# Developing Robust Audio Deepfake Detector

Emanuel Mendiola-Ortiz[1,2], Zirui Zhang[1], Chengzhi Mao[1] Junfeng Yang[1]
[1]Columbia University, [2]Penn State University

**Introduction:** The proliferation of deepfake technology has raised significant concerns about the authenticity of audio content, necessitating robust detection mechanisms. Our research focuses on generating high-quality audio deepfakes and developing effective detection models. With the hopes it will help counterattack adversaries using it for bad intentions. This study aims to develop a real-world deepfake audio detector.

**Methods:** To collect the deepfake audio our team used open source deepfake and commercial deepfake generation sources (ElevenLabs, Soundboard101, Genny, Resemble, PlayHT and LipSynthesis), and generated audio from published datasets (WaveFake, ASVSpoof2019). We collected audio samples from online videos and imported published datasets (Narration, VCTK, In the Wild, VoxCeleb1, VoxCeleb2, ASVSpoof2021).

I utilized VokanTTS, an advanced fine-tuned StyleTTS2 model, renowned for its authentic and expressive zero-shot performance. VokanTTS is trained on a combination of AniSpeech, VCTK, and LibriTTS-R datasets, ensuring naturalness across various accents and contexts. The model generated deepfake audio from the following datasets bonafide, VCTKS and LibriSpeech ASR Corpus. Moreover, it was also used to generate deepfake audio from popular public figures and celebrities. With all the deepfake generation from the model and real audio extraction we combined them together to create a custom training dataset of 1.2 million data points with 50% real and 50% fake.

For detection I fully designed and developed an LSTM-based model for detecting deepfake audio, leveraging its ability to capture temporal dependencies in audio data. We started by converting raw audio files into Constant-Q Transform spectrograms, effectively capturing the changes in the distribution and changes in the frequencies present in the audio signal over a period. These deepfake audios often exhibits indistinct anomalies in the frequency patterns. The model contains multiple layers to process these frequencies, and dropout layers to prevent overfitting. Training is conducted by using binary cross entropy-loss and Adam optimizer.

Replicating the methodology from “Does Audio Deepfake Detection Generalize?” By Muller et al. We trained the model on the “In the Wild” dataset tested on the ASVSpoof2019 dataset and outputted a classification report showcasing our results and seeing the robustness before we test on our large custom dataset. We trained the model on “In the Wild” dataset and tested on the ASVSpoof2019 dataset and outputted a classification report showcasing our results.

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Bonafide** | 0.00 | 0.00 | 0.00 | 1031 |
| **Spoof** | 0.90 | 1.0 | 0.95 | 9014 |
| **Accuracy** | - | - | 0.90 | 10045 |
| **Macro AVG** | 0.45 | 0.5 | 0.47 | 10045 |
| **Weighted AVG** | 0.81 | 0.90 | 0.85 | 10045 |

**Figure 1**. LSTM-CQT-spec test model results

**Results:** Our results show an accuracy of 89.74%, a precision of 81%, a recall of 90% and an f1-score of 85%. The detection model achieved comparable performance metrics, validating its robustness and generalization capability. It also shows that by training and tuning the detection model on our diverse dataset it will make it more robust for out-of-distribution data.

**Future Work:** We will train and test the model on our newly created custom dataset and evaluates the results. The overall study promotes further development of the audio detector model and fine-tuning it to improve the accuracy.

**Acknowledgements:**
I would like to thank the Software Systems Lab Department of Computer Science at Columbia University for giving me the opportunity and space to conduct this work and in particular the principal investigator Junfeng Yang and mentors Zirui Zhang and Chengzhi Mao for their time, help and patience.

**References:**

N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, “Does Audio Deepfake Detection Generalize?,” arXiv preprint arXiv:2203.16263, 2022. [Online]. Available: https://arxiv.org/abs/2203.16263