DEEPFAKE VOICE DETECTION USING ML

Chaitanya Alhat[1], Hemant Aher[2], Omkar Autade[3] ,Prof Manisha Mehrotra[4]

[1,2,3]BE Scholar,Dhole Patil College of Engineering, Pune, India, India.

[4]Professor, DPCOE, Savitribai Phule Pune University, Pune, Maharashtra, India.

DOI: https://www.doi.org/10.58257/IJPREMS36602

ABSTRACT

Recent advancements in AI-generated human voices have raised significant concerns regarding impersonation and the spread of disinformation, highlighting the need for effective methods to identify synthetic voices. This research presents a novel strategy for detecting synthetic human voices by identifying artifacts produced by vocoders in audio signals. Many DeepFake audio synthesis techniques utilize neural vocoders—neural networks designed to convert temporal-frequency representations, such as mel-spectrograms, into waveforms. By recognizing the processing effects of neural vocoders in audio samples, we can ascertain whether a voice is artificially created. To facilitate the detection of synthetic human voices, we propose a multi-task learning framework that employs a binary classification RawNet2 model, which integrates a vocoder identification module with a shared feature extractor. By framing vocoder identification as a pretext task, we guide the feature extractor to concentrate on the distinctive artifacts left by vocoders, thus enhancing the feature set available for the final binary classification. Our experimental results indicate that the enhanced RawNet2 model, which incorporates vocoder identification, demonstrates superior classification performance on the binary detection task.

Keywords: - Deep Learning , Rawnet2, Sincov, Vocoder Artifacts, PyTorch, TensorFlow.

1. INTRODUCTION

Artificial intelligence (AI) has dramatically reshaped various fields, with media synthesis—particularly deepfake audio—emerging as a technology with both promising applications and significant risks. Deepfake audio involves AI-generated voices that can convincingly mimic real individuals, providing benefits in personalized voice assistants, entertainment, and accessibility for speech-impaired individuals. However, this realism also introduces serious security threats. AI-synthesized voices have been used in scams, impersonations, and misinformation campaigns, with real-world incidents illustrating the potential harm from voice deepfakes. Unlike deepfake images or videos, which can be visually scrutinized, synthetic audio presents unique challenges due to its continuous, intricate structure, making it difficult to detect with conventional methods Our study presents a novel approach to synthetic voice detection using machine learning, specifically focusing on the artifacts introduced by neural vocoders—a type of neural network component crucial in transforming spectrograms into audible waveforms in text-to-speech (TTS) and voice conversion (VC) systems. Vocoder artifacts represent a unique fingerprint left by the synthesis process, and our approach leverages these artifacts within a multi-task learning framework to achieve high detection accuracy. This paper demonstrates that by including a vocoder identification module as a secondary task, the model can generalize across a range of vocoder architectures, addressing a critical gap in current detection capabilities..

2. PROBLEM STATEMENT

As voice synthesis technology advances, distinguishing between real and synthetic voices becomes increasingly difficult. Traditional deepfake audio detection methods, which often rely on statistical analysis or superficial audio features, struggle to detect synthetic audio produced by new and unfamiliar vocoder architectures. Additionally, the inherent nature of audio data complicates the detection process; audio signals lack the clear, structured patterns found in images or videos, making it difficult to apply visual-based deepfake detection techniques directly to audio.

This research targets the challenge of creating a detection model that can generalize across vocoder types by training it to recognize artifacts unique to the vocoding process itself. Our goal is to build a detection system that can identify synthetic audio based on these artifacts, irrespective of the specific vocoder architecture used, enhancing its adaptability to new synthesis technologies.

SCOPE AND OBJECTIVE

This study explores the detection of AI-generated voices, specifically those synthesized through neural vocoders, across various vocoder types and architectures. The primary objective is to develop a machine learning model that identifies synthetic audio by leveraging vocoder artifacts, a generalizable feature that persists across vocoder architectures. By focusing on vocoder-specific patterns, the model aims to enhance detection accuracy and ensure adaptability to future advancements in voice synthesis technology.

Potential applications of this detection system include secure voice-based authentication systems, media verification processes, and cybersecurity, where distinguishing real voices from synthetic ones is crucial. This approach is particularly valuable for protecting voice-driven services, such as customer support, secure communications, and automated transactions, from impersonation attacks. Furthermore, this research aims to contribute to the broader field of synthetic media detection, setting a foundation for future models that integrate multiple audio synthesis techniques

3. LITERATURE SURVEY

3.1 Human Voice Synthesis Technique

The field of human voice synthesis has advanced through two primary methods: Text-to-Speech (TTS) and Voice Conversion (VC).

Text-to-Speech (TTS): TTS systems transform textual input into spoken language, providing a critical backbone for applications like virtual assistants and accessibility tools. Modern TTS models involve three main stages: (1) Text Analysis, which converts text into linguistic features; (2) Acoustic Modeling, where these features are transformed into acoustic representations, typically mel-spectrograms; and (3) Vocoding, where neural vocoders convert the mel-spectrograms into audio waveforms. Deep neural networks, such as WaveNet, Tacotron, and FastSpeech, have driven significant progress in TTS, enabling high-quality speech synthesis with natural intonation and timing. Each TTS model architecture leaves distinct artifacts, especially in the final vocoding stage, due to the model's specific approach to waveform reconstruction.

Voice Conversion (VC): VC techniques modify one individual’s voice to sound like another, using style transfer techniques to replicate voice characteristics. VC is often achieved through neural networks that map features of the source voice to those of a target voice, and it plays an important role in applications like voice dubbing and voice cloning. VC models frequently use variational autoencoders (VAEs) and generative adversarial networks (GANs), which, despite high synthesis quality, leave subtle artifacts in the resulting audio. These artifacts, often resulting from the vocoding process, serve as distinguishing features that detection models can exploit.

3.2 Neural Vocoder and Artifact Formation

Neural vocoders are essential for transforming intermediate audio representations into time-domain waveforms in both TTS and VC models. There are three main categories of neural vocoders, each of which introduces distinct synthesis artifacts:

Autoregressive Models: These models, such as WaveNet and WaveRNN, generate each sample based on previous samples, resulting in high-fidelity but slower synthesis. Autoregressive vocoders create minor amplitude and phase artifacts due to the sequential dependency of samples. These artifacts are often subtle but detectable by well-designed classifiers.

Diffusion Models: DiffWave and WaveGrad, examples of diffusion-based vocoders, generate audio through an iterative noise removal process. While diffusion models are efficient, they sometimes leave artifacts in the higher frequencies, as the noise-removal process can introduce signal degradation. These high-frequency irregularities can be used as indicators of synthetic origin.

GAN-based Models: GAN-based vocoders, including Mel-GAN and Parallel WaveGAN, utilize adversarial networks for rapid synthesis. GAN vocoders often prioritize speed over precision, leading to spectral inconsistencies, particularly in high-frequency bands. These inconsistencies are generally imperceptible to humans but detectable by a trained detection model.

3.3 Existing Detection Techniques

Early detection methods for synthetic audio relied on statistical analysis, focusing on phase mismatches and spectral inconsistencies. As synthetic voices improved, deep learning-based methods became more prominent.

Statistical Analysis: Traditional techniques, such as bi-spectral analysis, capture phase inconsistencies in synthetic audio by analyzing the spectral relationship across multiple frequencies. These methods work well for simple synthesis models but often fail when applied to advanced vocoders.

Deep Learning Approaches: Advanced models, such as RawNet2 and LFCC-LCNN, use neural networks to learn complex spectro-temporal features directly from audio waveforms. RawNet2, for instance, employs convolutional and recurrent layers that enable it to capture detailed patterns in raw waveforms, making it particularly effective in detecting synthetic audio artifacts. However, these models are often limited in their ability to generalize, as they tend to overfit to vocoder types seen during training.

4. METHODOLOGY

To address the limitations of existing methods, we propose a multi-task learning framework that incorporates vocoder artifact detection into the primary task of synthetic voice classification.

4.1 Model Architecture and Algorithmic Approach

Our detection model is based on the RawNet2 architecture, which processes raw audio waveforms through convolutional and recurrent layers. This design enables it to capture both short-term and long-term dependencies in audio, enhancing its ability to identify vocoder-specific artifacts.

Binary Classification Module: This module focuses on detecting whether a given audio sample is real or synthetic. RawNet2's convolutional layers process the raw waveform, extracting localized features, while the GRUs capture temporal dependencies, allowing the model to learn subtle vocoder artifacts that indicate synthesis.

Auxiliary Vocoder Identification Module: This module classifies synthetic samples by their vocoder type, using a multi-class classification objective. By training on specific vocoder categories, the vocoder identification module encourages the shared feature extraction layers to focus on vocoder-specific details. This setup improves the model's ability to generalize, as it can leverage vocoder-specific artifacts even when faced with vocoders not seen during training.

The multi-task objective is formulated as:

Total Loss=$\lambda L_{binary}+(1-\lambda)L_{vocoder}$

where L_{binary} is the binary cross-entropy loss for real vs. synthetic classification, $L_{vocoder}$ is the categorical cross-entropy loss for vocoder identification, and λ is a hyperparameter balancing the tasks. This setup allows the model to concentrate on features relevant to both vocoder-specific patterns and synthetic audio detection.

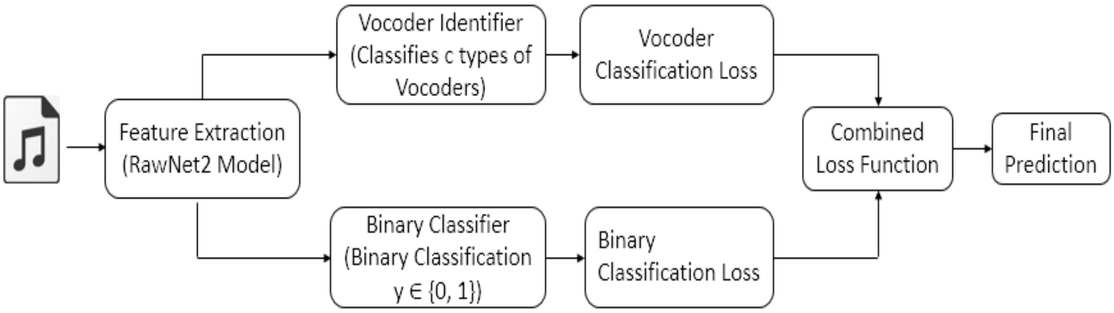
Vocoder Identifier (Classifies c types of Vocoders)
Vocoder Classification Loss
Feature Extraction (RawNet2 Model)
Combined Loss Function
Final Prediction
Binary Classifier (Binary Classification y ∈ {0, 1})
Binary Classification Loss

Fig 4.1

The proposed framework for detecting synthetic human voices is illustrated in the Fig 5.1. It begins with the Raw Audio Input **(x)**, which consists of the raw waveform of human voice signals. This input is fed into the Feature Extraction Module (RθR(x)), utilizing the RawNet2 model specifically designed to operate on raw audio. This approach preserves critical information necessary for detecting vocoder artifacts that may be present in synthetic voices.

The feature-extracted output is then directed towards two parallel classifiers. The first is the Vocoder Identifier (MθM), responsible for classifying the synthetic audio into one of the c possible neural vocoder models. This classification task is crucial for identifying specific vocoder artifacts that are indicative of synthetic audio generation. The second classifier is the Binary Classifier (BθB), which predicts whether the audio sample is authentic (y = 0) or synthetic (y = 1).

Both classifiers employ distinct loss functions for optimization. The Vocoder Classification Loss (L_m) quantifies the error in classifying the vocoder type, while the Binary Classification Loss (L_b) measures the accuracy in distinguishing between real and synthetic audio. The overall optimization objective combines these two loss functions into a Combined Loss Function, represented as $\mathrm{Min}_{\theta b,\, \theta r}\lambda L_b+\mathrm{Min}_{\theta m,\, \theta r}(1-\lambda)L_m$ an adjustable hyper-parameter that balances the contribution of each task.

Finally, the outcome of the classifiers leads to the Final Prediction (yˆ = Fθ(x)), which determines the label of the input audio, identifying it as either real or synthetic. This multi-task learning approach emphasizes the importance of vocoder artifact detection, enhancing the classifier's sensitivity to the unique statistical characteristics of synthetic audio generated through neural vocoders.

4.2 Training Algorithm and Dataset

Dataset: LibriSeVoc: The LibriSeVoc dataset was developed from the LibriTTS corpus to capture vocoder artifacts. It includes real and synthetic audio samples generated by six vocoders: WaveNet, WaveRNN, WaveGrad, DiffWave, Mel-GAN, and Parallel WaveGAN. Each vocoder introduces unique artifacts into the audio, and the dataset spans various durations (5-34 seconds) to ensure diversity in training.

Training Process: The model is trained using the Adam optimizer with a learning rate of 0.0001, dynamically reduced through cosine annealing for convergence. Data augmentation techniques, including noise addition and resampling, are used to simulate real-world conditions, enhancing robustness. Training runs for 100 epochs, with early stopping based on validation accuracy.

5. RESULT AND ANALYSIS

5.1 Intra-dataset Evaluation

On the LibriSeVoc dataset, the multi-task learning framework achieved a low Equal Error Rate (EER) of 0.13%, significantly outperforming models that rely solely on binary classification. By integrating vocoder identification, the model learned to prioritize vocoder-specific patterns, enhancing classification accuracy across different vocoders.

5.2 Cross Dataset Evaluation

To evaluate generalization, the model was tested on the WaveFake dataset, which includes synthetic samples from GAN-based vocoders not present in training. The model achieved an EER of 26.95%, showing strong generalization, particularly on FullBand-MelGAN and HiFi-GAN vocoders.

5.3 Robustness Evalution

The model's robustness to post-processing transformations, such as resampling and noise addition, was evaluated. An EER of 2.73% was maintained in these conditions, indicating resilience to common real-world audio modifications

6. CONCLUSION

This paper presents a multi-task learning framework that leverages vocoder artifacts for synthetic voice detection. Our approach integrates binary classification with vocoder identification, improving accuracy and generalizability across different vocoder types. The results demonstrate the model's effectiveness for real-world detection, providing a promising solution for audio authentication.

7. FUTURE SCOPE

Future work will focus on expanding vocoder coverage, integrating multi-modal cues, and optimizing for real-time applications to enhance the detection of synthetic audio across various platforms

REFERENCES

[1] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 9231–9235. 1, 3

[2] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, "Learning from yourself: A self-distillation method for fake speech detection," arXiv preprint arXiv:2303.01211, 2023. 1, 3

[3] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1859–1872, 2014. 2

[4] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multiresolution spectrogram," in ICASSP, 2020. 3

[5] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "Deepsonar: Towards effective and robust detection of ai-synthesized fake voices," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1207–1216. 3

[6] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in S

[7] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, "Deepfake audio detection by speaker verification," in 2022 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2022, pp. 1–6. 3

[8] Ousama a. Shaaban, Remzi Yildirim, and Abubaker a. Alguttar. "Audio Deepfake Approaches" IEEE Base Papers

[9] Daniele Ugo Leonzio, Luca Cuccovillo, Paolo Bestagini Marco Marcon Patrick Aichroth, Stefano Tubaro "Audio Splicing Detection and Localization Based on Acquisition Device Traces" IEEE Transactions on information forensics and security vol 18 2023

[10] Ghulam Ali, javed Rashid, Muhammad Rameez ul Hussnain, Muhammad Usman Tariq, Anwar Ghani, and Daehan Kwak. "Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection" IEEE TRANSACTIONS and JOURNALS

[11] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad s. Almadhor, Zunera Jalil and Rouba Borghol. "Deepfake Audio Detection via MFCC Features Using Machine Learning"

[12] Bismi Fathima Nasar, Sajini T and Elizabeth Rose Lason ." Deepfake Detection in Media Files - Audios, Images and Videos