Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

VocaTrust: Machine Learning Based Voice Integrity Assessment

Prof. Yashodara R[1], Prof. Divyashree K[2], Deeksha[3], Dimple G[4], Sinchana Gowda RP[5], Varsha NG[6]

[1,2] Assistant Professor, Information Science and Engineering.
[3,4,5,6] Undergraduate Student, Information Science and Engineering.

Abstract

The rapid advancement of artificial intelligence has enabled the creation of highly realistic synthetic voice recordings, making it increasingly difficult to distinguish between genuine and manipulated speech. The misuse of AI-generated voices in banking fraud, identity theft, misinformation, and unauthorized access has created a growing need for reliable deepfake voice detection systems. This paper proposes VocaTrust: Machine Learning-Based Voice Integrity Assessment, a machine learning framework for detecting manipulated audio and assessing voice authenticity.

The proposed system utilizes genuine audio from the LibriSpeech dataset and synthetic audio from the ASVspoof dataset. Audio samples are pre-processed and converted into standardized formats before extracting Mel-Frequency Cepstral Coefficients (MFCCs). These features are used to train a Convolutional Neural Network (CNN) model that classifies audio as genuine or fake and generates a Voice Integrity Score to indicate the authenticity of the input speech.

The proposed framework is lightweight, efficient, and designed for real-time voice analysis using only audio input. It provides reliable detection of deepfake speech while reducing computational complexity, making it suitable for applications such as voice authentication, banking security, digital forensics, and voice assistant protection.

Keywords: Deepfake Voice Detection, Voice Integrity Assessment, Machine Learning, CNN, MFCC, Speech Spoofing.

1. Introduction

Artificial Intelligence (AI) has significantly transformed the field of speech synthesis, enabling the generation of highly realistic synthetic voices using advanced deep learning techniques. Modern text-to-speech (TTS) and voice cloning models can accurately imitate a person's tone, pitch, pronunciation, and speaking style, producing speech that is often indistinguishable from genuine human voices. While these technologies have numerous beneficial applications in virtual assistants, accessibility, entertainment, and communication systems, they have also introduced serious security and ethical challenges. AI-generated voices are increasingly being exploited for identity impersonation, financial fraud, misinformation, social engineering attacks, and unauthorized access to voice-based authentication systems.

The rapid growth of deepfake audio has created significant concerns in cybersecurity and digital forensics. Traditional voice authentication systems primarily rely on speaker recognition techniques and often struggle to distinguish between genuine and AI-generated speech. Existing deepfake detection methods generally perform binary classification, identifying an audio sample as either real or fake. However, many of these systems require large computational resources, depend on multimodal data such as audio and video, or show reduced performance when processing noisy recordings and unseen spoofing attacks. These limitations reduce their effectiveness in practical real-world environments.

To address these challenges, this paper proposes VocaTrust: Machine Learning-Based Voice Integrity Assessment, a lightweight and efficient framework for detecting manipulated voice recordings using machine learning techniques. The proposed system utilizes genuine speech samples from the LibriSpeech dataset and spoofed speech samples from the ASVspoof dataset. The audio data undergoes preprocessing followed by the extraction of Mel-Frequency Cepstral Coefficients (MFCCs), which effectively capture the acoustic characteristics of speech. These features are used to train a Convolutional Neural Network (CNN) capable of distinguishing genuine and manipulated audio samples with high reliability.

Unlike conventional deepfake detection systems, the proposed framework introduces a Voice Integrity Score, which provides a confidence-based assessment of the authenticity of an input voice recording instead of only producing a binary classification result. The system is designed as an audio-only solution, making it computationally efficient and suitable for real-time deployment. Its lightweight architecture and robust feature extraction process enable reliable performance under practical recording conditions while supporting future extensions for multilingual voice analysis. The proposed framework has potential applications in voice authentication, banking security, digital forensics, call verification systems, and voice assistant protection, contributing to the development of secure and trustworthy voice-based technologies.

2. Literature Survey

Several researchers have proposed different machine learning and deep learning techniques for detecting AI-generated speech and audio deepfakes. These studies mainly focus on improving detection accuracy, robustness, and generalization using various feature extraction methods and neural network architectures. The following literature review summarizes some of the significant contributions related to deepfake voice detection.

1. Naveed, S. J., et al. (2026)

This study proposed DEEPSHIELD, a multimodal deep learning framework that combines audio and video information using CNN, BiLSTM, GRU, and cross-modal attention mechanisms. The model achieved high detection accuracy by learning both facial and speech features but required both audio and video inputs, increasing computational complexity.

2. Jung, J. W., et al. (2025)

This research introduced the SpoofCeleb dataset for speech deepfake detection using large-scale real and synthetic voice recordings. The study evaluated several deep learning models and demonstrated improved robustness for detecting spoofed speech in real-world environments.

3. Schäfer, K., and Steinebach, M. (2025)

The authors evaluated Automatic Speech Recognition (ASR) encoder-based feature extraction techniques using Whisper, SpeechT5, Wav2Vec2, and CNN models. Their results showed that deep learning-based acoustic representations improved deepfake detection performance compared to traditional handcrafted features.

4. Guragain, A., et al. (2024)

This work proposed an ensemble framework using self-supervised learning models and AASIST for singing voice deepfake detection. The combination of multiple speech representations improved detection accuracy and reduced error rates across different evaluation datasets.

5. Zhang, Y., et al. (2024)

The authors introduced the SVDD2024 challenge and benchmark datasets for singing voice deepfake detection. The study provided a standardized evaluation framework for comparing different deep learning models under realistic recording conditions.

6. Chaiwongyen, A., et al. (2023)

This research presented a deepfake speech detection method using pathological speech features such as jitter, shimmer, and harmonic-to-noise ratio combined with a Multilayer Perceptron (MLP) classifier. The proposed approach achieved good detection accuracy while remaining computationally efficient.

7. Yadav, A. K. S., et al. (2023)

The proposed PS3DT framework utilized Mel-spectrogram patches and Transformer networks for synthetic speech detection. The model demonstrated improved generalization and robust performance against different speech synthesis techniques.

8. Yan, R., et al. (2022)

This study developed an audio deepfake detection system using ResNet-34, multi-head attention, neural stitching, and LFCC features. The proposed framework enhanced detection accuracy and improved robustness against unseen spoofing attacks.

3. Methodology

The proposed VocaTrust framework detects deepfake speech using a machine learning pipeline consisting of data collection, preprocessing, feature extraction, CNN-based classification, model evaluation, and voice integrity scoring.

Data Collection

Genuine speech samples are collected from the LibriSpeech dataset, while fake speech samples are obtained from the ASVspoof dataset. A balanced dataset of real and fake audio is prepared to improve model training and detection performance.

Audio Preprocessing

The audio files are converted to mono format, resampled to 16 kHz, normalized, and segmented into fixed-duration clips. Data augmentation is applied to improve robustness against noise and recording variations.

Feature Extraction

The system extracts 13 Mel-Frequency Cepstral Coefficients (MFCCs) from each audio sample. These features capture important speech characteristics and serve as input to the classification model.

CNN-Based Classification

The extracted MFCC features are fed into a Convolutional Neural Network (CNN) for classifying audio as genuine or fake. The CNN automatically learns discriminative patterns from speech features to improve detection accuracy.

Model Training and Evaluation

The dataset is divided into 80% training and 20% testing data. The model is trained using the Adam optimizer and Binary Cross-Entropy loss function and evaluated using standard performance metrics.

Voice Integrity Scoring

The trained model generates a Voice Integrity Score ranging from 0 to 100. Based on this score, the input audio is classified as Genuine, Suspicious, or Fake, providing a confidence-based assessment of voice authenticity.

4. EXPECTED CONCLUSION

1. **Detection Performance**: The proposed VocaTrust system is expected to accurately distinguish between genuine and AI-generated speech using MFCC features and a CNN-based model. This improves the reliability of deepfake voice detection.

2. **Voice Integrity Assessment:** The system generates a Voice Integrity Score (0–100) to measure the authenticity of an input voice sample. This provides a confidence-based assessment instead of a simple real/fake classification.

3. **Practical Applications:** The proposed framework can be applied in voice authentication, banking security, digital forensics, and fraud detection. Its lightweight architecture makes it suitable for practical deployment.

4. **Final Assessment:** The proposed system provides an efficient and reliable solution for deepfake voice detection. It also offers scope for future enhancements such as multilingual support and real-time voice analysis.

References

1. Naveed, S. J., Alif, A. I., & Saha, S. (2026). DEEPSHIELD: A Multimodal Deep Learning Model for Audio-Visual Deepfake Detection. Proceedings of the 5th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), pp. 1–6.
2. Wani, T. M., & Amerini, I. (2026). Multi-Scale Self-Supervised Learning for Efficient Audio Deepfake Detection. IEEE Signal Processing Letters, 33, pp. 46–50.
3. Jung, J. W., et al. (2025). SpoofCeleb: Speech Deepfake Detection and SASV in the Wild. IEEE Open Journal of Signal Processing, Vol. 6, pp. 68–77.
4. Schäfer, K., & Steinebach, M. (2025). Generalization in Audio Deepfake Detection: Evaluating ASR Encoder-Based Feature Extraction. Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–5.
5. Coletta, E., Salvi, D., Negroni, V., Leonzio, D. U., & Bestagini, P. (2025). Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching. Proceedings of the European Signal Processing Conference (EUSIPCO), pp. 820–824.
6. Pham, L., Tran, D., Lam, P., Skopik, F., Schindler, A., Poletti, S., Fischinger, D., & Boyer, M. (2025). DIN-CTS: Low-Complexity Depthwise-Inception Neural Network with Contrastive Training Strategy for Deepfake Speech Detection. Proceedings of the European Signal Processing Conference (EUSIPCO), pp. 556–560.
7. Chauhan, S., Sethi, K., & Anand, A. (2025). Audio Deepfakes Detection: A Comprehensive Literature Review of Voice Cloning Fraud Detection Techniques. Proceedings of the IEEE 3rd International Symposium on Sustainable Energy, Signal Processing and Cybersecurity (iSSSC), pp. 1–5.
8. Guragain, A., Liu, T., Pan, Z., Sailor, H. B., & Wang, Q. (2024). Speech Foundation Model Ensembles for the Controlled Singing Voice Deepfake Detection (CTRSVDD) Challenge 2024. Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 774–781.
9. Zhang, Y., Zang, Y., Shi, J., Yamamoto, R., Toda, T., & Duan, Z. (2024). SVDD 2024: The Inaugural Singing Voice Deepfake Detection Challenge. Proceedings of the IEEE Spoken Language Technology

Workshop (SLT), pp. 782–787.
10. Chiddarwar, P. (2024). Real-Time Detection of AI-Generated Deepfake Audio: A Novel Approach. Proceedings of the IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), pp. 1–5.
11. Jain, E., & Singh, A. (2024). Deepfake Voice Detection Using Convolutional Neural Networks: A Comprehensive Approach to Identifying Synthetic Audio. Proceedings of the 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS), pp. 1–5.
12. Chaiwongyen, A., Duangpummet, S., Karnjana, J., Kongprawechnon, W., & Unoki, M. (2023). Deepfake Speech Detection with Pathological Features and Multilayer Perceptron Neural Network. Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 2182–2188.
13. Salvi, D., Bestagini, P., & Tubaro, S. (2023). Reliability Estimation for Synthetic Speech Detection. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5.
14. Singh Yadav, A. K., Xiang, Z., Bhagtani, K., Bestagini, P., Tubaro, S., & Delp, E. J. (2023). PS3DT: Synthetic Speech Detection Using Patched Spectrogram Transformer. Proceedings of the International Conference on Machine Learning and Applications (ICMLA), pp. 496–503.
15. Yan, R., Wen, C., Zhou, S., Guo, T., Zou, W., & Li, X. (2022). Audio Deepfake Detection System with Neural Stitching for ADD 2022. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9226–9230.
16. Cui, S., Huang, B., Huang, J., & Kang, X. (2022). Synthetic Speech Detection Based on Local Autoregression and Variance Statistics. IEEE Signal Processing Letters, 29, pp. 1462–1466.
17. Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M. C., & Tubaro, S. (2022). Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8962–8966.