Differential-based deepfake speech detection

Ing. Vojtěch Staněk | Supervisor: Ing. Anton Firc

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

1. Motivation

Deepfake speech technology can create highly realistic audio usable for entertainment purposes, but also *scams and criminal activities*. Moreover, it can be used to *undermine the reliability of legal media evidence* crucial in many investigations - Police or defendants might then have to prove if the recorded speech is real or fake before the court may accept it as a proof.

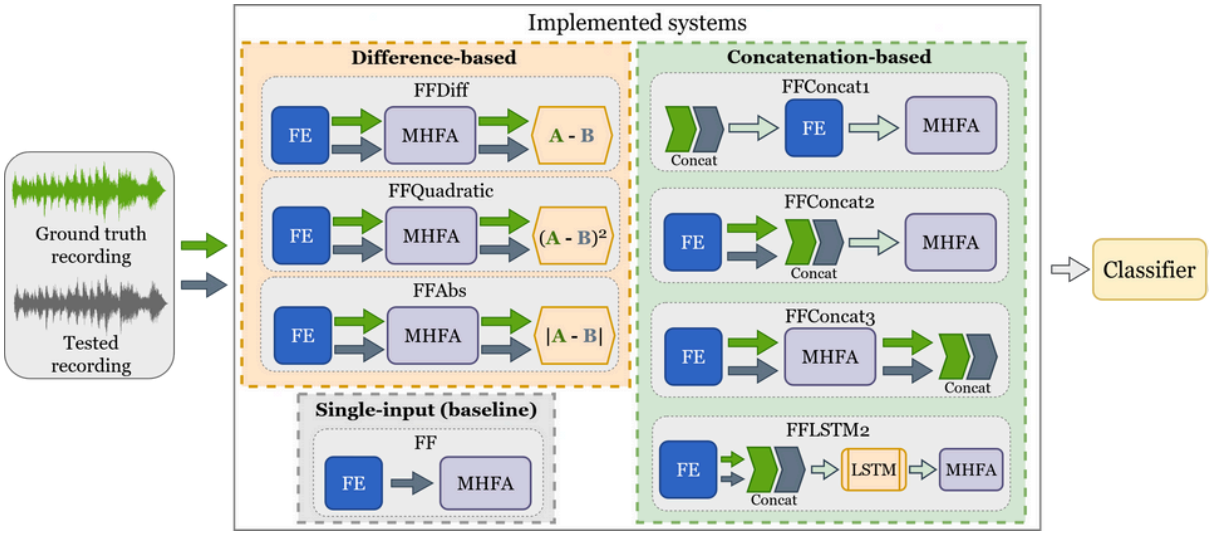
In such cases, one of the available deepfake speech detectors can be used to determine authenticity of the speech. However, these tools generally analyze recordings in isolation, with little awareness of the speaker's actual voice, as they lack direct comparison to verified reference samples of *how the person truly sounds*.

This poses a question: **could the inclusion of a trusted real recording help the detector more reliably and accurately detect real and fake speech?** A real recording should be easily obtainable in such cases, i.e. by recording the questioning at a police station or court hearing.

2. Our approach: Design & Implementation

We follow a standard system design in the field of deepfake speech detectors and enhance it with a reference to a trusted real recording of the speaker. Firstly, we extract defining features (FE) of the recordings. In our experiments, we used the *Self-supervised learning* model *XLSR-300M* [1] based on the *Wav2Vec 2.0* [2] architecture due to its superior performance and wide language support. Secondly, we use the *Multi-head factorized attentive pooling (MHFA)* [3] that contextually processes the extracted features to obtain maximum information possible. Finally, the extracted and pooled features are fed from MHFA to a *simple neural classifier* which decides if the tested recording is real or fake.

We explored two main strategies of combining the feature vectors as presented in the diagram below. **Difference-based** approach uses a metric to determine the difference between the two recordings. On the other hand, **concatenation-based** strategy utilizes concatenating the feature vectors at various places in the processing pipeline. We also implemented a single-input baseline system behaving just like other state of the art systems for result comparison. Additionally, we experimented with various system fusions to further enhance the detector's performance.



The models are implemented in the PyTorch framework, all code is freely available in the project's GitHub repository.

3. Results

We train and evaluate the detectors on two datasets commonly used for comparing deepfake detectors from the well-known *ASVspoof challenges* [4, 5]. To asses their reliability we also evaluate on *In-the-Wild* dataset [6] which contains real-world deepfakes collected from the Internet, mainly YouTube and social sites.

Equal Error Rate (EER)	Best to date	Implemented
ASVspoof2019 [4]	5.74% (single T04) 0.22% (ens. T05)	**0.18% (FF)** **0.22% (FFConcat3)**
ASVspoof2021 [5]	15.64% (T23) 2.85% (W2V2 + GAT)	**5.31% (FFLSTM2)** **2.92% (ensemble)**
In-the-Wild [6]	33.94% (RawNet2)	**12.28% (FFLSTM2)** **9.02% (ensemble)**

Equal Error Rate (EER)	Voice Conversion	Text to Speech
Single input	6.99% (FF)	**1.22% (FF)**
Difference-based	6.15% (FFDiff)	1.38% (FFQuadratic)
Concat-based	**4.79% (FFLSTM2)**	1.29% (FFConcat3)

4. Key contributions

- We show that **providing a trusted real recording enhances deepfake detectors**
- We **exceed state of the art performance**, especially with our fusion models
- **Our models perform better on unseen data** when compared to single-input detectors
  - More robust to overfitting
  - Very good performance on real-world deepfakes
- Up to date, *text-to-speech* and *voice conversion* were considered equal for detection, however **our results indicate that some methods are better at detecting specific types of deepfakes**
  - Single-input detector better at detecting Text-to-speech fakes
  - Difference-based and Concatenation-based detectors superior in detecting *Voice Conversion* deepfakes
- Combination of difference-based and concatenation-based models yields the best results

5. References

[1] A. Babu et al., XLS-R: self-supervised cross-lingual speech representation learning at scale. CoRR, abs/2111.09296, 2021
[2] A. Baevski et al., Wav2vec 2.0: a framework for self-supervised learning of speech representations. 2020. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20).
[3] J. Peng et al., An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification, 2022.
[4] Todisco et al., ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection, Interspeech 2019
[5] Xuechen Liu et al., ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild, 2023, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31(1):2507–2522
[6] Nicolas M. Müller et al., Does audio deepfake detection generalize?, Interspeech 2022