Nicolas Müller

Replay Attacks Against Audio Deepfake Detection

Fraunhofer AISEC

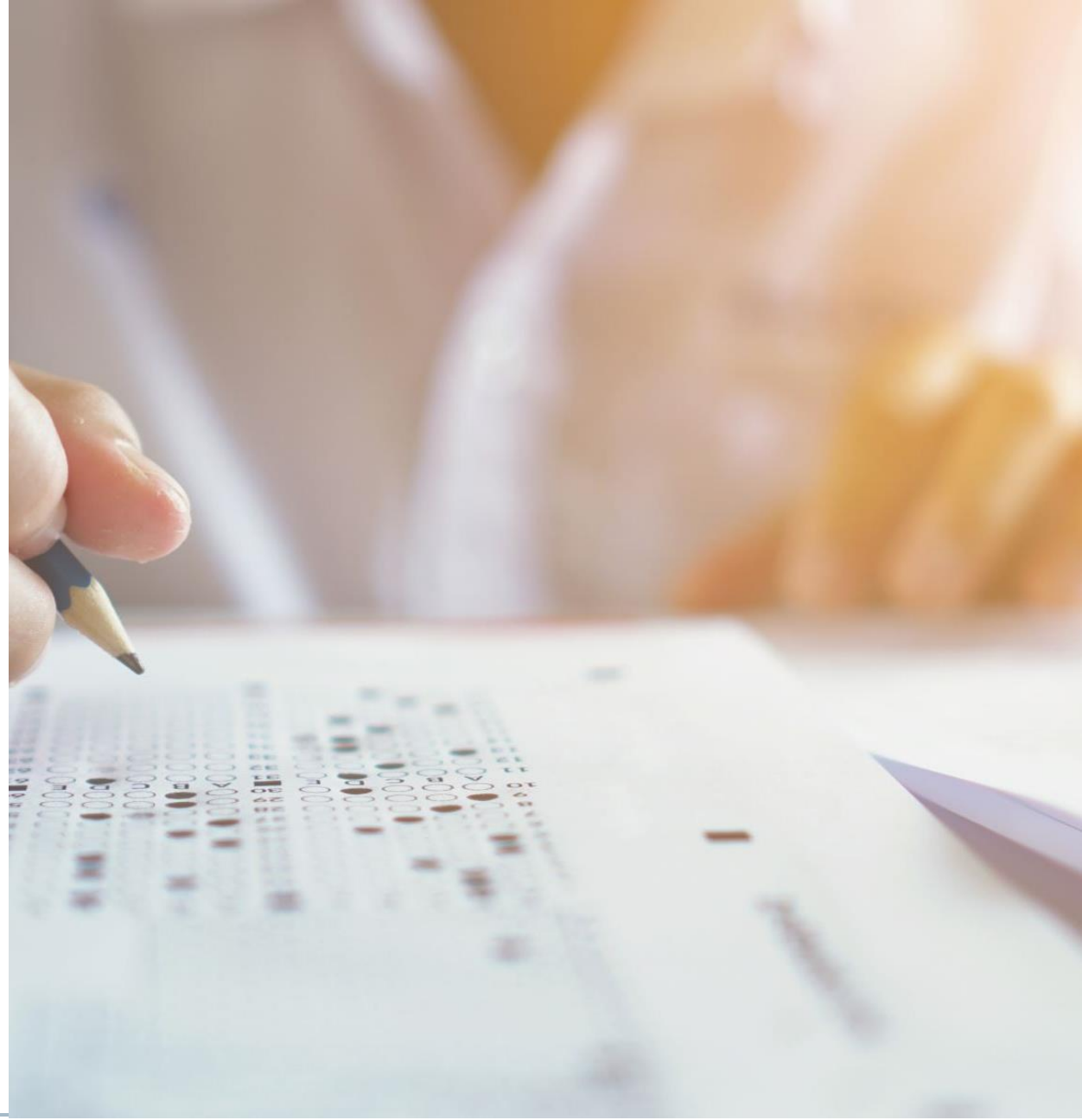
Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Agenda

Introduction & Problem Statement

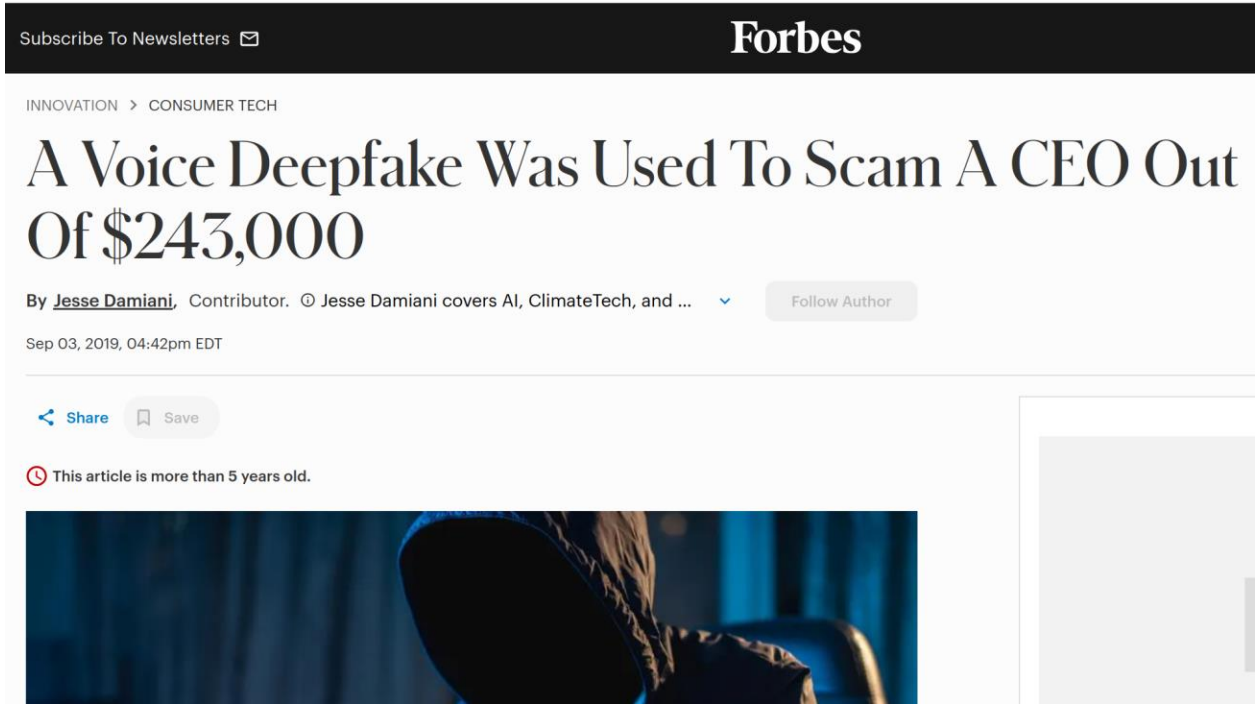
Proposed Solution

Evaluation and analysis



Audio Deepfake Detection

- Text-to-Speech allows to easily clone anybody's voice
- Simple, fast, cheap via SaaS providers
- Been used for misinformation, slander, fraud, …
- Urgent need to develop AI-driven audio deepfake detection tools


Subscribe To Newsletters

Forbes

INNOVATION > CONSUMER TECH

A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

By Jesse Damiani, Contributor. Jesse Damiani covers AI, ClimateTech, and ... Follow Author

Sep 03, 2019, 04:42pm EDT

Share Save

This article is more than 5 years old.

DEEPFAKE TOTAL

Analyze suspicious audio files to detect deepfakes, and automatically share them with the security community.

Youtube | Twitter / X | File Upload

Enter a Youtube URL

Analyze

Problem Statement

Replay attack: Play audio through loudspeaker and re-record

Makes audio deepfakes much harder to detect by ML-based classifiers

Proposed Solution: ReplaySpoof

Approach:
- Create a dataset of recordings of both fake and real audio
- Using a diverse set of microphones and loudspeakers

Use this to:
- Evaluate Vulnerability of Deepfake Detect (DFD) models against replay attacks
- Devise mitigation strategy

Approach

Systematic approach:

1) Set-up Speaker and Microphone
2) Run script `python main.py --mic "Microphone A42" --speaker "Speaker Denon H1337"`
3) This will:
   - Capture Room Impulse Response of Room
   - Play and record audio files (see next slide)

Approach

Audio files:

- Use M-AILABS (https://github.com/imdatceleste/m-ailabs-dataset) for real and MLAAD (https://deepfake-total.com/mlaad) for fake audio files

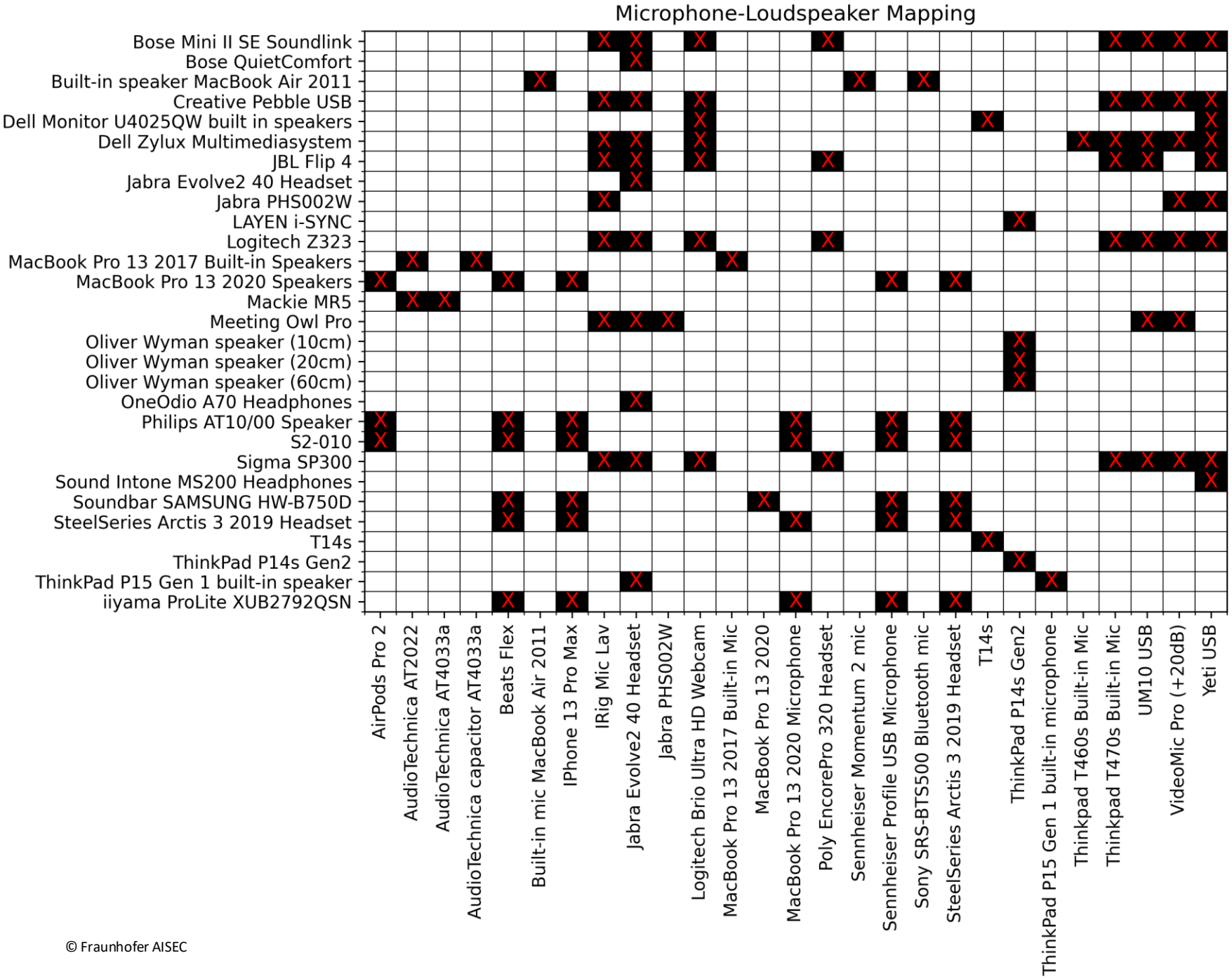
- Play 10 fake and 10 real audio files, for 6 languages (EN, DE, FR, PL, IT, ES) and four TTS models (XTTS v1.1, XTTS v2.0, Bark, Vits)

- Results in (10+10)*6*4 = 480 audio files per speaker/microphone combination

- 109 speaker/mic combinations

Algorithm 1 The data creation pipeline for *ReplayDF*. For each setup (i.e. per combination of loudspeaker and microphone), we select $n = 10$ instances for each language and TTS model from both MLAAD v5 (spoof) and M-AILABS (bona fide). Recordings and original audio files are stored to create a balanced dataset of air-gapped vs. non-air-gapped data.

$I = [id_0, id_1, ..., id_{108}]$ ▷ Speaker/microphone used.
$n = 10$
$R = \{\}$ ▷ List to hold recorded audio files.
$O = \{\}$ ▷ List to hold original audio files.
for id ∈ I **do**
  for lang ∈ {en, de, fr, it, pl, es} **do**
    for model ∈ {bark, vits, xtts_v1.1, xtts_v2.0} **do**
      $A \leftarrow$ choose n from `M-AILABS(lang)`
      $B \leftarrow$ choose n from `MLAAD(lang, model)`
      for w ∈ $A \cup B$ **do**
        r = `play_and_record(id, w)`
        $R \leftarrow R \cup \{r\}$
        $O \leftarrow O \cup \{s\}$
      end for
    end for
  end for ▷ $6 \cdot 4 \cdot 2 \cdot 10 = 480$ recordings per id.
end for ▷ Total number of recordings: $|I| \cdot 480 = 52320$.



Fraunhofer AISEC

Learnings

Insight 1: ReplaySpoof consistently makes classifiers perform worse

	Accuracy (%) ↑		EER (%) ↓	
	Baseline	ReplayDF	Baseline	ReplayDF
Whisper [27]	57.9	50.0	44.7	49.5
Raw PC Darts [28]	69.4	56.6	32.1	43.9
RawNet2 [15]	74.3	57.1	25.9	43.1
TCM ADD [18]	73.5	59.6	13.3	37.3
RawGAT-ST [29]	79.4	58.7	19.8	40.2
W2V2-AASIST [30]	90.0	74.2	10.6	24.8

Table 2: *Performance of Open-Source models with publicly available checkpoints in mean accuracy and EER over ReplayDF, as well as the original audio files (Baseline). The threshold for accuracy computation is as specified in the respective original publication. In all scenarios, replay attacks deteriorate model performance.*

Learnings

Insight 2: This holds true regardless of training data

Training Dataset	Accuracy (%) ↑		EER (%) ↓	
	Baseline	ReplayDF	Baseline	ReplayDF
ASVspoof 19	74.6 ± 6.6	55.1 ± 3.2	14.3 ± 7.7	34.7 ± 3.1
ASVspoof 5	67.6 ± 8.6	52.8 ± 1.4	12.5 ± 4.1	34.8 ± 3.5
Fake-or-Real	54.2 ± 0.9	49.5 ± 0.0	28.4 ± 1.6	45.0 ± 1.3
In-the-Wild	66.8 ± 3.1	52.9 ± 1.4	30.2 ± 3.2	42.4 ± 1.2
ODSS	94.7 ± 0.7	77.7 ± 2.4	4.7 ± 0.9	18.2 ± 1.5

Table 3: *Performance of W2V2-AASIST, trained on five different datasets, and evaluated on* ReplayDF. *Results computed over three independent trials, with mean and standard deviation shown.*

Learnings

Insight 3: Performance drops only for spoofed instances: ReplayAttacks seem to remove "fakeness" traces, possibly fingerprint of TTS model

Attack	No Augmentation		With Augmentation	
	Baseline	ReplayDF	Baseline	ReplayDF
Bark	82.6	40.7	83.0	56.9
VITS	82.2	53.4	75.7	65.8
XTTS v1.1	100.0	66.3	99.7	76.2
XTTS v2	100.0	59.4	99.8	73.6
bona fide	98.3	97.7	99.9	98.6

Table 4: *Detection accuracy [%] (↑) of W2V2-AASIST on the* Baseline *and* ReplayDF*, with and without RIR augmentation. While TTS-generated spoofs cause substantial accuracy reductions (up to* -43.6 *percentage points), bona fide detection remains unaffected. Incorporating RIR augmentation during training reduces the effect of the replay attacks.*

Learnings

Insight 4: Room Impulse Response (RIR) Augmentations help mitigate attack, but not fully

Attack	No Augmentation		With Augmentation	
	Baseline	ReplayDF	Baseline	ReplayDF
Bark	82.6	40.7	83.0	56.9
VITS	82.2	53.4	75.7	65.8
XTTS v1.1	100.0	66.3	99.7	76.2
XTTS v2	100.0	59.4	99.8	73.6
bona fide	98.3	97.7	99.9	98.6

Table 4: *Detection accuracy [%] (↑) of W2V2-AASIST on the* Baseline *and* ReplayDF*, with and without RIR augmentation. While TTS-generated spoofs cause substantial accuracy reductions (up to* -43.6 *percentage points), bona fide detection remains unaffected. Incorporating RIR augmentation during training reduces the effect of the replay attacks.*

Replay quality and detection performance

Does replay quality correlate with detection performance?

=> Have human evaluator listen to recordings and evaluate their quality

Rating	Description	Speech Quality	Distortion (background noise, overdrive, etc.)
5	Excellent	Clear	Imperceptible
4	Good	Clear	Slightly perceptible, but not annoying
3	Fair	Understandable	Perceptible and slightly annoying
2	Poor	Understandable	Perceptible and annoying
1	Very Poor	Barely understandable	Very annoying and objectionable
f	Failure	Inaudible	Heavy

Insight 5: Detection Performance (red) correlates with recording quality (blue, green). Pearson correlations of 0.423 (MOS) and 0.509 (PESQ).

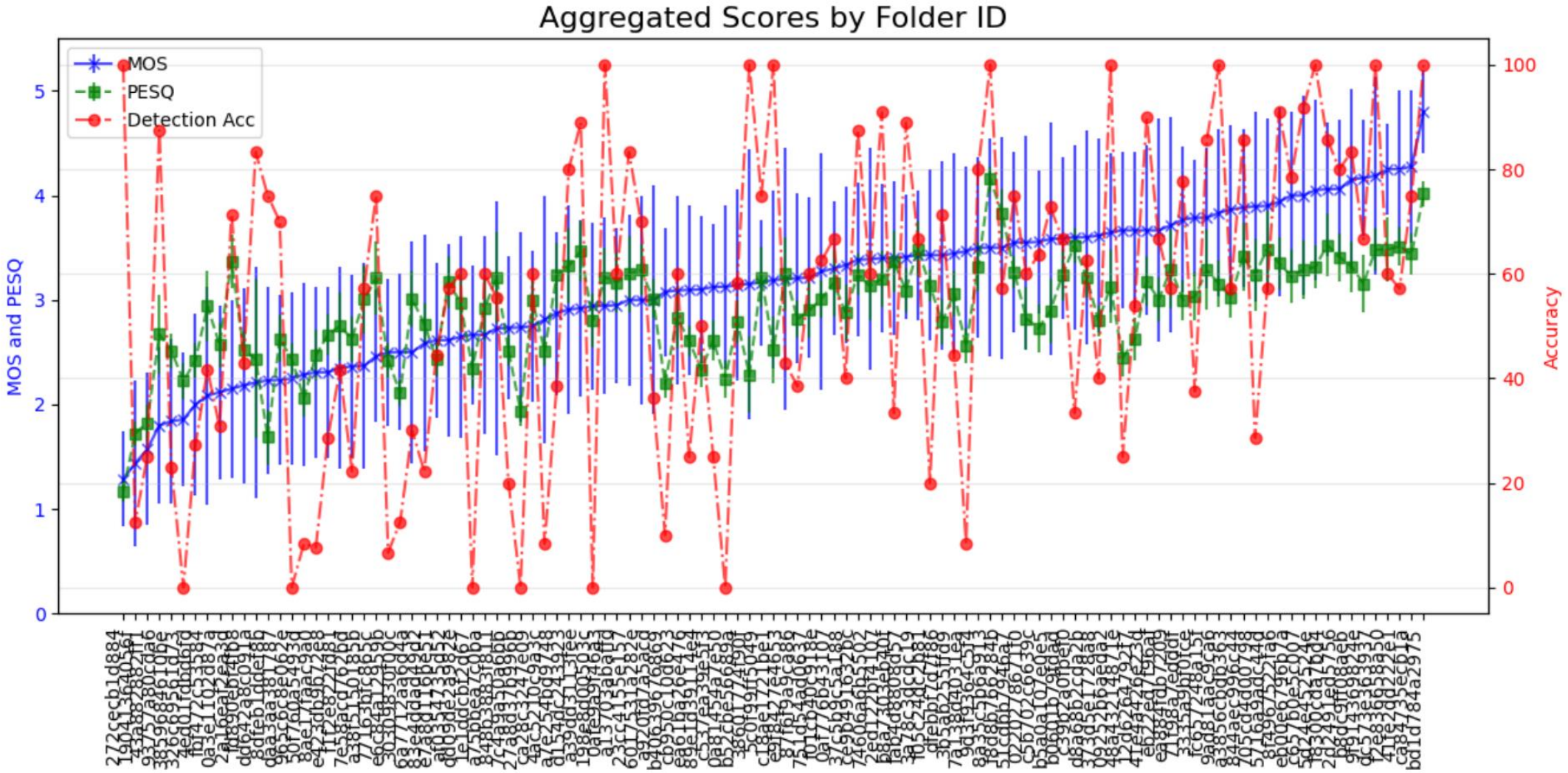


Figure 2: *Overview over the recording quality in* ReplayDF *as per (PESQ and MOS), as well as the detection performance (Accuracy) on only the spoofed samples.*

Replay quality and detection performance

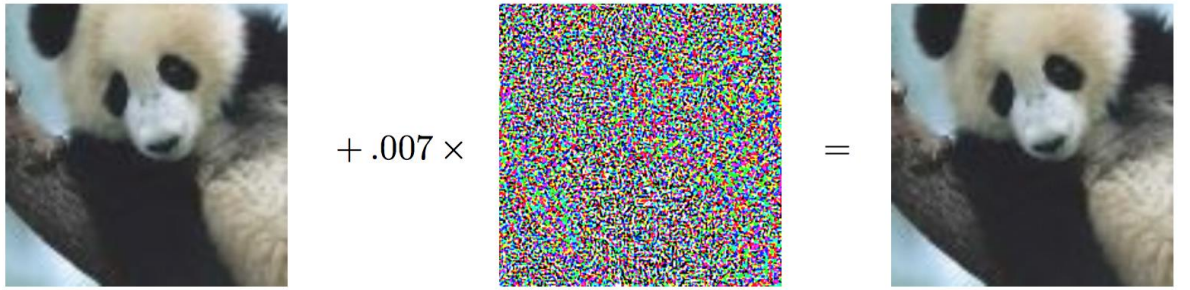
Maybe this correlation is because:

- The worse the recording, the more the "characteristics" of spoof-ness are removed
- Thus, the worse detection performance
- (Remember: performance on bona-fide unaffected by replayed audio)

Replay quality and detection performance

Adversarial Examples:

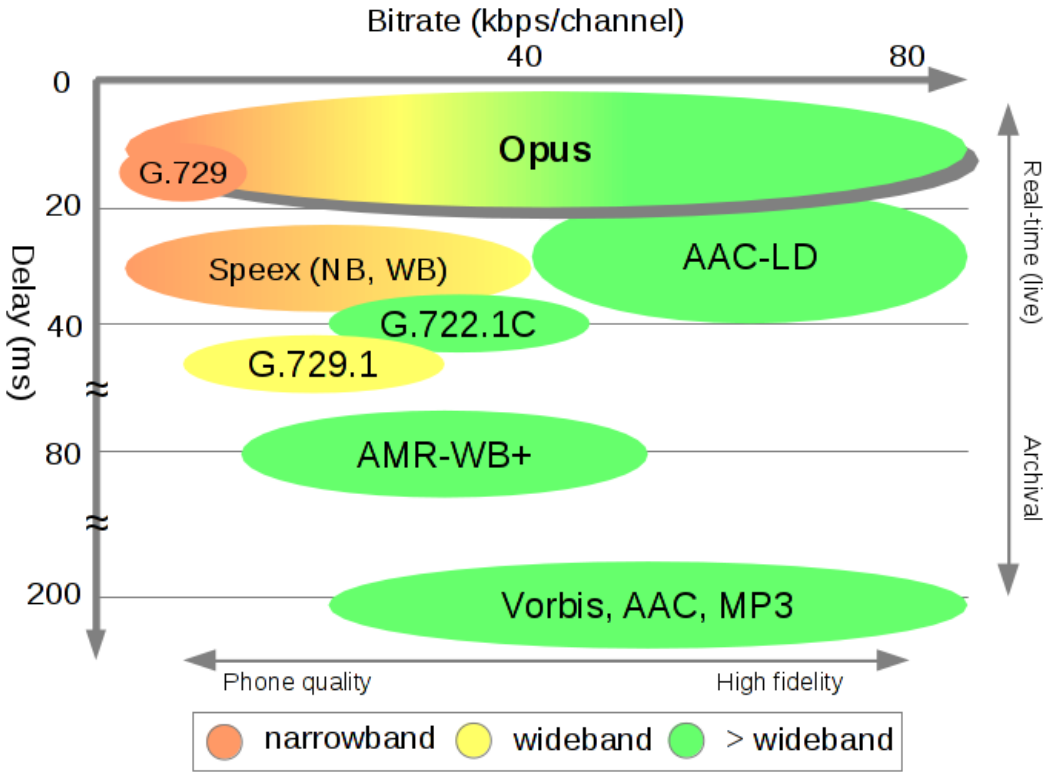
- Similar phenomenon for over-the-air adversarial attacks (the worse airgap, the worse adv-example works)



Conclusion

Codecs:

- We find that performance deteriorates on codec'ed audio data
- Easy fix: Add codecs to train (some computational overhead)



Next steps

Next steps: simulate a replay attack more faithfully, and use for data augmentation

- RIR ✅
- Loudspeaker transfer function 🟡
- Microphone transfer function 🟡
- Air-path attenuation: distance-dependent loss; high-frequency rolloff 🟡
- Room absorption & reflection characteristics: materials, surfaces, furniture 🟡
- Background noise model: environmental noise, mic self-noise 🟡

Goal:

- Simulate removal of air-gapped features; force the model to learn more robust ones
- Might even help the model generalize?

Conclusion

Summary:

- ReplayAttacks effectively break ML-driven audio deepfake detection
- Hypothesis: Airgap removes traces of "fakeness", e.g. characteristics of TTS models
- Some mitigation possible via Room Impulse Response data augmentation
- Next steps: Better replay attack simulation?

Contact

Fraunhofer AISEC

Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Dr. Nicolas Müller

Cognitive Security Technologies
nicolas.mueller@aisec.fraunhofer.de

Resources:

- deepfake-total.com/