BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

# Resistance of Deepfake Detectors to Modified Audio Recordings

Eva Trnovská*

**Abstract**
Deepfakes present a powerful tool for influencing public opinion or bypassing security measures. Audio deepfake detectors are publicly available but in most cases, they were only tested on a small number of datasets. As such, the effects of even a small modification to the recordings were unknown. We took an existing dataset and tested the detectors on various modifications. The results have shown that most modifications significantly lower success rates across the detectors. Consequently, the experiment has demonstrated the need for more robust detection systems.

*xtrnov01@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

## 1. Introduction

Deepfakes present a powerful tool for influencing masses. Anyone with a basic knowledge of computers can create a synthetic voice or clone a real person's voice with tools available on the internet. Speaker verification is used as a security measure not only for unlocking phones but also for accessing bank accounts. These are just a few examples of why deepfake detection plays an important role in today's society.

Publicly available deepfake detectors are usually tested on a limited number of datasets (mostly one or two datasets). The testing does not acknowledge the possibility of artificial modifications that could possibly fool the deepfake detector.

We made a list of available deepfake detectors and selected a wide range of modifications. These deepfake detectors were trained on a normalized version of the Fake or Real dataset [1] and confronted with modified versions of the dataset. This enabled us to compare the reliability of the detectors in case of a range of attacks.

The results have shown that most deepfake detectors are immune to some modifications while extremely susceptible to being fooled by others. Testing different architectures have shown that some modifications pose threat to most systems. This information can be used for designing a new generation of robust deepfake detectors.

## 2. Tested detectors

The detectors used in this experiment were presented in two papers. A total number of 13 detectors were trained. Since some configurations and their results were very similar, this paper concentrates on 5 of them:

- LFCC SpecRNet [2]
- LFCC LCNN [2]
- STFT LCNN [3]
- STFT CGNN [3]
- STFT ResNet [3]

The detectors introduced by Kawa et al. [2] are trained using LFCC and MFCC as feature extraction methods. Yang et al. [3] chose an image-based approach; the recordings are converted to STFT-spectrograms and trained on the images.

## 3. Dataset and modifications

The normalized version of the Fake or Real dataset [1] was used for training the detectors. The dataset contains 17,870 normalized utterances cropped to 2 seconds.

The detectors were then confronted with a wide range of modifications to this dataset. The goal is to determine which have the most potential to fool audio deepfake detectors. The list of modifications includes:

- rerecording
- changing volume, bitrate, and sampling rate
- conversion to lossy formats (MP3 and M4V) and back
- adding white or street noise

## 4. Methodology

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

## 5. Experiment results

As seen on Figure 1, classifications by feature-based detectors were very accurate, almost binary. Conversely, image-based detectors provided a continuous evaluation range. The EER of all detectors tested on the original validation set did not exceed 1%.

### 5.1 Modified recordings

Figure 2 shows that all modifications caused an increase in inaccurate classifications. Overall, the feature-based detectors proposed by Kawa et al. [2] were more efficient in evaluating modified recordings.

Rerecording resulted in an average of 15% EER of image-based detectors and 3% EER of the rest. Conversion to MP3 made specifically the image-based detectors unusable.

Adding artificial street and white noise fooled all tested detectors to different degrees, the least prone to failure being the LFCC LCNN model. 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.

### 5.2 EER threshold

Figure 1 suggests that in some cases, the modifications can significantly raise the EER threshold. Without considering the possibility of manipulated recordings, spoofed recordings could be assessed as genuine using the original threshold.

## 6. Impact

This experiment demonstrates the need for more robust audio deepfake detection, as the detectors currently available are susceptible to deception by easily reproduced modifications, such as rerecording or conversion to MP3.

## Acknowledgements

I would like to thank my supervisor Kamil Malinka for his help and Anton Firc for provided materials and additional information.

Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

## References

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

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

[3] 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.

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