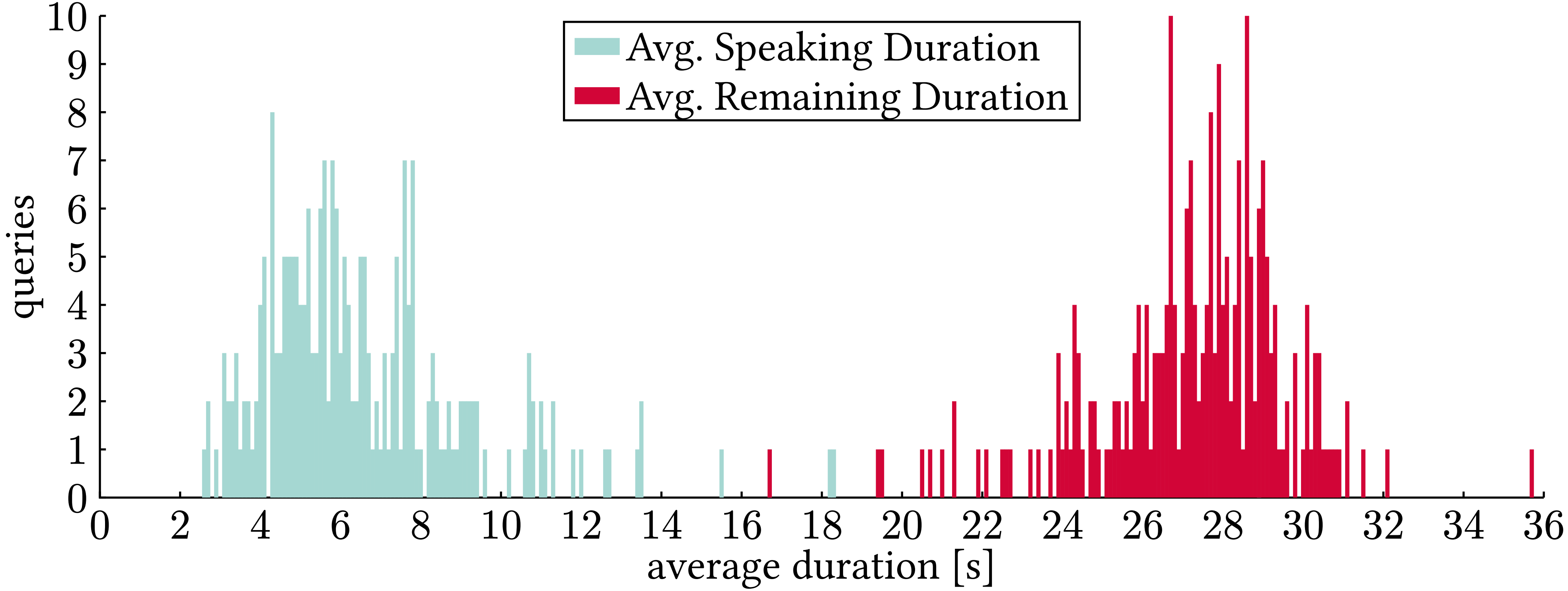
Results

Datasets

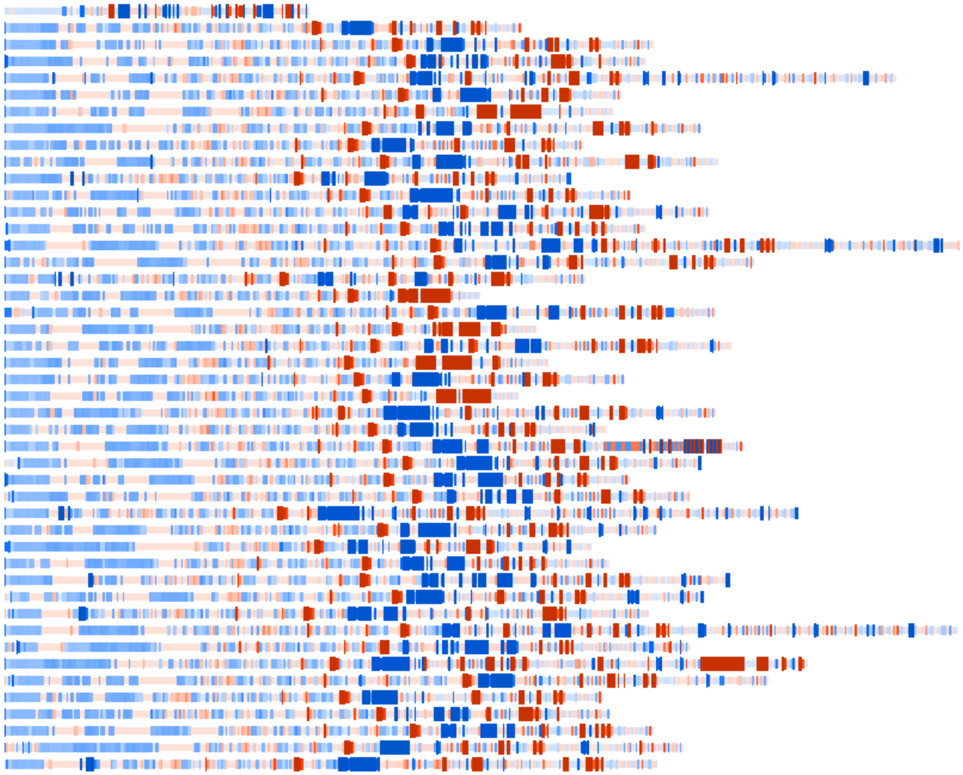
~800h, ~70′000 interactions

`large`	227 queries, 140 interactions each
`small`	13 queries, 2400 interactions each
`binary`	*“Call John Doe”* and *“Call Mary Poppins”*, 1500 interactions each

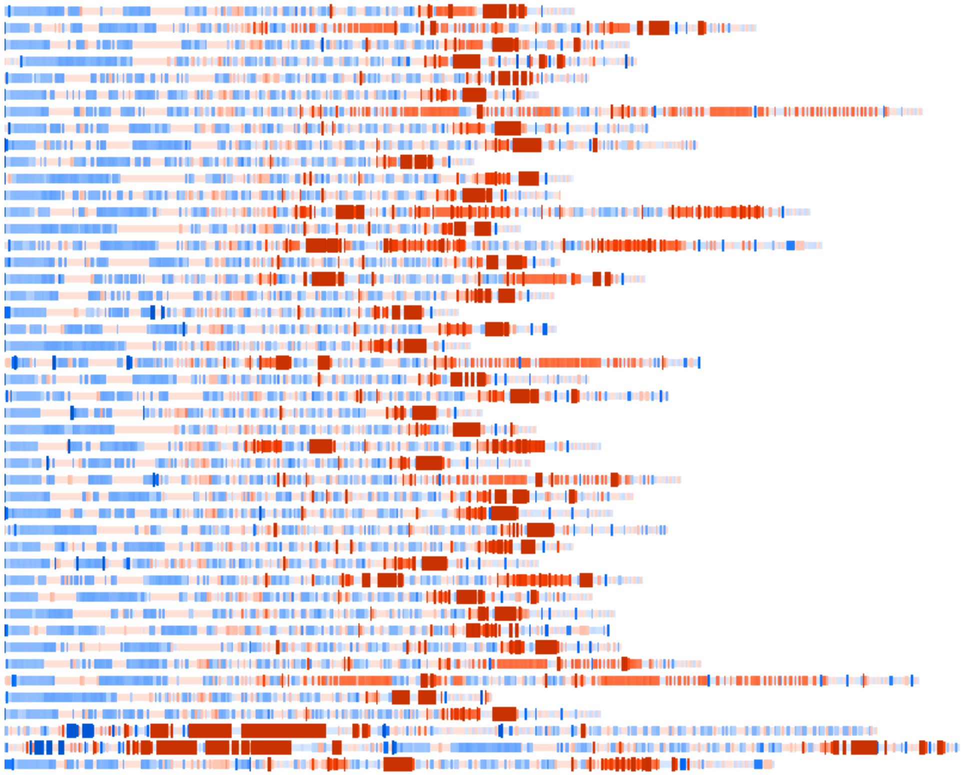
Efficiency


Avg. Speaking Duration
Avg. Remaining Duration
queries
0
1
2
3
4
5
6
7
8
9
10
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
average duration [s]

Traffic Trace Examples



"Any missed calls?"



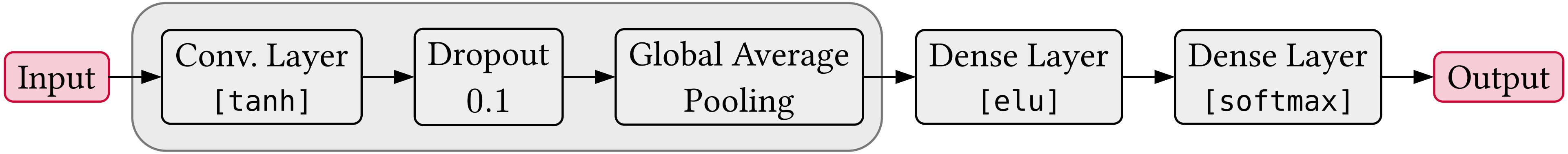
"What day was 90 days ago?"

Fingerprinting Model

Feature extraction (packet size $s \in [0, 1500]$ and direction $d \in \{0, 1\}$):

$$(s, d) \to (-1)^d \cdot \frac{s}{1500}$$

CNN adapted from Wang et al.[1]:



[1]Chenggang Wang et al., *"Fingerprinting Encrypted Voice Traffic on Smart Speakers with Deep Learning"*, May 2020, Available: https://doi.org/10.1145/3395351.3399357

Classification

Accuracy on test sets:

`large`	~40.40% (random choice ~0.44%)
`small`	~86.19% (random choice ~7.69%)
`binary`	~71.19% (random choice 50%)

Demo

```
./target/release/varys -i ap1 analyse demo data/ml/test_5_13\ queries_0.86 f4:34:f0:89:2d:75
```

“Hey Siri, any missed calls?”

“Hey Siri, remind me to wash the car.”

It is unlikely this will work...

Timeline References

- D. Wagner and B. Schneier, *"Analysis of the SSL 3.0 Protocol"*, November 1996, Available: https://dl.acm.org/doi/10.5555/1267167.1267171
- H. Cheng and R. Avnur, *"Traffic Analysis of SSL Encrypted Web Browsing"*, 1998
- K. Abe and S. Goto, *"Fingerprinting Attack on Tor Anonymity using Deep Learning"*, August 2016, Available: https://core.ac.uk/display/229876143
- S. Kennedy et al., *"I Can Hear Your Alexa: Voice Command Fingerprinting on Smart Home Speakers"*, June 2019, Available: https://doi.org/10.1109/CNS.2019.8802686
- Chenggang Wang et al., *"Fingerprinting Encrypted Voice Traffic on Smart Speakers with Deep Learning"*, May 2020, Available: https://doi.org/10.1145/3395351.3399357
- Jianghan Mao et al., *"A novel model for voice command fingerprinting using deep learning"*, March 2022, Available: https://doi.org/10.1016/j.jisa.2021.103085
- D. Ahmed, A. Sabir, and A. Das, *"Spying through Your Voice Assistants: Realistic Voice Command Fingerprinting"*, August 2023, Available: https://www.usenix.org/conference/usenixsecurity23/presentation/ahmed-dilawer