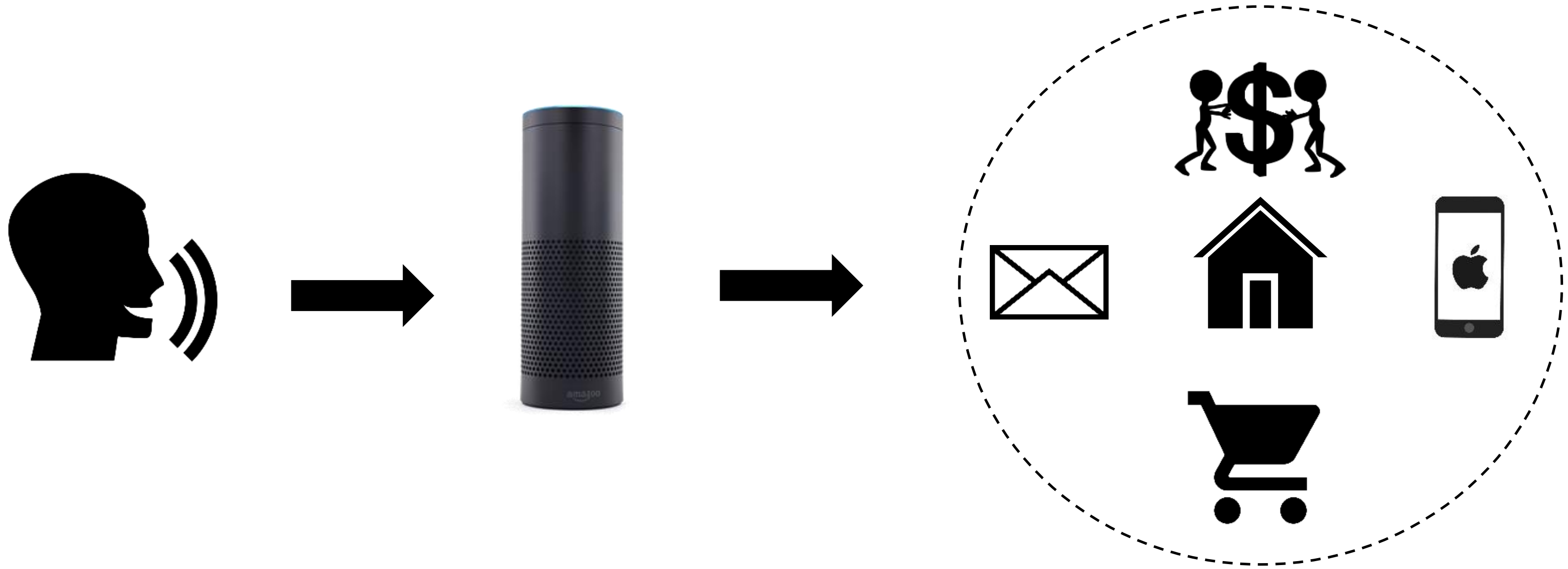
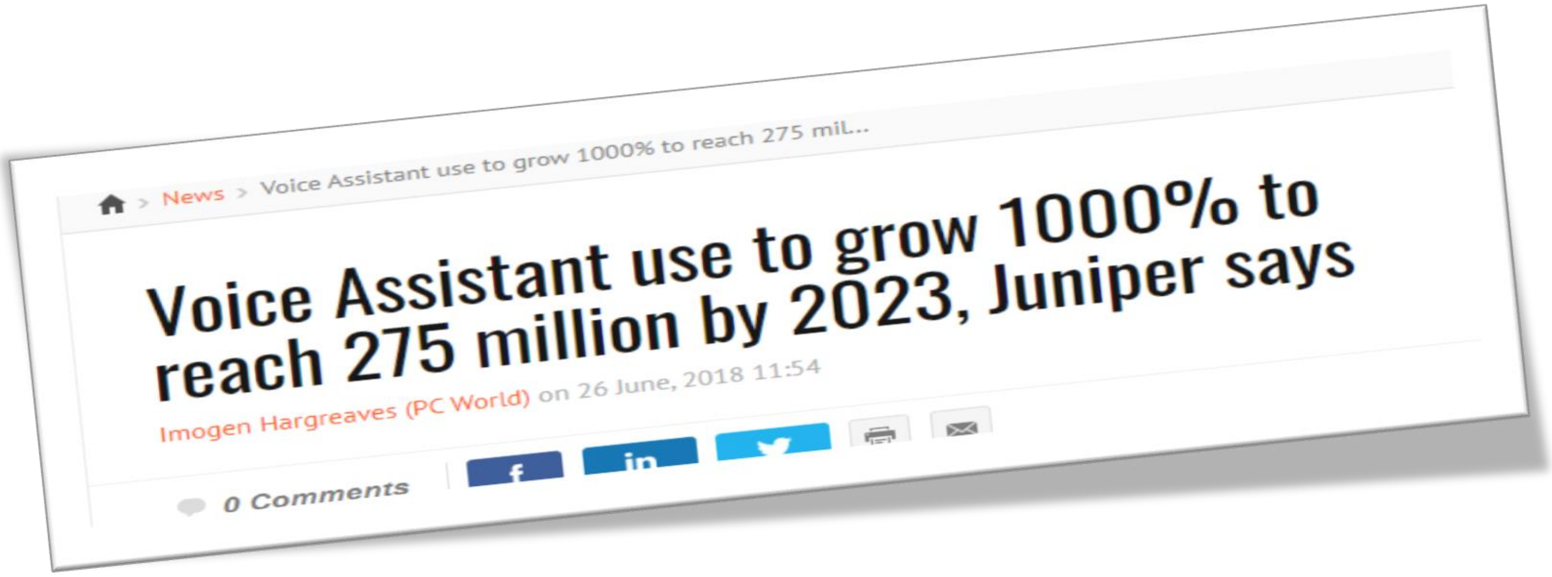
Void: A fast and light voice liveness detection system

Muhammad Ejaz Ahmed[1,2], Il-youp Kwak[4], Jun Ho Huh[3], Iljoo Kim[3], Taekyung Oh[2, 5], and Hyoungshick Kim[1, 2]

[1] CSIRO Data61, Australia
[2] Sungkyunkwan University, Korea
[3] Samsung Research, Korea
[4] Chung-Ang University, Korea
[5] KAIST, Korea





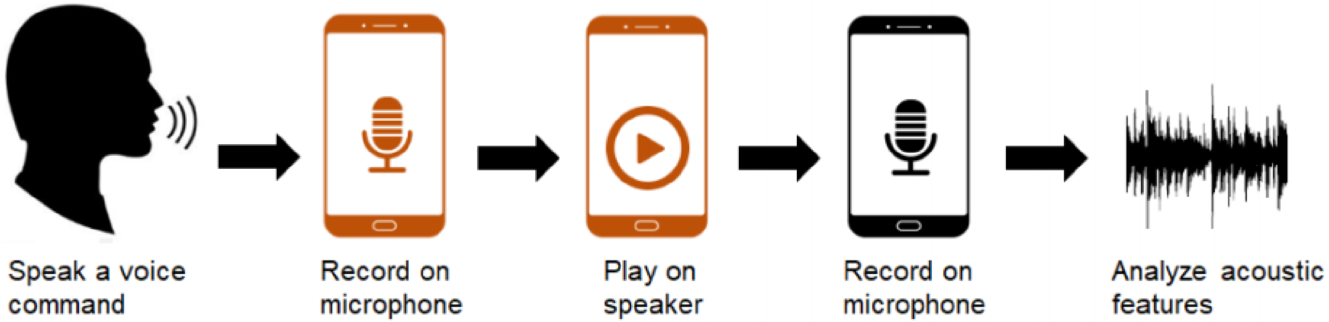
News > Voice Assistant use to grow 1000% to reach 275 mil...

Voice Assistant use to grow 1000% to reach 275 million by 2023, Juniper says

Imogen Hargreaves (PC World) on 26 June, 2018 11:54

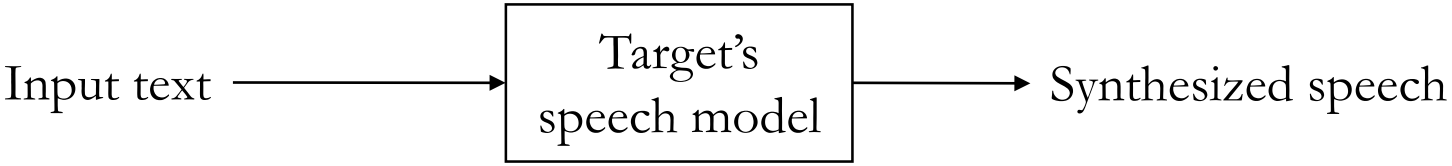
0 Comments

Voice replay attack

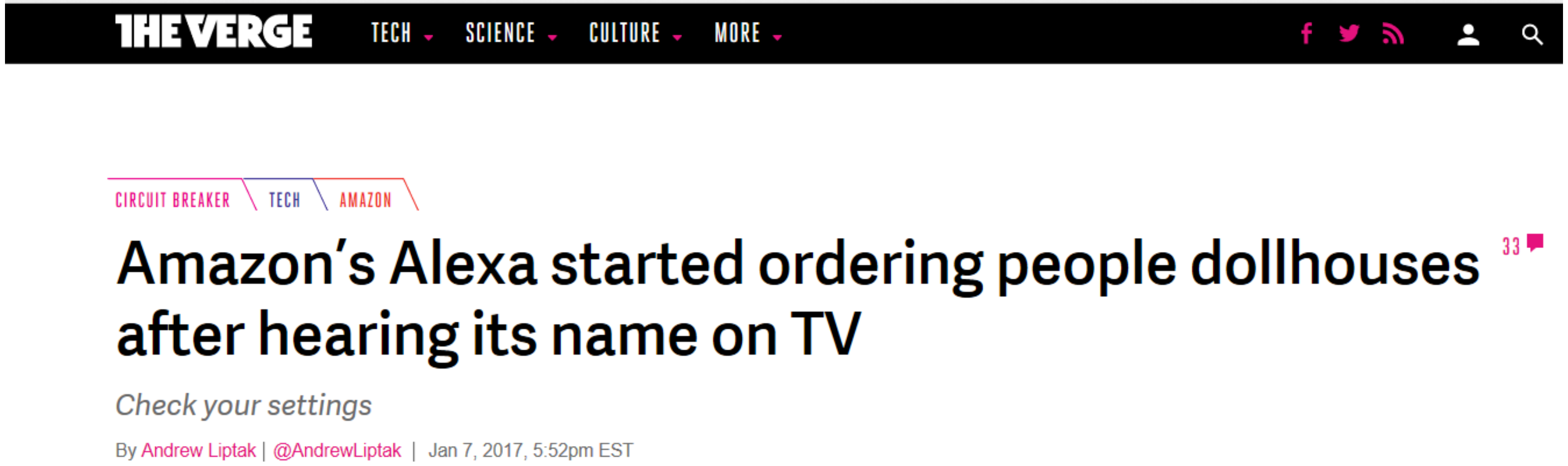


Speak a voice command → Record on microphone → Play on speaker → Record on microphone → Analyze acoustic features

Voice synthesis attack



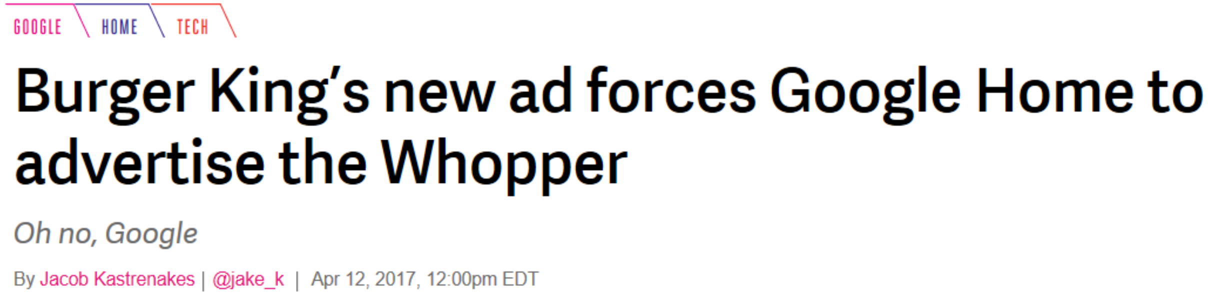
Input text → Target's speech model → Synthesized speech



Amazon's Alexa started ordering people dollhouses after hearing its name on TV

Check your settings

By Andrew Liptak | @AndrewLiptak | Jan 7, 2017, 5:52pm EST

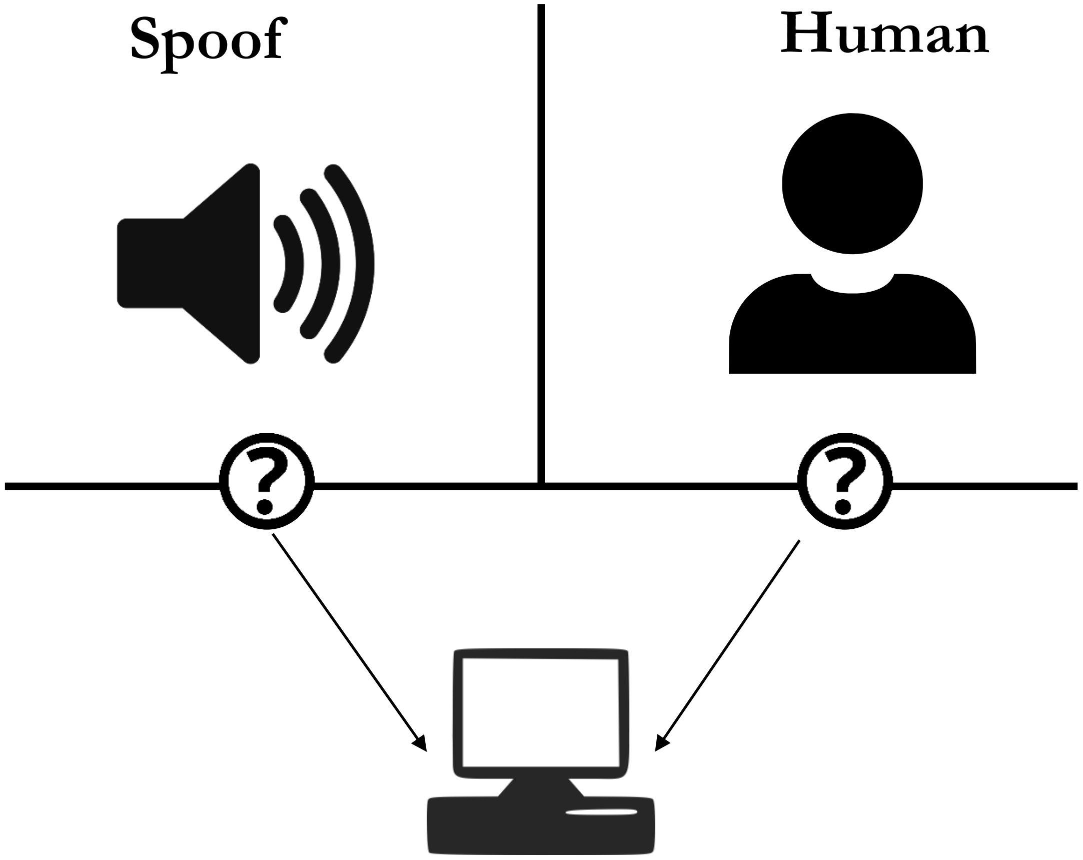


Burger King's new ad forces Google Home to advertise the Whopper

Oh no, Google

By Jacob Kastrenakes | @jake_k | Apr 12, 2017, 12:00pm EDT

Voice liveness detection



Voice liveness detection system

Requirements

- Latency and model size
  - Processing delay must be less than 100 milliseconds.
  - A single GPU may be expected to concurrently process 100 or more voice sessions.
  - On-device implementation without need to communicate with remote servers.

- Detection accuracy
  - Around 10% or below EER to be considered as a usable solution.

Our contributions

- Void: a fast, lightweight and easily implementable in commercial voice assistants.
    - Provide key insights for attack detection.
    - Single classification model with just 97 features.
    - Void is robust under numerous environmental settings.
- Evaluation using two large datasets consisting of:
    - 255,173 voice samples collected from 120 participants. EER achieved is 0.3%.
    - 18,030 ASVspoof competition voice samples collected from 42 participants. EER achieved 11.6% (second best-performing approach).
    - Void is about 8 times faster and uses 153 times less memory in detection compared to best-performing.
- Resilient against adversarial attacks. We evaluated it on:
    - Hidden voice attack: Accuracy 99.7%
    - Inaudible voice command (Dolphin attack): accuracy 100%
    - Voice synthesis attacks: accuracy 90.2%
    - Equalization manipulation attacks: accuracy 86.3%

Key insights

Key insight 1: Decay patterns in spectral power

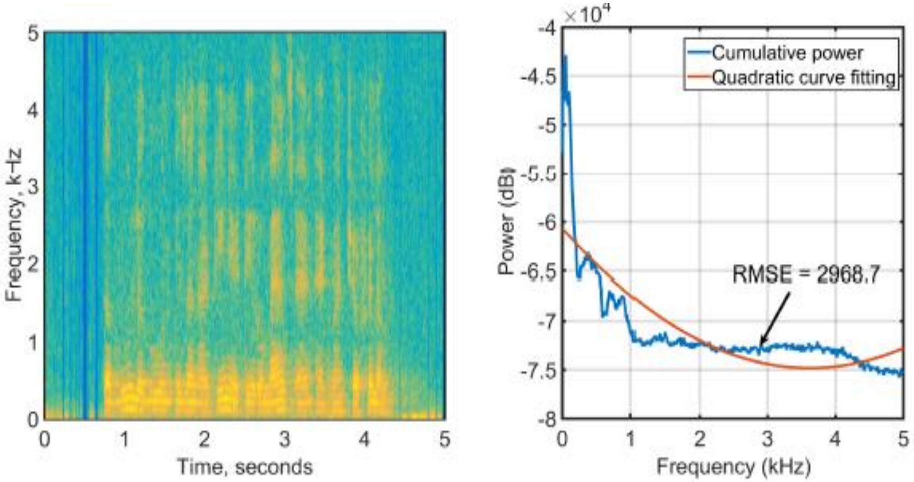


Fig. 1. Live-human voice sample

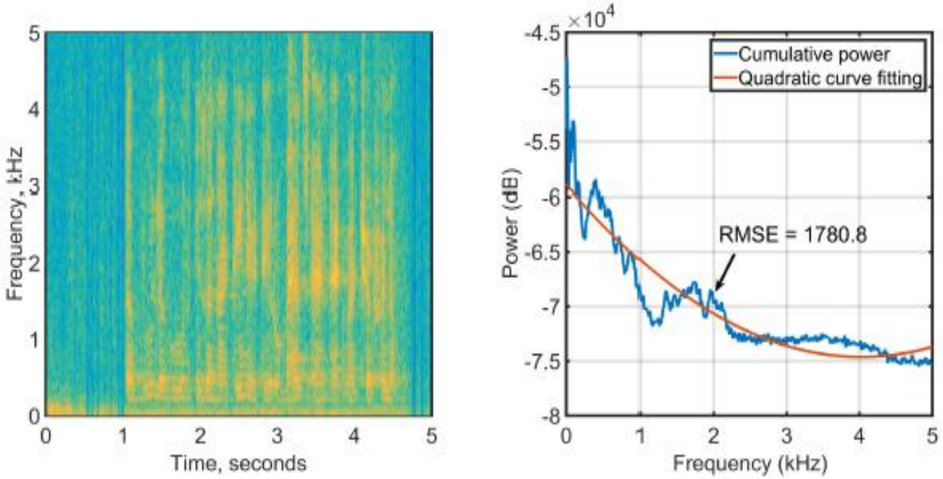
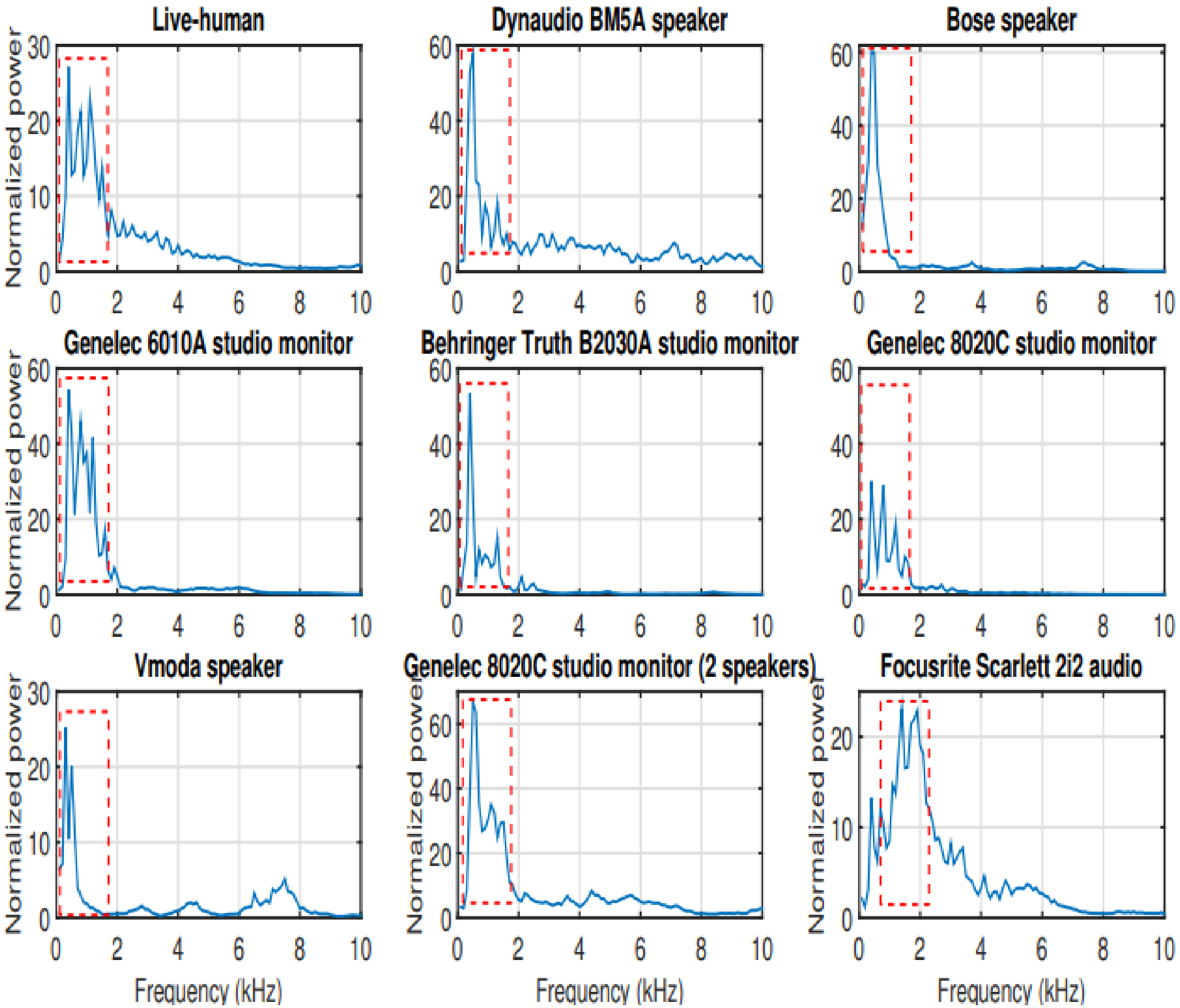


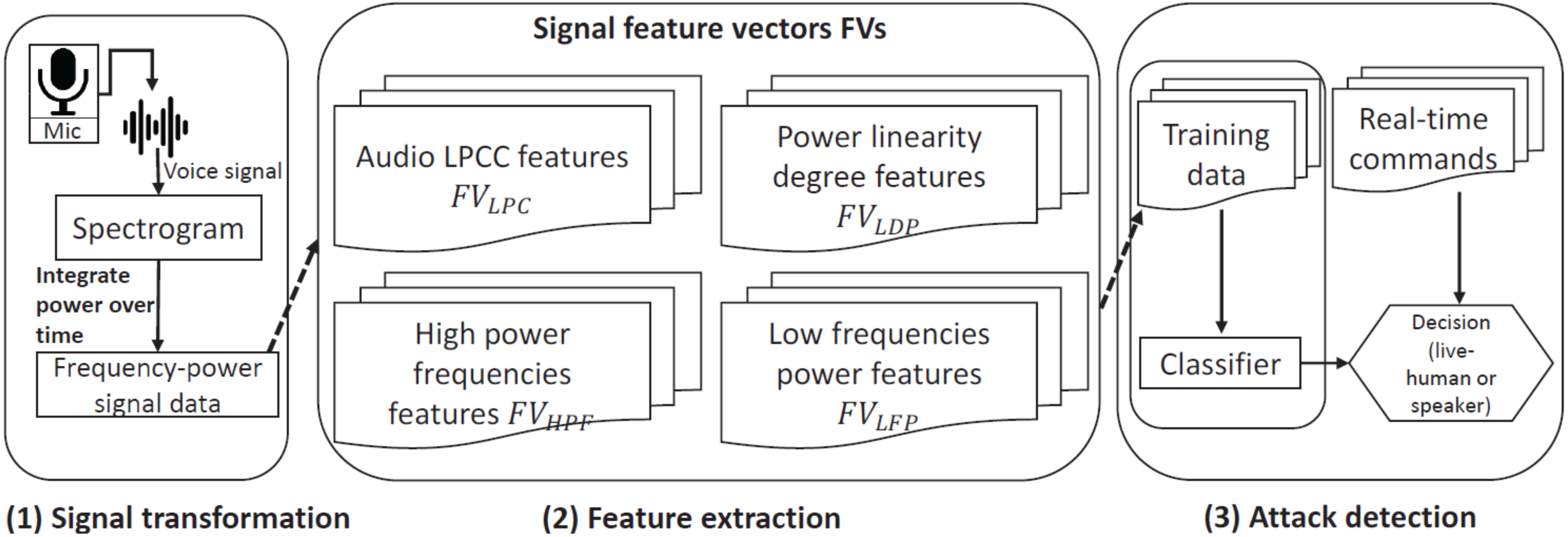
Fig. 2. Replayed using smartphone loudspeaker

Key insight 2 : Peak patterns in spectral power



Power patterns of live-human and different loudspeakers.

High level overview of Void


Signal feature vectors FVs
Mic
Voice signal
Spectrogram
Integrate power over time
Frequency-power signal data
Audio LPCC features FV_{LPC}
Power linearity degree features FV_{LDP}
High power frequencies features FV_{HPF}
Low frequencies power features FV_{LFP}
Training data
Real-time commands
Classifier
Decision (live-human or speaker)
(1) Signal transformation
(2) Feature extraction
(3) Attack detection

Data collection

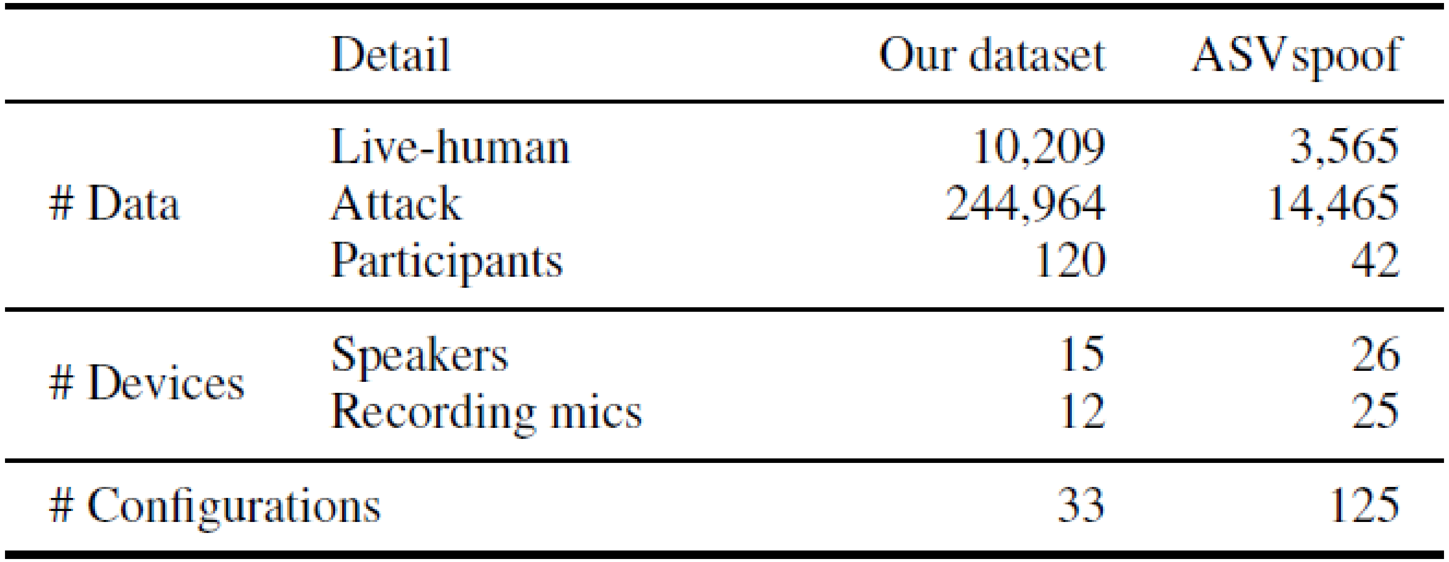
Our dataset

- 120 participants recruited for data collection. 53% of the participants were male.
- 50 commands from a prepared list of real-world voice assistant commands.
- Participants were in the 40-49 (13%), 30-39 (62%), and 20-29 (25%) age groups.

ASVspoof 2017 competition dataset

- Dataset contain three sets (training, development, and evaluation).
- Voice sample were collected from numerous environments such as balcony, bedroom, canteen, home, office, and lab space.

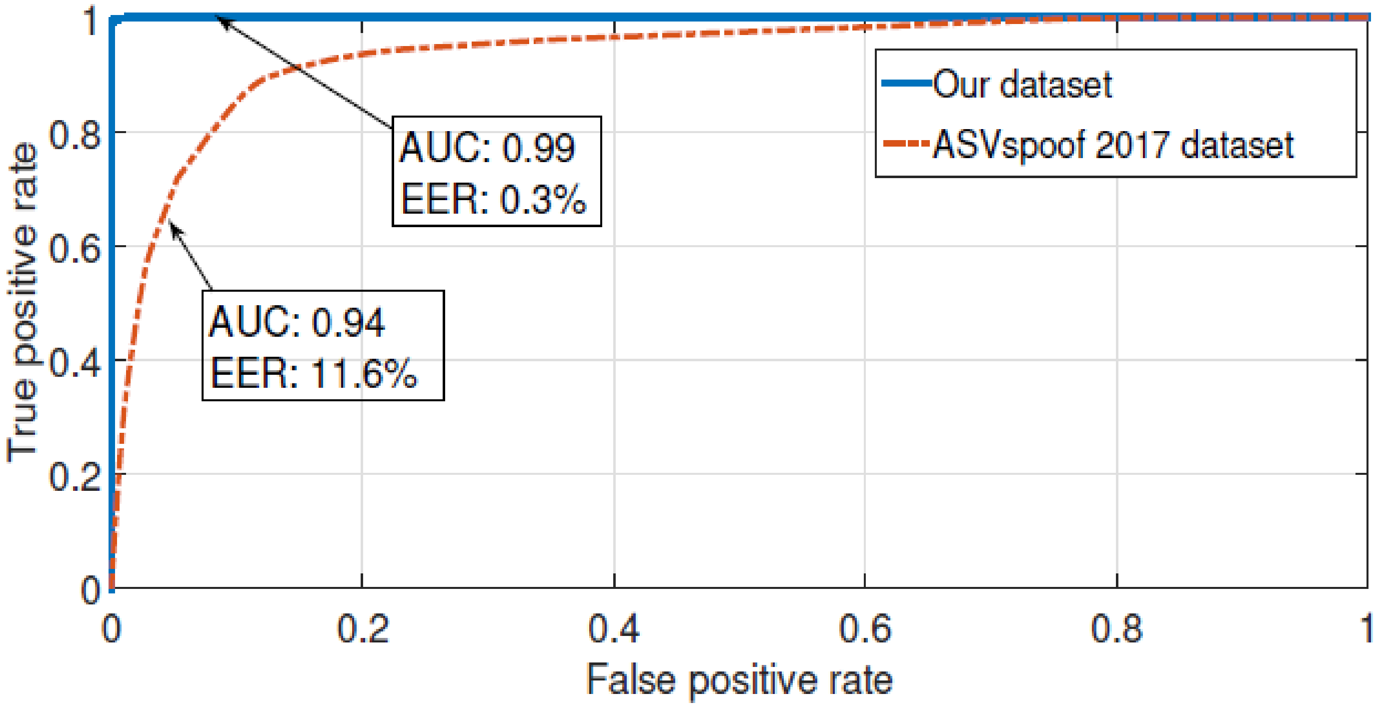
Datasets



	Detail	Our dataset	ASVspoof
# Data	Live-human	10,209	3,565
	Attack	244,964	14,465
	Participants	120	42
# Devices	Speakers	15	26
	Recording mics	12	25
# Configurations		33	125

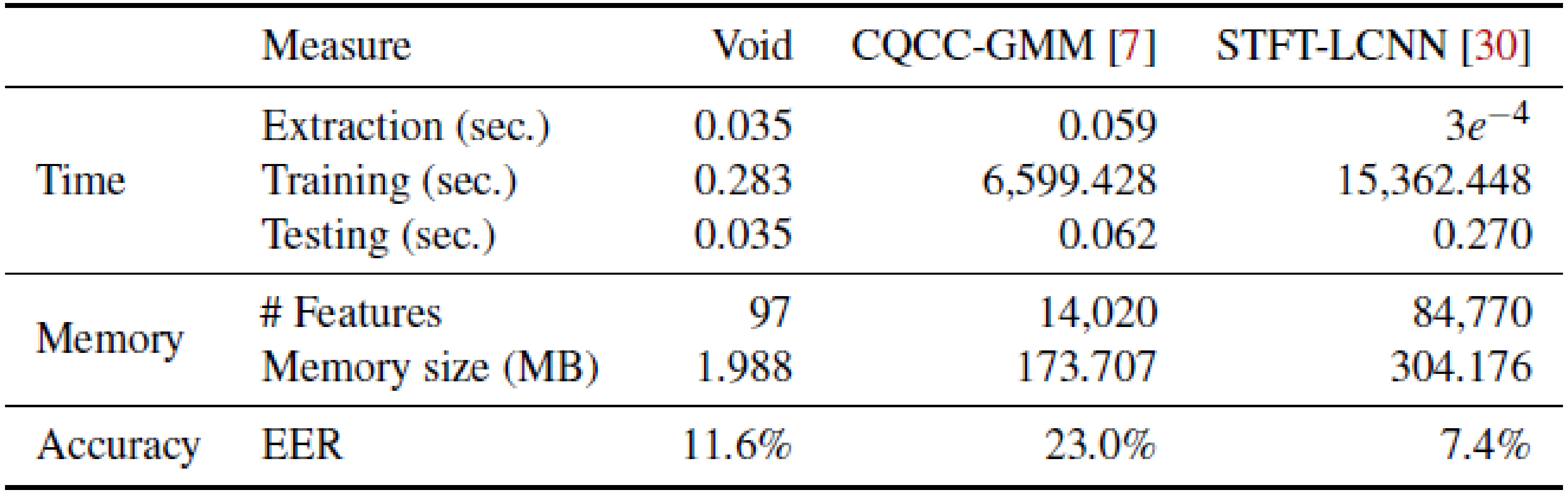
Evaluation

Overall performance



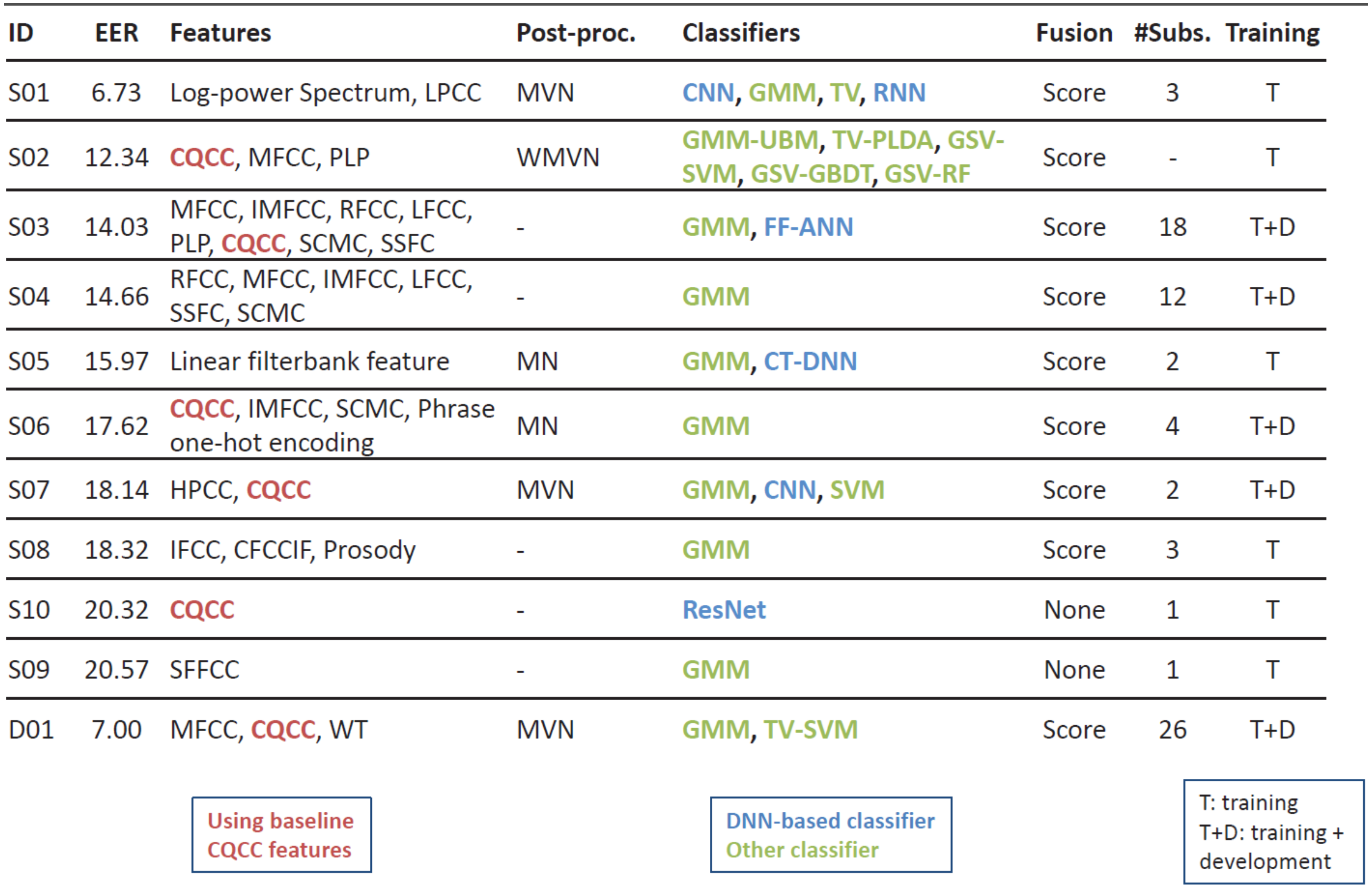
The EER reported in evaluation is computed using Bosaris toolkit (recommended by ASVspoof challenge competition).

Lightweight nature of Void



	Measure	Void	CQCC-GMM [7]	STFT-LCNN [30]
Time	Extraction (sec.)	0.035	0.059	$3e^{-4}$
	Training (sec.)	0.283	6,599.428	15,362.448
	Testing (sec.)	0.035	0.062	0.270
Memory	# Features	97	14,020	84,770
	Memory size (MB)	1.988	173.707	304.176
Accuracy	EER	11.6%	23.0%	7.4%

Void



ID	EER	Features	Post-proc.	Classifiers	Fusion	#Subs.	Training
S01	6.73	Log-power Spectrum, LPCC	MVN	CNN, GMM, TV, RNN	Score	3	T
S02	12.34	CQCC, MFCC, PLP	WMVN	GMM-UBM, TV-PLDA, GSV-SVM, GSV-GBDT, GSV-RF	Score	-	T
S03	14.03	MFCC, IMFCC, RFCC, LFCC, PLP, CQCC, SCMC, SSFC	-	GMM, FF-ANN	Score	18	T+D
S04	14.66	RFCC, MFCC, IMFCC, LFCC, SSFC, SCMC	-	GMM	Score	12	T+D
S05	15.97	Linear filterbank feature	MN	GMM, CT-DNN	Score	2	T
S06	17.62	CQCC, IMFCC, SCMC, Phrase one-hot encoding	MN	GMM	Score	4	T+D
S07	18.14	HPCC, CQCC	MVN	GMM, CNN, SVM	Score	2	T+D
S08	18.32	IFCC, CFCCIF, Prosody	-	GMM	Score	3	T
S10	20.32	CQCC	-	ResNet	None	1	T
S09	20.57	SFFCC	-	GMM	None	1	T
D01	7.00	MFCC, CQCC, WT	MVN	GMM, TV-SVM	Score	26	T+D

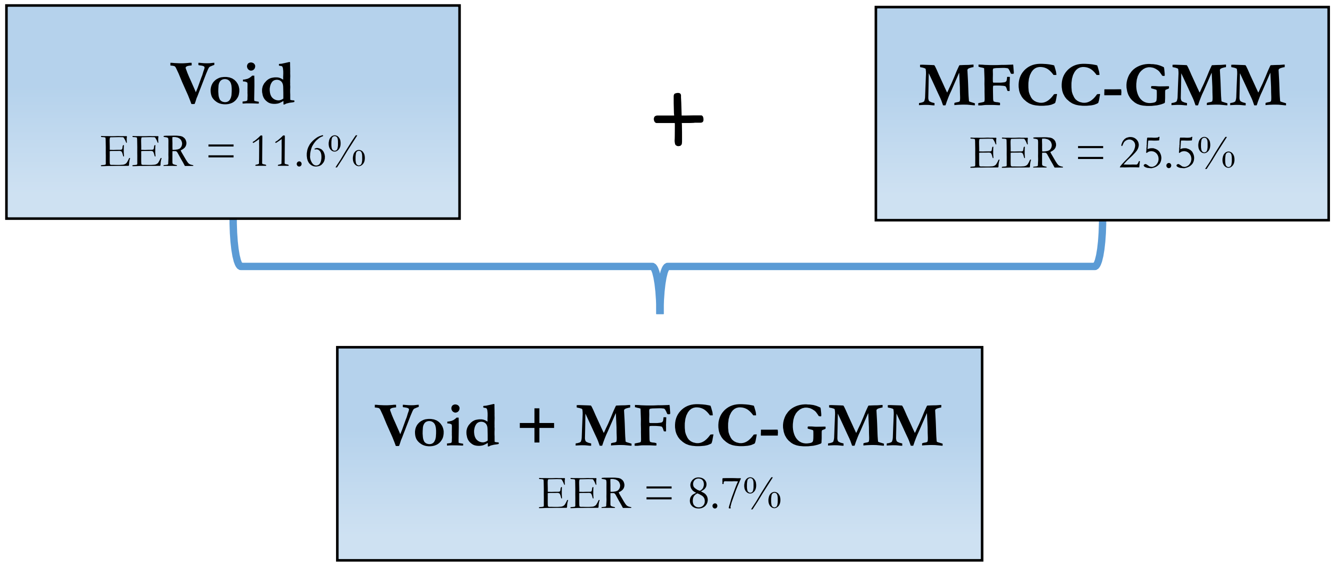
Using baseline CQCC features

DNN-based classifier
Other classifier

T: training
T+D: training + development

ASVspoof 2017 competition results [https://www.asvspoof.org/ slides_ASVspoof2017_Interspeech.pdf].

Void as an ensemble solution



Void

EER = 11.6%

+

MFCC-GMM

EER = 25.5%

Void + MFCC-GMM

EER = 8.7%

Adversarial attacks against Void

Void’s resilience against adversarial attacks

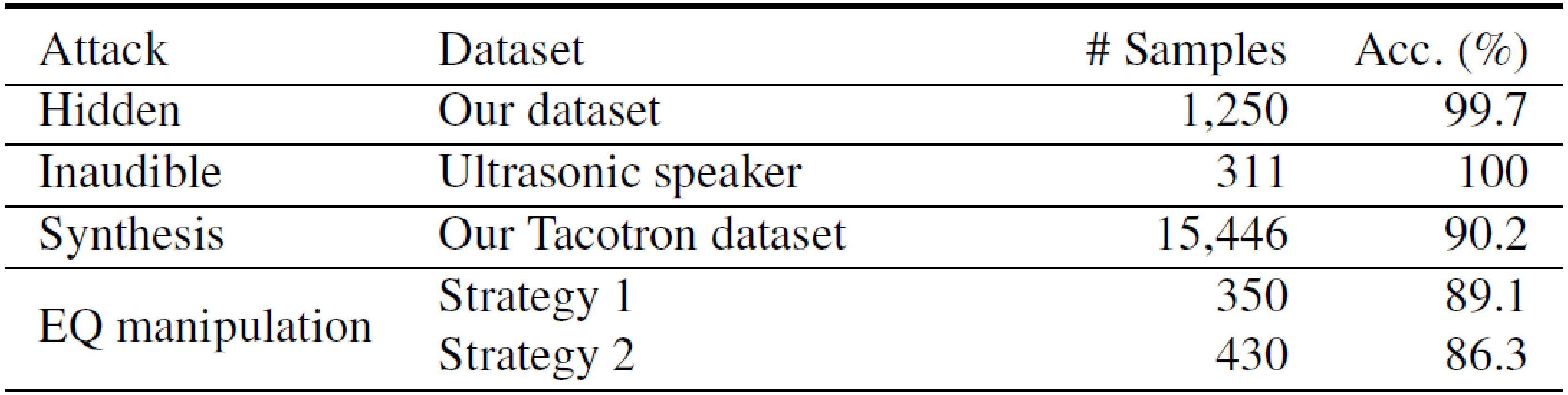
Hidden voice command: Hidden voice commands refer to commands that can not be interpreted by human ears but can be interpreted and processed by voice assistant services.

Inaudible voice command (Dolphin attack): Inaudible voice command attacks involve playing an ultrasound signal with spectrum above 20kHz, which would be inaudible to human ears.

Voice synthesis attack: Open source voice modeling tools called “Tacotron” and “Deepvoice 2” to train a user voice model with 13,100 publicly available voice samples.
We then used the trained model to generate 1,300 synthesis voice attack samples by feeding in commands as text inputs.

EQ manipulation attacks: EQ manipulation is a process commonly used for altering the frequency response of an audio system by leveraging linear filters.
By leveraging audio equalization, an attacker could intentionally manipulate the power of certain frequencies to mimic spectrum patterns observed in live-human voices.

Void’s resilience against adversarial attacks



Attack	Dataset	# Samples	Acc. (%)
Hidden	Our dataset	1,250	99.7
Inaudible	Ultrasonic speaker	311	100
Synthesis	Our Tacotron dataset	15,446	90.2
EQ manipulation	Strategy 1	350	89.1
	Strategy 2	430	86.3

Limitations of Void

- Void performance against high-quality speakers may degrade.

- EQ attack results show that carefully crafted voice samples can bypass Void. However, such attack would require strong signal processing expertise.

Conclusions

- Lightweight:
  - Void runs on single efficient classification model with 97 features and does not require addiction hardware.
  - Void is 8 times faster and 153 times lighter than top performing solution of ASVspoof competition.
  - On average Void took 35 milliseconds to classify a voice sample and just 1.98MB memory.
  - On-device implementation possible.
- Efficient:
  - Our evaluation on two large datasets, Void achieves 0.3% and 11.6% EER, respectively.
- Void is also resilient against various adversarial attacks.

Thank you!