Deepfake Audio Detection using WavLM, Wav2Vec2 and Whisper Model

Prateek Korat, Avinash Saxena, Gauri Naik

December 17, 2024

Abstract

The following research work proposes an enhanced deepfake audio detection system, powered by the power of several state-of-the-art pre-trained models: WavLM, Whisper, and Wav2Vec2. The proposed system was trained and evaluated on the ASVspoof 2019 LA dataset, which has been carefully resampled into equal numbers of spoofed and bonafide audio recordings for training, validation, and test splits. This project aims to accurately identify genuine and artificially generated audio with a high degree of accuracy by fine-tuning such complex models on a judiciously selected dataset. A relative study of the three models will shed enough light on their efficiency in detecting Deepfake Audio, which is vital in the various strands of ongoing efforts meant for solving the ever-growing dangers from audio deepfakes toward the security and authentication frameworks in various applications.

1 Introduction

Deepfake is a category of digitally produced material where computer-generated faces or speech have been used in place of the real human faces in a picture, video, or recording [1, 2]. The rapid advancement of artificial intelligence and deep learning technologies has led to the emergence of increasingly sophisticated audio deepfakes, posing significant challenges to security, authentication, and trust in digital communications. Deepfake audio, which involves the artificial manipulation or generation of human speech, has become a growing concern due to its potential for misuse in various domains, including identity fraud, misinformation campaigns, and social engineering attacks. As the technology behind deepfakes continues to evolve, it is crucial for detection methods to keep pace, ensuring the integrity and trustworthiness of digital audio content.

This project focuses on developing an advanced deepfake audio detection system by leveraging state-of-the-art pretrained models, including WavLM [3], Whisper [4], and Wav2Vec2 [5]. These models have demonstrated remarkable performance in various speech processing tasks and offer promising potential for deepfake detection. By fine-tuning these models on the ASVspoof 2019 Logical Access (LA) dataset, we aim to achieve high accuracy in distinguishing between authentic and synthetically generated audio. Our approach involves carefully resampling the dataset to ensure balanced representation of both spoof and bonafide classes across training, validation, and test sets, mitigating potential biases and enhancing the generalizability of our models. General overview of the proposed system is shown in Fig.1.

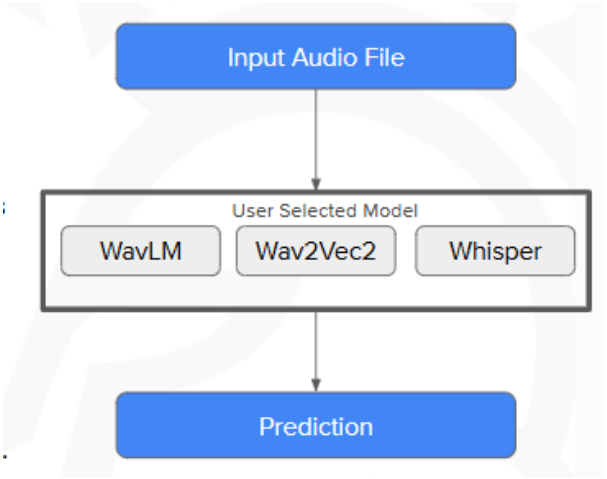


Figure 1: Overview of the System

2 Related Work

Deepfake audio detection lately has turned out to be a domain that gained huge momentum due to the thrusting threat created by synthetic media. Researchers leveraged several state-of-the-art speech processing models and novel architecture improvements. This has been one of the most promising approaches based on the use of self-supervised speech representations with WavLM and Wav2Vec2.

A recent work by Chen et al. [6] compares WavLM-based features against other pre-trained representations for deepfake detection. The authors have proposed an MFA classifier which captures complementary information across time and feature layers and yield state-of-the-art results on the ASVspoof 2021 dataset. Another exciting possible front-end for deepfake detection is Whisper, an automatic speech recognition model. Whisper-based features have been shown in [7] to further improve performance, lowering the Equal Error Rate by 21% on an in-the-wild dataset compared to prior work operating on multiple model architectures.

Ensemble methods have also been found to increase the detection rate significantly. Combined spectrogram-based features were extracted by [8] through an ensemble of deep learning architectures: CNN, RNN, and transfer learning techniques from computer vision models. It turned out that the multi-model approach showed strong performance under various audio transformations.

Despite these advances, a number of challenges persist in deepfake audio detection. One of the major challenges is generalization to unseen deepfake generators. Recent benchmarks have shown that state-of-the-art detectors often struggle with out-of-distribution content generated by novel methods [9]. Another challenge is that the detectors shall perform well irrespective of the quality of the audio, background noise, or artifacts introduced by compression [10]. As deepfake methods are extensively increasing in other domains such as image, video, and music, the detection of deepfakes requires serious attention to both audio and visual clues [11].

Other future research in this direction can be done by developing more robust feature representations, studying self-supervised learning methods specifically tailored for deepfake detection, and curating larger, more diverse datasets representative of evolving dynamics within the landscape of synthetic media generation techniques [9, 11].

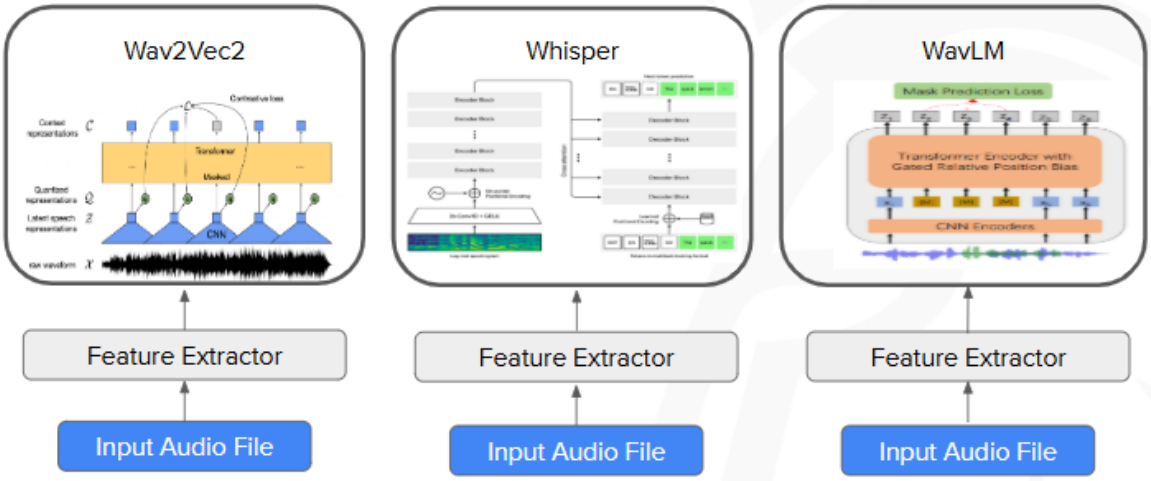


Figure 2: Proposed Solution

3 Methodology

Three advanced pretrained models were used for the detection of deepfake audio: Whisper, WavLM, and Wav2Vec2. Consequently, the fine-tuning of these models using the ASVspoof 2019 dataset resulted in the development of strong detectors that could efficiently tell between authentic and synthetic audio samples.

3.1 Dataset

We used ASVspoof 2019 dataset [12], whose speech samples are in the English language and approximately 6 seconds in length. We resampled the dataset to have an equal number of samples between the spoof and bonafide samples:

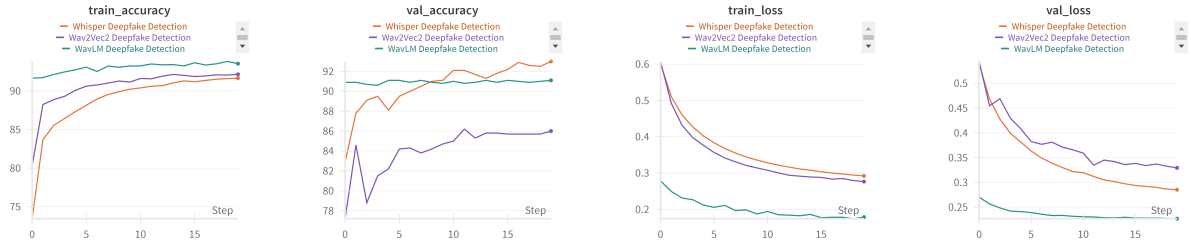
- Training set: 5,160 samples
- Validation set: 1,000 samples
- Test set: 2,000 samples

3.2 Model

We have used a multiple model approach, allowing users to choose from different pre-trained architectures for deepfake prediction. A brief overview of proposed solution is shown in Fig.2. We have performed fine-tuning on pre-trained Whisper, WavLM and Wav2Vec2 models, a brief summary of these models is presented below:

Whisper is a strong speech model that was trained on diverse multilingual audio data. It is transformer-based in encoder-decoder architecture and has self-supervised pretraining [4]. The key innovation is its ability to perform various speech processing tasks without fine-tuning, leveraging large-scale pre-training on weakly labeled audio data.

WavLM is based on the transformer-based architecture, which integrates masked language modeling and contrastive learning objectives [3]. Among its salient features, WavLM captures both local and global contexts in speech. Unlike the others, WavLM implements a multi-scale modeling strategy in order to obtain the representations from different levels of temporal resolution.



(a) Train Accuracy Curve (b) Validation Accuracy Curve (c) Training Loss Curve (d) Validation Loss Curve

Figure 3: Accuracy and Loss Curves

Wav2Vec2 uses a multilayer convolutional neural network to encode raw audio into latent audio representations, followed by a transformer network, which models local and global audio contexts. This involves a contrastive training objective, which discriminates between the true future audio sample and the distractors [5] . Wav2Vec2's innovation lies in its ability to learn powerful speech representations from unlabeled audio data.

4 Experiment

The mentioned three models were trained and tested on the same data under the same environment to avoid any bias in comparison. The key experimental settings are mentioned below:

- Loading and pre-processing data using custom DataLoader classes.
- Model initialization with pre-trained weights.
- Fine-tune using binary cross-entropy loss with Adam optimizer.
- Leverage mixed-precision training to optimize GPU memory usage and accelerate the process.
- Performed training for 20 epochs on T4 GPU.
- Each model was fine-tuned on our prepared dataset, with the final layer modified to perform binary classification (spoof vs. bonafide audio).
- To assess model performance, we utilized Accuracy, Precision, Recall and F1-score

5 Result

The experimental results shown in Fig. 3 demonstrate the comparative performance of three deepfake detection models: WavLM, Wav2Vec2, and Whisper. WavLM consistently outperformed the other models, achieving the highest training accuracy of approximately 93% and maintaining stable validation accuracy around 91%.

The accuracy for Whisper increased slowly from its low value of about 92% to the values around those obtained. In general, the model was very smooth in terms of convergence: its best training accuracy was around 0.91, and a validation loss smoothly diverged from 0.5 down to around 0.29.

Wav2Vec2 had a strong start but was more volatile in its validation accuracy, ranging from 78% up to 86%. While it reached 91% accuracy during training, it stabilized at about 86% on validation, therefore being a bit more overfitted than the other models.

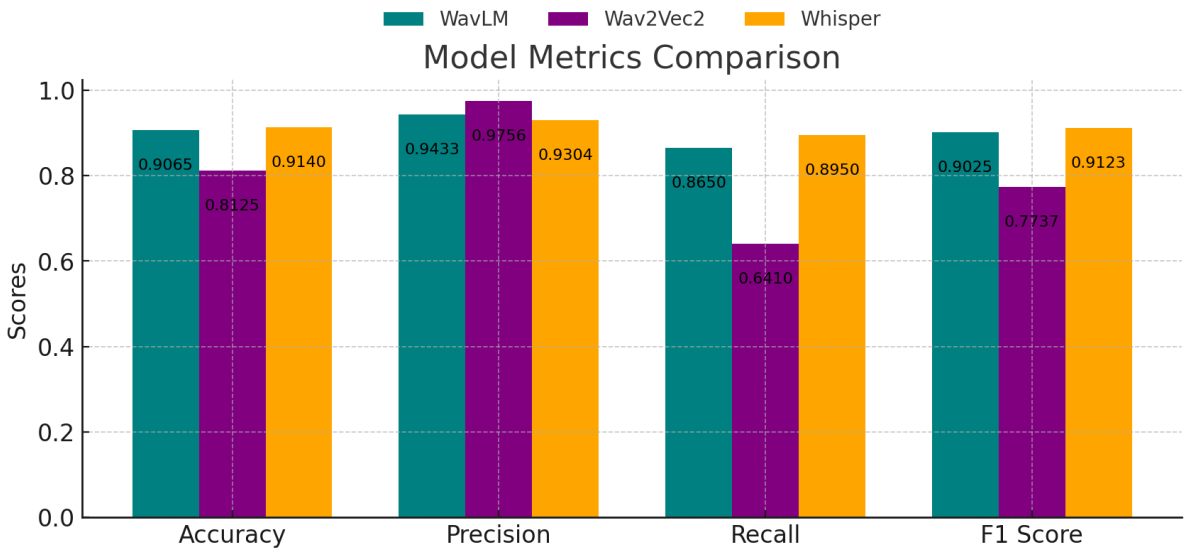


Figure 4: Model Metrics Comparison

Loss curves show that WavLM had the most stable training process, maintaining the lowest losses throughout training at about 0.2. On the other hand, Whisper and Wav2Vec2 started off high in the beginning with losses of about 0.6 and converged gradually towards 0.3. Still, Wav2Vec2 behaved a bit more unsteadily during the validation process.

The key metrics used for evaluation are shown in Fig.4. Where, WavLM consistently had higher accuracy, precision, recall, and F1 scores. Training and validation accuracy graphs showed stability for WavLM while Whisper plotted linear increases, whereas Wav2Vec2 was observed more volatile as compared to others.

6 Conclusion

In conclusion, this study compared the performance of three state-of-the-art models—WavLM, Wav2Vec2, and Whisper—for deepfake audio detection. The results obtained shows that, among the models tested, the WavLM model has constantly provided better performance along all the metrics of comparison-accuracy, precision, recall, and F1 score.

Lastly, WavLM provided the best stability during the training and validation process, always with high performance.Whisper shows a continuous improvement and has competitive finals, especially regarding validation accuracy, while Wav2Vec2, though it improves at a rapid rate in the beginning, has greater fluctuation in validation accuracy and possible overfitting.

These confirm that WavLM indeed works well for deepfake audio detection and hence is promising for reliable deployment in real-world applications. At the same time, however, the very strong performance of all three shows that one good approach to this important task is to use a pre-trained speech processing framework. Future work should be more directed at improving model generalization by considering ensembling strategies with a view toward realizing the maximum strengths of each model.

References

[1] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics," *IEEE Access*, vol. 10, pp. 38885–38894, 2022.

[2] A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, "A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions," *IEEE Access*, vol. 10, pp. 11065–11089, 2022.

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.

[5] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[6] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, "Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier," 2024.

[7] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," 2023.

[8] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," 2024.

[9] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea, P. Irofti, M. Popescu, C. Rusu, R. T. Ionescu, F. S. Khan, and M. Shah, "Deepfake media generation and detection in the generative ai era: A survey and outlook," 2024.

[10] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," 2022.

[11] Y. Li, M. Milling, L. Specia, and B. W. Schuller, "From audio deepfake detection to ai-generated music detection – a pathway and overview," 2024.

[12] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," 2020.