Creating More Meaningful Explanations for Audio Deepfake Detection

Bachelorarbeit

im Studiengang
Wirtschaftsinformatikim 6. Semester

vorgelegt von

Jaocb LaRock
Matr.-Nr.: 1667321
am 31.03.2025

Universität Siegen

Fakultät III: Wirtschaftswissenschaften, Wirtschaftsinformatik und Wirtschaftsrecht

Erstprüfer: Gunnar Stevens Gunnar.Stevens@uni-siegen.de

1 Abstract

Abstract filler text blablabla!

2 Introduction

As the use of generative methods for the creation of synthetic voices becomes more widespread, so grows the need for reliable and usable detection methods for the protection of the security of people and businesses alike. In particular, the rise of deepfake technology has led to concerns about its potential misuse in areas such as politics, entertainment, and national security by, for instance, potentially allowing malicious actors to create fake audio recordings that appear to be genuine statements made by public figures or to create genuine-seeming recordings of events that never happened, which could have significant consequences if used to spread misinformation, from defaming individuals to creating political tension [10, 1]. The field of Explainable AI (XAI) shows promise for producing useful and interpretable results from such models. Explainable AI refers to the ability of artificial intelligence systems, such as machine learning models and neural networks, to provide understandable and interpretable explanations for their decisions or predictions. This means that XAI systems can articulate why they made a particular decision, what factors influenced it, and how confident they are in their conclusion [4]. For the detection of audio deepfakes, this could mean that the model can justify its classification with feature-based evidence, allowing for verification of the result by the end user. Much of the existing research, further discussed in the next section, follows the path of producing a model with the best possible benchmarks against datasets of samples, without taking into account if the results can be usefully interpreted, while existing explorations into making audio deepfake detection explainable have produced results that often require a lot of background knowledge to meaningfully interpret With the context of the research discussed in the next section, I would like to use new combinations of features with a novel

model architecture to generate explanations for fake or real audio classification that are potentially more useful to their end user.

3 Related Work

4 Experiments

4.1 Dataset

The dataset used for these experiments is the "In-the-Wild" dataset, one which focuses on the generalization of audio deepfake detection models by collecting real-world data, in comparison to other examples that use more controlled laboratory conditions [8].

4.2 Black-Box Model

The model used for this experiment was made, trained and evaluated using the tools available in the TensorFlow and Keras libraries [7]. This section will go into further detail about the experimental setup for the black-box model as well as the way in which explanations were created and summarized.

4.2.1 Features

Several features were used for the final version of the black-box model. Some are features that can be called generic, widely used in research involving the detection of synthetic while others are more specific and are cited accordingly. The features that are not summarized for a whole sample are extracted using a sliding window method, preventing a loss of fidelity that can be caused by compression of the feature to a standard size. The features used are as follows:

- Harmonic-to-noise ratios: Inspired by previous research [2, 3, 6] but slightly modified from the original concept, the HNRs are the ratios of the strengths of harmonic frequencies to the strengths of "noise", the to-

tal strength outside the harmonic frequencies. Unlike some previously seen examples [2, 3], I calculate this ratio for each fundamental frequency length, instead of using one value for the whole sample. Given that γ_i is the harmonic energy in a given fundamental frequency cycle and ι_i is the residual energy in a given fundamental frequency cycle, the HNR at that cycle is calculated as follows:

$$20log\frac{\gamma_i}{\iota_i}$$

- Mel-spectrograms: A widely-used feature in synthetic audio detection, the mel-spectrogram is a spectrographic representation of an audio sample transformed in a manner to better represent human perception of frequencies [9].
- Mel-frequency cepstral coefficients: The derivative cepstral coefficients of the mel-spectrogram, also widely-seen in research on synthetic audio detection.
- Fundamental frequency lengths: The lengths of every fundamental frequency cycle in the sample. This has also been previously used in certain research on synthetic audio detection [11]. The output is one-dimensional with time as the axis.
- Onset strength: As used in previous research [6], the strengths of each onset in the audio sample, where an onset is point where there is a sudden rise in energy across the audio spectrum. This results in a one-dimensional output with time as the axis.
- Intensity: Also inspired by previoous research [6], intensity is the total power at each point in the audio sample, given in db. This is calculated by creating a fourrier transformation of the sample and then summing across the frequency-axis for every point on the time-axis. Resulting in a one-dimensional output.
- Pitch-fluctuations: Calculated for every sample in the audio as the dif-

ference between pitch at the given point and the pitch at the previous point. Pitch is estimated based on the maximum power harmonics. This feature is not identical to, but is inspired by the use of summarized pitch fluctuations in previous research [5]. This feature is also one-dimensional with time as the axis. Given that H_i is the set of harmonic frequencies at fundamental frequency cycle i and $s(h_i)$ is the power of a given harmonic frequency at cycle i, the pitch can be estimated as follows:

$$p_i = s(max(H_i))$$

Then, given an offset x, the pitch fluctuations can be calculated as follows:

$$p_i - p_{i-x}$$

- Jitter features: As defined in previous attempts to identify synthetic audio [3] with perceptible features, jitter-based features measures the absolute variations in fundamental frequency cycles in comparison to the nearest x neighbors using the following method:

$$\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i(\frac{1}{x}\sum_{n=i-m}^{i+m}T_n)|}{\frac{1}{N}\sum_{i=1}^{N}T_i}$$

  where T_i represents the extracted fundamental frequency lengths and N is the number of fundamental frequency periods.

- Shimmer features: Similarly defined to the jitter features, shimmer features instead use the amplitudes at each fundamental frequency period. Their purpose is to capture irregular vocal fold vibrations which may be an indication but not a guarantee of a synthetic voice [3].

$$\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i(\frac{1}{x}\sum_{n=i-m}^{i+m}A_n)|}{\frac{1}{N}\sum_{i=1}^{N}A_i}$$

4.2.2 Model Architecture

The general architecture is designed to allow the combination of diverse features into one final prediction, regardless of whether the individual features have the

same shape. This is achieved by using a separate model for each feature, with its own input and output layers (further referred to as sub-models), which are then concatenated together and pooled into a final output value (in what I will further refer to as the terminus). This modular structure allows not only better performance and less memory use, due to the fact that fewer transformations are required in input preparation, but also allows for more flexibility in the construction of the model, making it easy to add and remove features using pre-defined functions depending on feature type. This kind of architecture also provides an advantage to the generation of explanations, as only the terminus part of the entire must be explained to interpret the importance of each input feature. Further explanation is in the following section. Figure 1 represents the layout of the model layers.

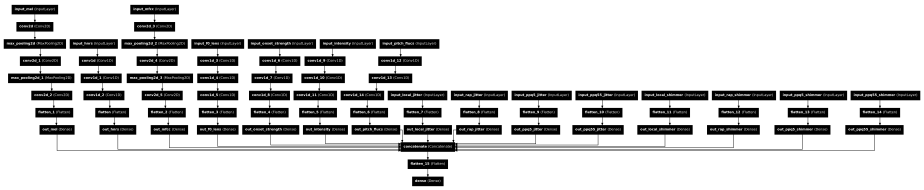

Figure 1: The architecture of the model used for the experiment.

4.2.3 Training and Evaluation

The black-box model described in the previous section was trained on the first 10,000 samples out of the aforementioned dataset, with evaluation being performed on the remaining 21,778. Training batches consisted of 100 samples each with every window position for each sample and an upper limit of 1,000,000 lines per batch. For each batch, two epochs were performed. Because a single sample can have multiple input lines, evaluation cannot be performed on a per-line basis, but must instead be summarized. Two methods are possible: median and mean, with the difference in results being negligible in my experiments. The default threshold is 0.5, meaning that a result under 0.5 is a false result and over 0.5 is a true result. This threshold can however be adjusted, useful for

computing the EER.

4.3 Explanations

5 Results

5.1 Black-Box Model Results

5.2 Explanation Results

6 Discussion

7 Conclusion

References

[1] Marwan Albahar and Jameel Almalki. "DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW". In: . *Vol.* 22 (2005).

[2] Anuwat Chaiwongyen et al. "Contribution of Timbre and Shimmer Features to Deepfake Speech Detection". In: *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2022 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Chiang Mai, Thailand: IEEE, Nov. 7, 2022, pp. 97–103. ISBN: 978-616-590-477-3. DOI: 10.23919/APSIPAASC55919.2022.9980281. URL: https://ieeexplore.ieee.org/document/9980281/ (visited on 05/28/2024).

[3] Anuwat Chaiwongyen et al. "Deepfake-speech Detection with Pathological Features and Multilayer Perceptron Neural Network". In: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Taipei, Taiwan: IEEE, Oct. 31, 2023, pp. 2182–2188. ISBN: 9798350300673. DOI: 10.1109/APSIPAASC58517.2023.10317331. URL: https://ieeexplore.ieee.org/document/10317331/ (visited on 05/28/2024).

[4] Michael Hind. "Explaining explainable AI". In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (Apr. 10, 2019), pp. 16–19. ISSN: 1528-4972, 1528-4980. DOI: 10.1145/3313096. URL: https://dl.acm.org/doi/10.1145/3313096 (visited on 06/05/2024).

[5] Zahra Khanjani et al. "Learning to Listen and Listening to Learn: Spoofed Audio Detection Through Linguistic Data Augmentation". In: *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*. 2023 IEEE International Conference on Intelligence and Security Informatics (ISI). Charlotte, NC, USA: IEEE, Oct. 2, 2023, pp. 01–06. ISBN: 9798350337730. DOI: 10.1109/ISI58743.2023.10297267. URL: https://ieeexplore.ieee.org/document/10297267/ (visited on 05/28/2024).

[6] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. "A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection". In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. MM '22: The 30th ACM International Conference on Multimedia. Lisboa Portugal: ACM, Oct. 14, 2022, pp. 35–41. ISBN: 978-1-4503-9496-3. DOI: 10.1145/3552466.3556523. URL: https://dl.acm.org/doi/10.1145/3552466.3556523 (visited on 05/11/2024).

[7] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[8] Nicolas M. Müller et al. *Does Audio Deepfake Detection Generalize?* Apr. 21, 2022. arXiv: 2203.16263[cs,eess]. URL: http://arxiv.org/abs/2203.16263 (visited on 07/04/2024).

[9] Abu Qais et al. "Deepfake Audio Detection with Neural Networks Using Audio Features". In: *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*. 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP). Hyderabad, India: IEEE, July 21, 2022, pp. 1–6. ISBN: 978-1-66547-258-6. DOI: 10.1109/ICICCSP53532.2022.9862519. URL: https://ieeexplore.ieee.org/document/9862519/ (visited on 05/28/2024).

[10] Namosha Veerasamy and Heloise Pieterse. "Rising Above Misinformation and Deepfakes". In: *International Conference on Cyber Warfare and Security* 17.1 (Mar. 2, 2022), pp. 340–348. ISSN: 2048-9889, 2048-9870. DOI: 10.34190/iccws.17.1.25. URL: https://papers.academic-conferences.org/index.php/iccws/article/view/25 (visited on 09/23/2024).

[11] Jun Xue et al. "Audio Deepfake Detection Based on a Combination of F0 Information and Real Plus Imaginary Spectrogram Features". In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. MM '22: The 30th ACM International Conference on Multi-

media. Lisboa Portugal: ACM, Oct. 14, 2022, pp. 19–26. ISBN: 978-1-4503-9496-3. DOI: 10.1145/3552466.3556526. URL: https://dl.acm.org/doi/10.1145/3552466.3556526 (visited on 05/11/2024).