BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023

Eva Trnovská
Supervisor: Mgr. Kamil Malinka, Ph.D.

2022/23

CAN YOU FOOL A DEEPFAKE DETECTOR?

Resistance of Deepfake Detectors to Modified Audio Recordings

MOTIVATION

Deepfakes present a powerful tool for influencing public opinion or bypassing security measures. Several deepfake detectors are publicly available. They tend to work well for recognizing spoofed recordings from the datasets they were trained on. However, even a slight modification can dramatically reduce their ability to distinguish between real and fake samples.

TESTED DETECTORS

Total number of 13 detectors were trained. This poster shows the results for 5 of them:

- LFCC SpecRNet [1]
- LFCC LCNN [1]
- STFT LCNN (image-based) [2]
- STFT CGNN (image-based) [2]
- STFT ResNet (image-based) [2]

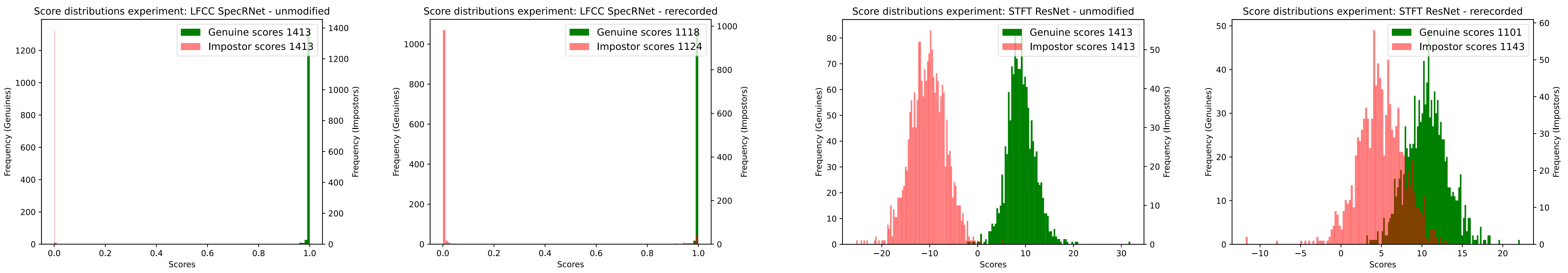


Figure 1: Distribution of scores predicted by the LFCC SpecRNet and STFT ResNet detectors. The 1st and the 3rd plot show the results for the unmodified validation set, the 2nd and the 4th for the rerecorded set.

DATASET AND MODIFICATIONS

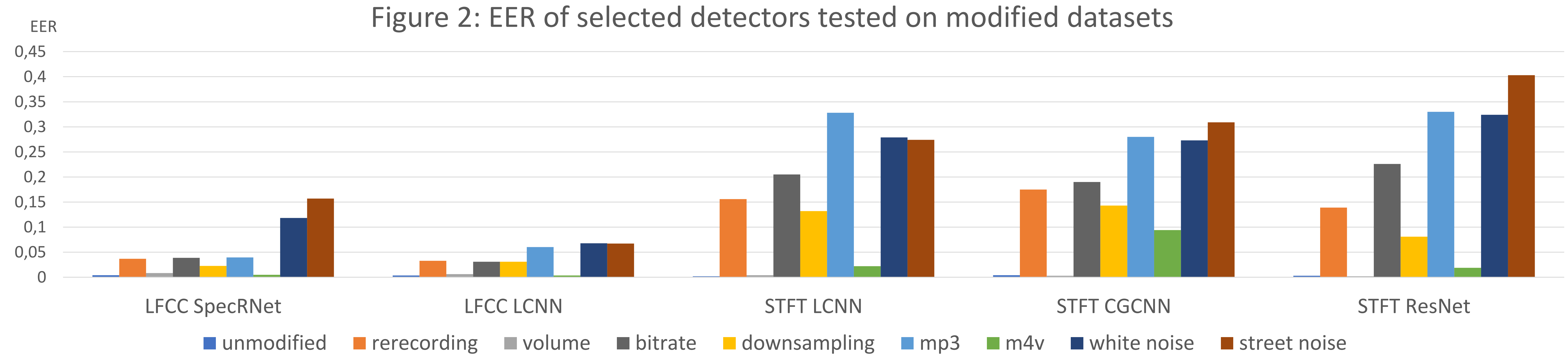
The normalised version of the Fake or Real dataset [3] was used for training the detectors. From 15 tested modifications, 8 were selected. The list of modifications includes:

- rerecording
- changing volume, bitrate, and sampling rate
- conversion to lossy formats and back
- adding artificial noise

METHODOLOGY

The detectors were trained in the testing environment and evaluated with the validation set of the normalised dataset and its modified versions. The evaluation metric is equal error rate (EER). The distribution graphs and EER were generated using Pyeer tool [4].

Figure 2: EER of selected detectors tested on modified datasets



RESULTS

Figure 2 shows that all modifications caused an increase in inaccurate classifications. Overall, the feature-based detectors were more efficient in evaluating modified recordings.

Adding artificial street and white noise fooled all tested detectors, while conversion to MP3 made specifically the image-based detectors unusable. Moreover, the image-based detectors showed a tendency to be deceived by changing bitrate and downsampling. Changing volume was the only modification that all detectors could easily pass.

REFERENCES

[1] Piotr Kawa, Marcin Plata, and Piotr Syga. Defense against adversarial attacks on audio deepfake detection, 2022

[2] Yexin Yang, Hongji Wang, Heinrich Dinkel, Zhengyang Chen, Shuai Wang, Yanmin Qian, and Kai Yu. The sjtu robust anti-spoofing system for the asvspoof 2019 challenge. Proc. Interspeech 2019, pages 1038–1042, 2019.

[3] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. Pages 1–10, 10 2019.

[4] Manuel Aguado Martinez. Pyeer, 2021.
https://github.com/manuelaguadomtz/pyeer