BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Differential-based Deepfake Speech Detection

Vojtěch Staněk*

Abstract

Deepfake speech technology, which can create highly realistic fake audio, poses significant challenges, from enabling multi-million dollar scams to complicating legal evidence's reliability. This work introduces a novel method for detecting such deepfakes by leveraging bonafide speech samples. Unlike previous strategies, the approach uses verified ground truth speech samples to identify spoofs, providing critical information that common methods lack. By comparing the bonafide samples with potentially manipulated ones, the aim is to effectively and reliably determine the authenticity of the speech. Results suggest that this innovative approach could be a valuable tool in identifying deepfake speech, especially recordings created using Voice Conversion techniques, offering a new line of defence against this emerging threat.

*xstane45@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

In the evolving landscape of digital forensics, the advent of deepfake speech technology poses unprecedented challenges. On the one hand, it offers innovative advancements in the fields of entertainment and education, on the other hand, its malicious applications have drawn significant concern – financial scams, privacy invasion or spreading of misinformation. [1]

The traditional approach to combat deepfakes may involve deploying state-of-the-art deepfake speech detectors. While effective, these tools typically operate by analyzing isolated input samples without the benefit of direct comparison to verified sources. This limitation poses a question: could the inclusion of reference or ground-truth speech samples as a basis for comparison enhance the accuracy and reliability of deepfake detection? This work presents the first differential-based deepfake speech detector, which incorporates trusted ground-truth speech samples to identify spoofs, providing critical information that common methods lack. A trusted sample can be easily obtained – such cases might include biometric check at an airport or police questioning [2].

The detection generally happens in three stages: *feature extraction*, *model training* and finally *classification* [3]. Popular features to describe speech are *cepstral coefficients*, such as MFCCs, which summarize the frequency content of sound signals. Modern systems utilize speaker embeddings extracted by a deep neural network [4]. The cutting-edge approaches employ novel techniques such as Self-Supervised Learning (SSL) with models like Wav2Vec2 to extract embeddings of the highest quality [5]. Regarding the classifiers, there are approaches utilizing classical (shallow) machine learning techniques such as GMM or SVM [6], as well as neural architectures, which are becoming more and more prevalent [7, 8].

This work utilizes both ground truth and tested recordings with the aim of finding the difference between them. The input recordings are first transformed into features using an SSL frontend: Wav2Vec2 [9] with Multi-head Factorized Attentive Pooling (MHFA) [10]. Next, these feature vectors are combined into a single feature vector, representing the difference between the two inputs, which is then fed into a feed-forward network for final classification.

Implemented systems successfully transfer the differential analysis based detection from facial to speech domain. Explored is the feasibility of multiple differential metrics as well as concatenation-based methods. Although the results indicate a need for refined parameterization, the foundational premise of utilizing differential analysis in deepfake speech detection showcases the potential for significant advancements. One of the major discovered benefits of differential deepfake speech detection is its superior ability to detect Voice Conversion (VC) samples.

2. Design and Implementation

As visible in the system overview on the poster, there are multiple systems designed and implemented:

1. **Difference-based** models
   (a) *FFDiff* using ordinary subtraction of extracted and pooled features
   (b) *FFQuadratic* uses squared subtraction
   (c) *FFAbs* uses absolute subtraction
2. **Concatenation-based** models
   (a) *FFConcat1* concatenates the recordings before feature extraction
   (b) *FFConcat3* concatenates extracted and pooled features
   (c) *FFLSTM* concatenates the pooled features and passes them to two LSTM cells
3. **Baseline** single input *FF* model for comparison

The main idea is that Difference- and Concatenation-based models take pairs of tested and ground-truth recordings as an input, while the *FF* model takes only a single tested recording as an input, operating the same way as classical detectors.

The feature extractor (FE) module is a pre-trained XLSR-300M model [9] based on the Wav2Vec2 architecture. The FE processes recordings by 50ms frames and extracts a 1024-value feature vector for all of them. To enable the processing of varied-length recordings, the features extracted from all 24 transformer layers of XLS-R were consequently pooled using Multi-head Factorized Attentive Pooling [10].

The resulting pooled feature vector was fed to a downstream feed-forward classification neural network with a simple architecture of three linear layers with batch normalization and ReLU activation function between them. Softmax function is applied to the resulting values to obtain classification probabilities for bonafide and spoofed classes.

3. Results

Equal Error Rate (EER) is used as the primary evaluation metric, as it's threshold-independent and a standard for evaluation of deepfake speech detection methods. Moreover, it provides reliable information about how well the detector can separate bonafide and spoof classes, i.e., the lower the EER, the better the detector.

As apparent from the result tables, implemented systems supersede other systems from the ASVspoof challenges [7, 8] or In-the-Wild [11] benchmark. However, it is necessary to declare that the proposed systems could not be submitted to the ASVspoof challenges, because the FE module was pre-trained on multiple datasets, which violates the challenge rules.

The models annotated by *ens.* or *ensemble* mark systems, which utilize a fusion of the resulting scores of individual systems to further boost performance. The motivation behind is that upon close examination of the score distributions, differential-based techniques performed better in accurately identifying bonafide samples, while concatenation-based systems demonstrated a heightened ability to distinguish spoofed samples.

The best performing system for the ASVspoof2021 challenge is a mean fusion of *FFConcat1*, *FFConcat3* and *FFQuadratic*. Similarly, for In-the-Wild dataset, a square root fusion of all individual models except *FFLSTM* yielded the best results.

Results of ASVspoof2021 were further explored – a significant portion of the dataset consists of Voice Conversion (VC) samples. The investigation uncovered that pair-input systems perform significantly better than single input baseline, which in contrast performed the best on Text-to-Speech samples. The hypothesis of this behavior lies in VC samples containing leaked speaker information from both the source and target utterances [12].

4. Conclusions

Performed research, implementation, and evaluation discovered several interesting points. Firstly, it shows that differential-based detection is a feasible approach for deepfake speech detection. Secondly, the pair-based models seem more robust to overfitting – there is a slight tradeoff between the performance on known data and robustness on unseen data. This tradeoff makes the pair-input models less efficient over the ASVspoof2019 (training) data, but in turn generalize better over ASVspoof2021 and In-the-Wild data. Finally, it is expected that the pair-based input will be extremely efficient on morphed speech, where the effect from VC is further amplified. This was however not possible to prove due to the lack of morphed speech data.

Acknowledgements

I would like to thank my supervisor, Ing. Anton Firc, for his wise guidance, consultations, reviews and all-around support.

References

[1] Anton Firc, Kamil Malinka, and Petr Hanáček. Deepfakes as a threat to a speaker and facial

recognition: An overview of tools and attack vectors. *Heliyon*, 9(4):e15090, 2023.

[2] C. Rathgeb, C.-I. Satnoianu, N. E. Haryanto, K. Bernardo, and C. Busch. Differential detection of facial retouching: A multi-biometric approach. *IEEE Access*, 8(1):106373–106385, 2020.

[3] Zaynab Almutairi and Hebah Elgibreen. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5), 2022.

[4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[5] Juan M. Martín-Doñas and Aitor Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9241–9245. ICASSP, 2022.

[6] Bhusan Chettri, Daniel Stoller, Veronica Morfi, Marco A. Martínez Ramírez, Emmanouil Benetos, and Bob L. Sturm. Ensemble models for spoofing detection in automatic speaker verification, 2019.

[7] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Hector Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection, 2019.

[8] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31(1):2507–2522, 2023.

[9] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296, 2021.

[10] Junyi Peng, Oldrich Plchot, Themos Stafylakis, Ladislav Mosner, Lukas Burget, and Jan Cernocky. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification, 2022.

[11] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize?, 2022.

[12] Jiangyi Deng, Yanjiao Chen, Yinan Zhong, Qianhao Miao, Xueluan Gong, and Wenyuan Xu. Catch you and i can: Revealing source voiceprint against voice conversion. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5163–5180, Anaheim, CA, August 2023. USENIX Association.