Deepfake Audio Detection: Differentiating Modulated and Deepfake Audio

Sai Sri Harsha, Chakravarthula
Dept. of Computer Science and Engineering
University of North Texas, USA
Email: SaiSriHarshaChakravarthula@my.unt.edu

Alakananda Mitra
Nebraska Water Center
University of Nebraska-Lincoln, USA
Email: amitra6@unl.edu

Saraju P. Mohanty
Dept. of Computer Science and Engineering
University of North Texas, USA
Email: saraju.mohanty@unt.edu

Elias Kougianos
Dept. of Electrical Engineering
University of North Texas, USA
Email: elias.kougianos@unt.edu

***Abstract*—The rise in deepfake audio and AI-generated speech has become a very real and serious security issue. The voices that are generated using deepfakes are becoming so realistic that it is difficult to differentiate between a real speech and a deepfake speech. Traditional detection techniques that rely only on audio features will fail in such cases. The paper introduced a hybrid approach to deepfake audio detection that combines modern deep learning with classic audio analysis. The paper uses a Wav2Vec2 embedding to capture speech features and MFCC features to highlight fine details in the sound. It shows that using Wav2Vec2 alone achieves about 81% accuracy, while adding MFCCs boosts performance to 85%, with noticeable improvements in detecting modulated speech. It demonstrates that combining deep embeddings, traditional features, and realistic data augmentations leads to a more reliable detection system.**

***Index Terms*—Audio deepfake detection, Wav2Vec2, MFCC, Multilayer Perceptron (MLP), modulated audio, speech synthesis.**

I. Introduction

The AI-based speech generation has changed the way audio content can be generated. Modern generative models like text-to-speech (TTS) systems and voice conversion frameworks can generate very realistic deepfake audio that can mimic a speaker's identity and speech content very accurately. Although the systems are very beneficial applications in assistive technologies, entertainment, and language learning, they also raise serious security and trust concerns. When misused, deepfake audio can be used to commit financial fraud by impersonating, spread misinformation in political or social contexts, or facilitate unauthorized access to systems that rely on voice-based authentication. [1], [2]. Detection of such deepfakes is hard and not so straightforward. The Figure. **??** shows how speech can be affected in different environments.

II. Literature Survey

The recent developments of deepfake audio techniques, including text-to-speech (TTS) and voice conversion (VC), have brought about new communication capabilities as well as considerable security issues. As the realism of synthetic

Figure 1. Illustration of Speech Generation in Different Environments such as Indoor, Outdoor, and Synthetic.

speech production improved, researchers have looked more to deep learning based detection methods [3], [4]. The existing research converts audio files into a spectrogram or MFCC and uses CNNs to detect the deepfake. Advanced models like hybrid CNN-RNN architectures [5] and transformers-based frameworks like Wav2Vec2 and RawNet2 [6] [7] will learn the speech patterns and achieve detection accuracy, but they overlook real-world distortions like change in pitch, variations in tempo, external white noise, etc. Previous spoofing models relied mostly on such techniques to create deepfake audio, whereas current models such as Wavenet, AutoVC, and VITS add errors into the speech which can only be found out in deeper linguistic layers [8], [9]. Despite progress, such as ASVspoof [10], limited attention to modulated audio reduces model robustness in real-world conditions. Bridging this gap is essential to developing audio detectors that reliably distinguish between genuine, modulated, and deepfake speech.

III. Novel Contribution and Proposed Solution

The novelty of the proposed work lies in combining deepfake audio detection with modulation analysis. This helps in lowering the gap in deepfake detection as natural signal modulations, such as pitch or frequency shifts, can mimic deepfake artifacts and confuse existing detectors. This will prevent and make it clear in distinguishing real, modulated, and deepfake speech.

The paper proposes a modulation augmented dataset that combines genuine, modulated, and deepfake audio, enabling fair and reproducible evaluation. The paper proposes a modulated class to the existing Fake-or-Real (FoR) dataset to build a deepfake detecting approach where it can detect genuine, modulated, and deepfakeclassess. Such detectors can be used for real world application. The modulated class is introduced by applying the pitch shift, volume variation, adding background noise, clipping, and speed variation. This will help the model to distinguish between all three classes as shown in Figure. **??**.

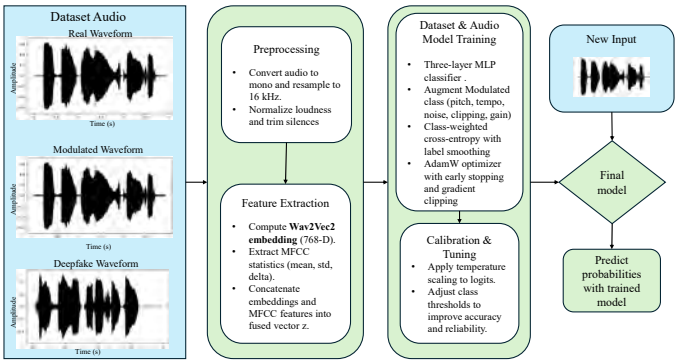

Figure 2. Overview of the Proposed Deepfake and Modulated Audio Differentiation System.

The audio clips are represented by two feature sets, which are a 768- dimensional Wav2Vec2 embedding that can capture contextual and semantic cues, and MFCC-based statistical features that can preserve the spectral characteristics. The features fused are passed through an MLP classifier for prediction. To mitigate the imbalances in the class and improve the model stability, the training process useclass-weighteded loss and balanced sampling with label smoothing technique for regularization, then temperature scaling with threshold tuning calibration to obtain the output probabilities. This helps in obtaining a lightweight detector model, which is useful in detecting deepfake speech from real speech across diverse recording conditions.

IV. Implementation and Results

The paper is implemented using PyTorch, where torchaudio is used for audio processing, transformers are used for Wav2Vec2 embeddings, librosa is used for MFCC feature extraction, and scikit-learn is used for evaluation metrics and probability calibration. The Input audio files were converted to 16 kHz mono, and the loudness is normalized for uniformity. Using the Fake-or-Real (FoR) corpus [11], we labeled Real and Fake samples and introduced a third Modulated class by applying small pitch and tempo shifts, light clipping, subtle gain adjustments, and limited noise injection to genuine speech. These controlled augmentations simulate realistic recording variations without changing the quality.

Each sample was represented using two complementary feature views: a 768-dimensional Wav2Vec2 embedding that captures contextual and semantic information, and MFCC-based statistical features that preserve fine-grained spectral patterns. These features were concatenated and fed into a compact three-layer MLP with GELU activations, Batch Normalization, Dropout, and a three-way softmax output. Class imbalance was addressed through class-weighted cross-entropy loss and a weighted random sampler. Training employed AdamW optimization, along with label smoothing, light mixup, mild feature noise, gradient clipping, and a cosine annealing scheduler with warm restarts. The validation macro-F1 score guided early stopping. The overall training pipeline is summarized in Algorithm **??**. After training, probability calibration was performed using temperature scaling, and decision thresholds were lightly tuned via coordinate search to identify the most accurate and stable operating point. As illustrated in Figure. 3, the model shows stable convergence, with the validation accuracy rising as the training loss decreases. This confirms the network's ability to generalize beyond the training data.

Algorithm 1 : Deepfake Detection (x)

Require: Audio clip x
Ensure: Label $y \in \{\text{REAL}, \text{FAKE}, \text{MODULATED}\}$
Preprocessing:
$x \leftarrow \text{resample}(x, 16\text{kHz})$
$x \leftarrow \text{mono}(x)$
$x \leftarrow \text{normalize}(x)$
$x \leftarrow \text{trim\_silence}(x)$
Feature Extraction:
$e \leftarrow \text{Wav2Vec2}(x)$ // 768-D embedding
$m \leftarrow \text{MFCC}(x)$ // statistics (mean, std, delta)
$z \leftarrow \text{concatenate}(e, m)$
Training:
for each epoch **do**
  Augment data with modulation (pitch, tempo, noise, clipping, gain)
  Compute loss $=$ weighted CE $+$ label smoothing
  Update MLP parameters using AdamW
  Apply mixup, dropout, and early stopping
end for
Calibration:
Calibrate probabilities using temperature scaling
Tune decision thresholds τ
Inference:
$p \leftarrow \text{softmax}(\text{MLP}(z)/T)$
$y \leftarrow \text{argmax\_with\_thresholds}(p, \tau)$
return y

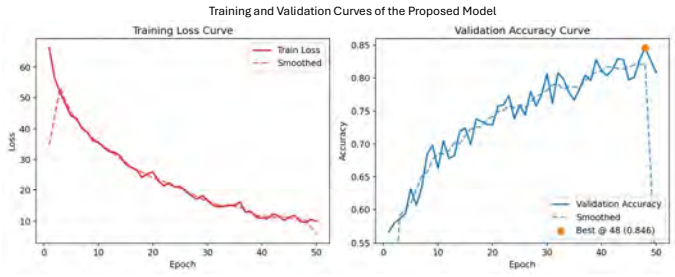

Figure 3. Training and Validation Performance of the Hybrid (W2V2+MFCC) Model.

TABLE I
PERFORMANCE METRICS OF THE MODEL

Model	Accuracy	Macro-F1	Per-class F1		
			Real	Modulated	Fake
Wav2Vec2 only	0.81	0.80	0.84	0.75	0.84
Hybrid (W2V2+MFCC)	**0.85**	**0.83**	**0.83**	**0.83**	**0.85**

As shown in Figure. 4, the hybrid *Wav2Vec2 + MFCC* model outperforms the Wav2Vec2-only baseline, with a more evident diagonal pattern and reduced off-diagonal confusion—particularly in distinguishing the *Modulated* class. The hybrid model achieved 85% accuracy, compared to 81% for the Wav2Vec2-only configuration, and a higher macro-F1 score.

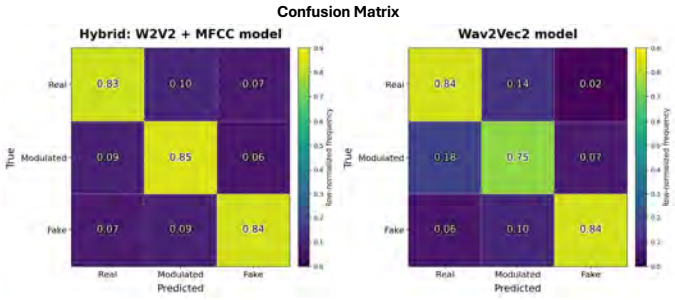

Figure 4. Confusion matrices for Hybrid (W2V2+MFCC) and Wav2Vec2-only models.

The experimental results in Table I show the efficiency of the proposed hybrid approach in showing the real-world audio uncertainty. The addition of the Modulated class helped the system learn minute acoustic differences between a real and AI-generated voice. With the integration of the contextual embeddings in Wav2Vec2 with detailed MFCC-based spectral cues, the model shows a strong understanding of both the acoustic and linguistic information. This allows the detector to remain more reliable even when the data is new, with environmental noise or altered playback conditions, which will make it a trustworthy deepfake audio detection that can be used for everyday applications.

V. CONCLUSION

This paper has presented a practical, adaptive approach for reliable deepfake audio detection in real-time environments, such as in noisy environments and phone calls. By adding a dedicated *Modulated* class, the model learns to distinguish natural distortions such as pitch or tempo changes from AI-generated speech. The proposed hybrid architecture integrates Wav2Vec2 embeddings, which capture linguistic cues with MFCC-based spectral features adds more important features for the model to become more reliable. This will help the lightweight MLP classifier to capture the details in the human speech better while maintaining the real-world changes. The hybrid approach has improved the classification accuracy from 81% to 85% compared to the baseline model. Overall, the proposed framework demonstrated that integrating deep contextual representations with the traditional audio features has led to a highly resilient and trustworthy deepfake detection system that moves closer to practical deployment in real-world voice verification scenarios.

VI. ACKNOWLEDGMENT

The Figure. **??** has been generated using ChatGPT with the prompt "Generate a cartoon of three small scenes in one frame: man on phone at home, man on phone in park with trees and birds, and masked man on computer screen with waveforms."

REFERENCES

[1] A. Qais, A. Rastogi, A. Saxena, A. Rana, and D. Sinha, "Deepfake audio detection with neural networks using audio features," in *Proceedings of the 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, Jul. 2022, pp. 1–6.

[2] A. Mitra, S. P. Mohanty, and E. Kougianos, "The World of Generative AI: Deepfakes and Large Language Models," 2024. [Online]. Available: https://arxiv.org/abs/2402.04373

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[4] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A novel machine learning based method for deepfake video detection in social media," in *Proceedings of the IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, Dec. 2020, pp. 91–96.

[5] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Nov. 2018, pp. 1–6.

[6] G. Lee, J. Lee, M. Jung, J. Lee, K. Hong, S. Jung, and Y. Han, "Dual-channel deepfake audio detection: Leveraging direct and reverberant waveforms," *IEEE Access*, vol. 13, pp. 18 040–18 052, 2025.

[7] O. Pascu, A. Stan, D. Oneata, E. Oneata, and H. Cucu, "Towards generalisable and calibrated synthetic speech detection with self-supervised representations," 2024. [Online]. Available: https://arxiv.org/abs/2309.05384

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *ArXiv*, vol. abs/2106.06103, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235417304

[9] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. A. Hasegawa-Johnson, "Zero-shot voice style transfer with only autoencoder loss," *ArXiv*, vol. abs/1905.05879, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:155091770

[10] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, A. L. Kong, V. Vestman, A. Nautsch *et al.*, "Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database," http://urn.fi/urn:nbn:fi:att:4e69bd10-66b8-4b0a-9854-c2b738ef721a, 6 2019.

[11] M. Abdeldayem. (2025) The fake-or-real (for) dataset: Deepfake audio. [Online]. Available: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset