SAMSUNG Research

Voice Presentation Attack Detection through Text-Converted Voice Command Analysis

Il-Youp Kwak, Jun Ho Huh, Seung Taek Han, Iljoo Kim and Jiwon Yoon

Motivation

Emergence of Voice assistants:

Samsung Bixby, Apple Siri, Amazon Alexa, etc

Threat Model

TV anchor says live on-air 'Alexa, order me a dollhouse' – guess what happens next

Story on accidental order begets story on accidental order begets accidental order

By Shaun Nichols in San Francisco 7 Jan 2017 at 00:58 244 SHARE ▼

A San Diego TV station sparked complaints this week – after an on-air report about a girl who ordered a dollhouse via her parents' Amazon Echo caused Echoes in viewers' homes to also attempt to order dollhouses.

Security Critical Commands

Voice assistants now support security-critical commands, making them attractive target for adversaries to exploit

"Open Samsung Pay and show me the registered credit card"

"Take a picture with the front camera"

"Open Facebook and post a recent picture"

"Change unlock password"

How to Defend?

- Voice Biometric Authentication

- Replay Attack Detection

- Voice-Command based Attack Detection

Voice-Command based Attack Detection

Start from global model that detect security critical commands

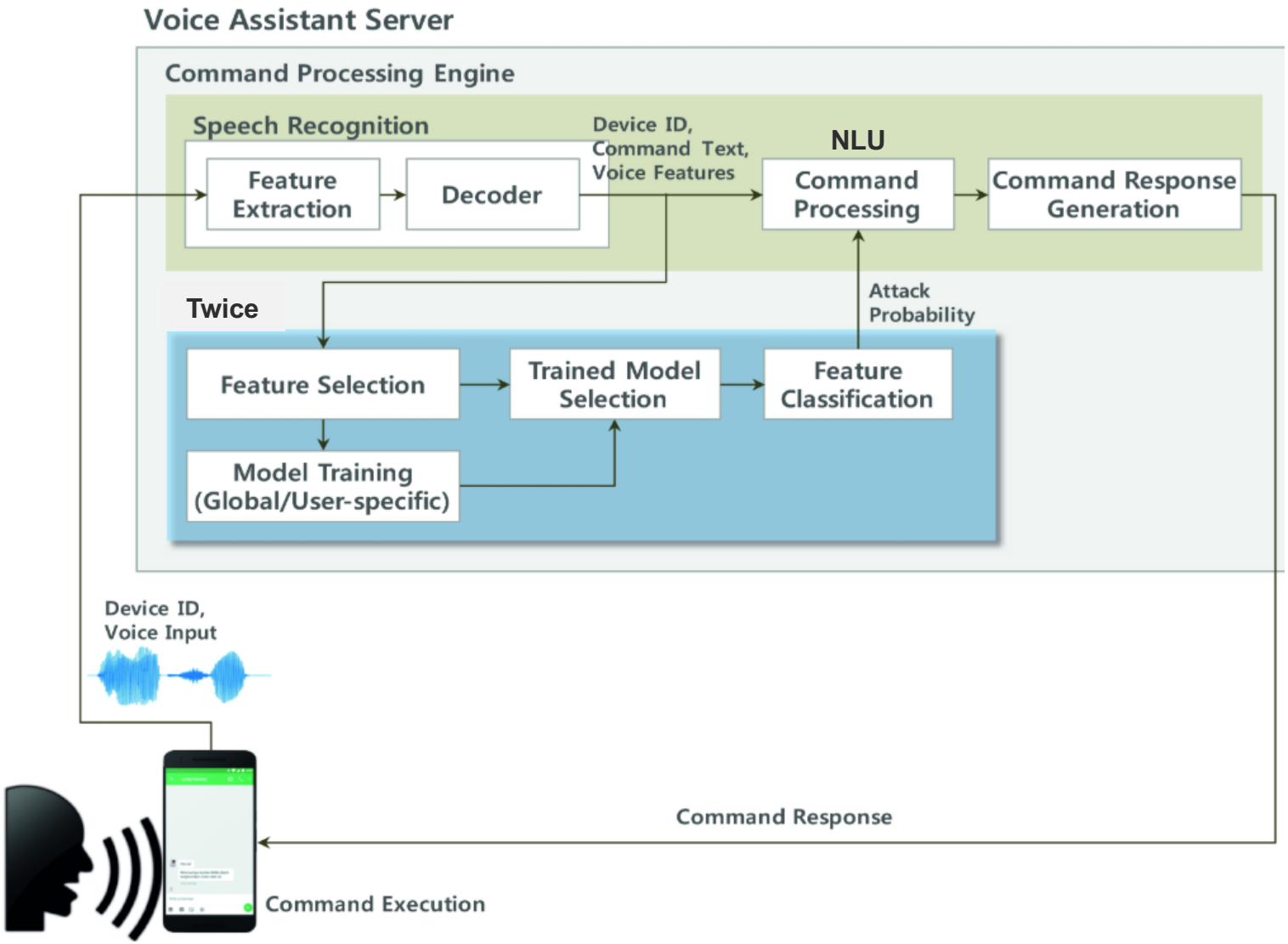
And grow to user model that detect security critical commands or user-specific commands

Security Critical Commands

Security critical command:
command that can be used in a voice presentation attack
to exploit one of the threats mentioned in
"I've Got 99 Problems" (Porter et al. 2012)

Synthetic attack set:
security-critical commands selected from
existing set of Bixby commands

System Overview



Modeling

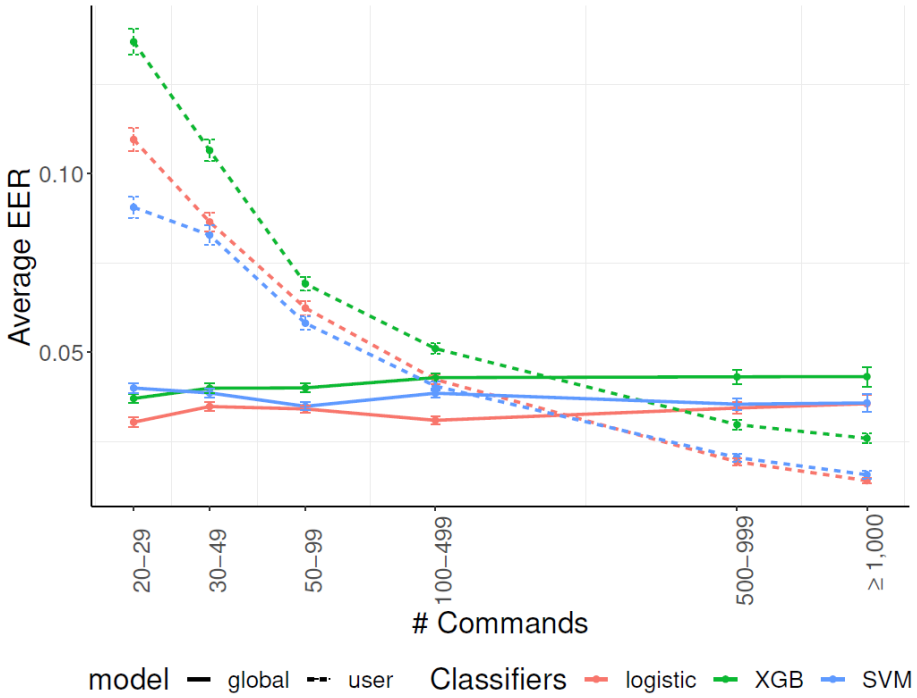
Considered BoW (Back of Words) feature, logistic regression model with LASSO penalty

Global model trained from random user commands vs security critical commands.

User specific model trained from user commands vs security critical commands

Evaluation

Start with the global model for new users (available immediately), and switch to the user-tailored models when users use about 500 commands



Accuracy against Unseen Data (FAR)

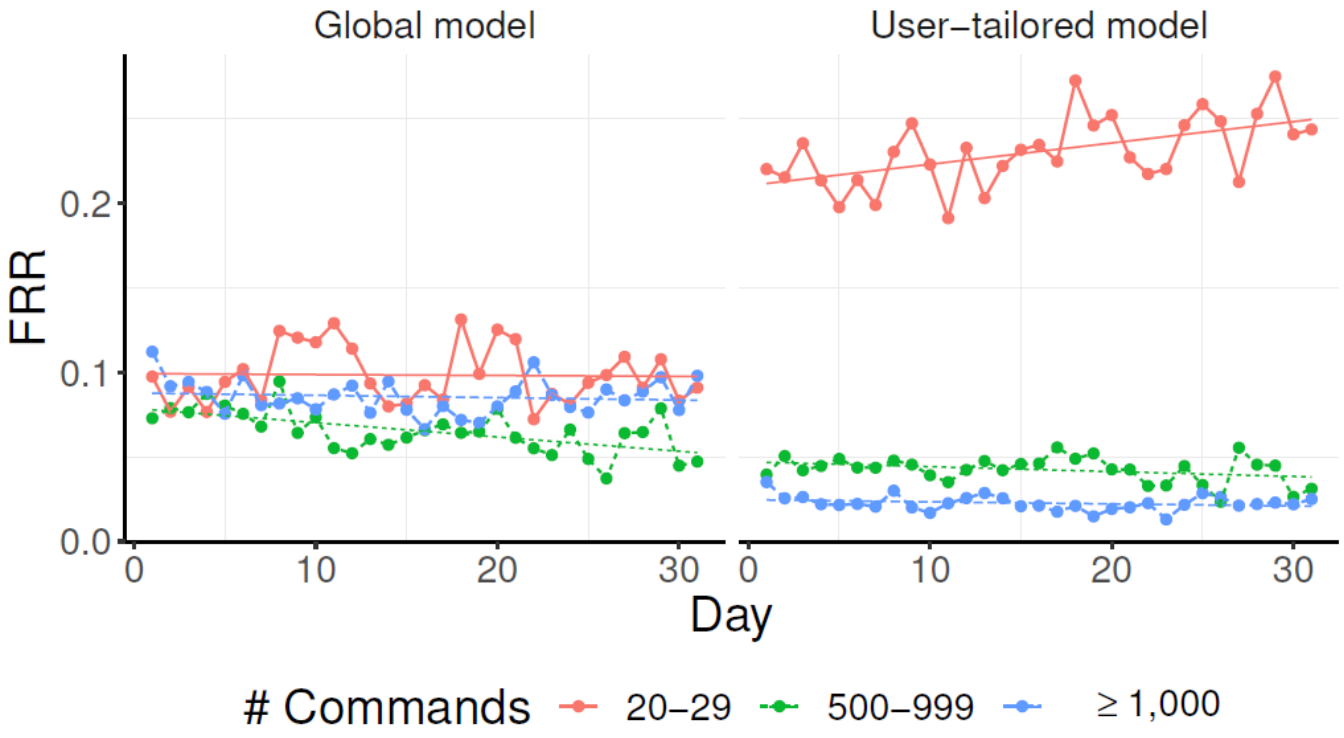
Global model achieves low average FAR at 4.3%

For those who have used 100 or more commands, the user-tailored models achieve lower FARs between 1.7–4.3%

	Global model	User-tailored model					
# Commands		20–30	30–50	50–100	100–500	500–1,000	>1,000
FAR	4.3% (0.03)	10.1% (0.10)	8.34% (0.08)	5.9% (0.06)	4.3% (0.04)	2.3% (0.02)	1.7% (0.02)

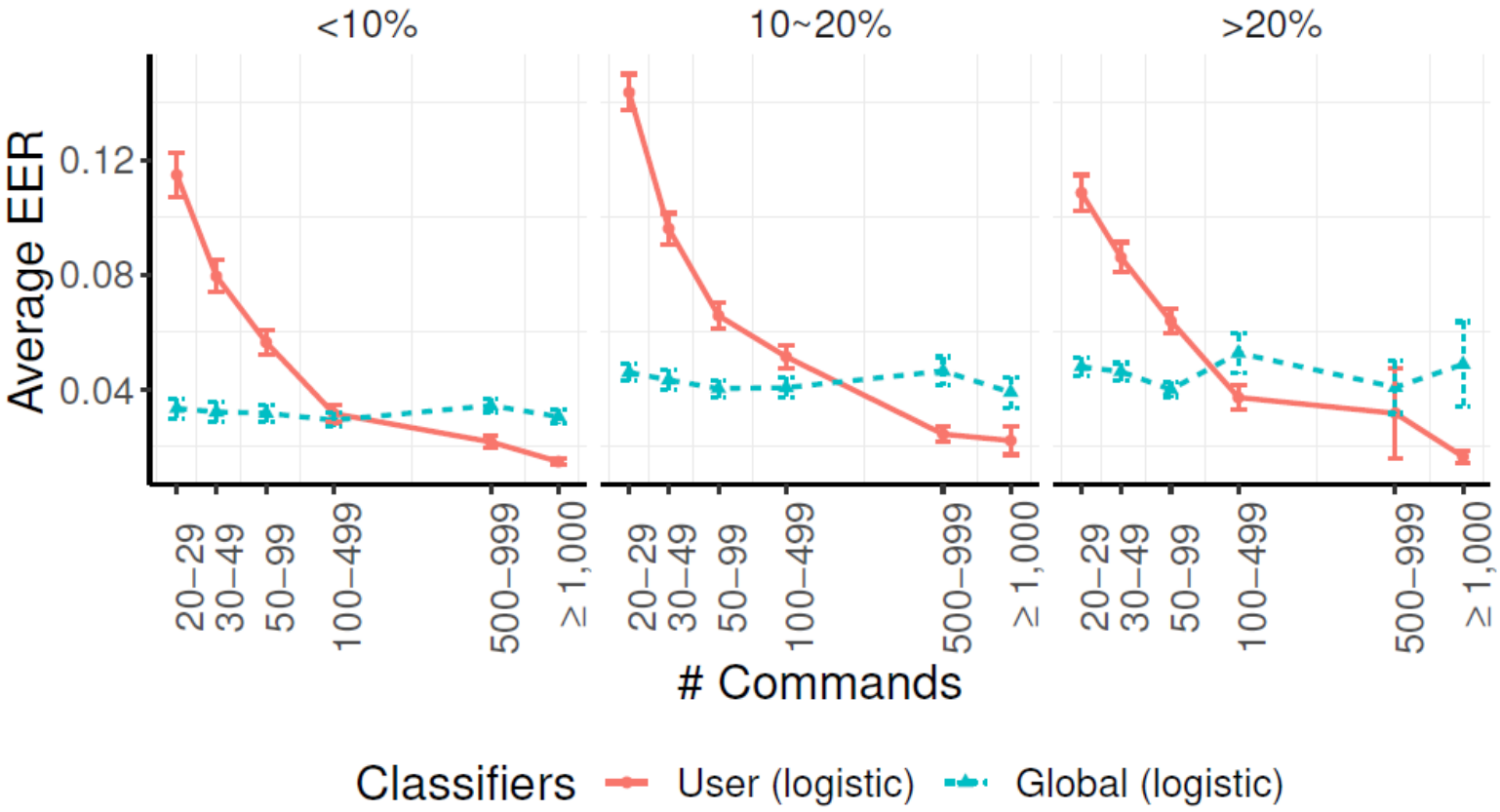
Accuracy against Unseen Data (FRR)

Global model should be used initially to maintain consistent FRRs below 10%, and eventually the user-tailored models would have to be used to achieve FRRs below 5%



Security-Critical Command Users

Twice will maintain average EERs below 5% even for those who frequently use security-critical commands.



Conclusion

Can be used as an effective complementary technology to further enhance user attack detection accuracy

Combined use of the global model and user-tailored models are integral in maintaining low and consistent EERs at around 3.4% for all users

To maximize detection accuracy and minimize false rejections, we recommend a switch (from global to user-tailored) when users have used up to about 500 commands.

Even for those users who frequently use security-critical commands, Twice is capable of achieving EERs below 5%.

Thank you

Shape the Future with Innovation and Intelligence