International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

An Edge Device for Real-Time Fake and Tampered Audio Detection

Dr.Ranjit Bobate[1], Himanshi Tirale[2], Kanika Arora[3], Avantika Admane[4], Falguni Jambhulkar[5]

Department of Electronics and Communication
Priyadarshini College of Engineering
Nagpur, Maharshtra , India
kanikaarora297@gmail.com

ABSTRACT— The rapid development of artificial intelligence enables realistic speech generation through text-to-speech, voice cloning, and speech synthesis, which can also be misused for impersonation, fraud, and misinformation. This paper presents "An Edge Device for Real-Time Fake and Tampered Audio Detection," a software-based system that uses Digital Signal Processing (DSP) techniques to analyze audio authenticity. The input audio is divided into short frames, and features such as Zero Crossing Rate (ZCR) variance, Auto-Correlation (AC) variance, and Root Mean Square (RMS) variance are extracted to capture variations in phonetic activity, pitch periodicity, and amplitude dynamics. A mathematical scoring and decision mechanism combines these features to estimate whether an audio signal is AI-generated or manipulated. The system aims to provide a lightweight and responsive solution for real-time audio authenticity analysis while considering challenges such as noise, compression, and recording conditions.

KEYWORDS — Audio Deepfake Detection, Tampered Audio, Digital Signal Processing, Voice Authenticity, Zero Crossing Rate, Auto-Correlation, Root Mean Square, Speech Analysis, Real-Time Detection.

I. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and speech synthesis technologies has made it possible to generate highly realistic human-like speech. Techniques such as Text-to-Speech (TTS), voice cloning, and speech synthesis can reproduce a person's voice with increasing accuracy. Although these technologies have useful applications, they can also be misused to create fake or manipulated audio for impersonation, fraud, misinformation, social engineering, or manipulation of digital evidence. As synthetic speech becomes increasingly difficult to distinguish from genuine speech through ordinary listening, automated audio authenticity analysis has become an important requirement.

Fake audio refers to speech artificially generated or synthesized using AI-based techniques, whereas tampered audio refers to genuine recordings altered through modification, insertion, removal, or other manipulation. Both can affect the reliability of voice-based information and create challenges in voice authentication, digital forensics, secure communication, and media verification. Therefore, a detection system must analyze the characteristics of an audio signal rather than relying only on human perception.

Digital Signal Processing (DSP) provides an efficient approach for analyzing speech signals through their temporal and acoustic characteristics. In the proposed system, the input audio is divided into short frames, and features such as Zero Crossing Rate (ZCR) variance, Auto-Correlation (AC) variance, and Root Mean Square (RMS) variance are extracted to study phonetic activity, periodicity, pitch-related characteristics, and amplitude dynamics. These features provide indications of differences between natural human speech and artificially generated or manipulated audio.

The proposed work focuses on a software-based real-time analysis framework that combines DSP-based feature extraction with a mathematical scoring and decision mechanism. The extracted features are evaluated across multiple short-time frames to estimate whether the input audio is likely to be genuine or AI-generated/manipulated. The system also includes an auxiliary AI-assisted analysis interface for additional acoustic and linguistic observations, while the primary detection mechanism is based on mathematical DSP analysis.

The main objective is to develop a responsive and lightweight framework for identifying suspicious characteristics in audio signals and supporting real-time voice authenticity analysis. The proposed system demonstrates how fundamental DSP features can be combined for an interpretable approach to fake and tampered audio detection while highlighting challenges caused by background noise, echo, audio compression, recording conditions, and modern synthetic speech.

II. PROBLEM STATEMENT

The increasing use of AI for speech synthesis, voice cloning, and audio manipulation makes it difficult to distinguish genuine speech from fake or tampered audio through human listening alone, creating risks in voice authentication, communication, fraud prevention, and digital forensics. Existing detection methods often rely on computationally intensive machine-learning or deep-learning models requiring trained parameters, computational resources, and suitable datasets. Therefore, this work aims to develop a lightweight, real-time software-based system that analyzes speech using Digital Signal Processing

techniques. The proposed system examines frame-level variations in Zero Crossing Rate (ZCR), Auto-Correlation (AC), and Root Mean Square (RMS) and combines these characteristics through a mathematical scoring mechanism to provide an authenticity assessment with low computational complexity.

III. RELATED WORK

Audio authenticity and speech deepfake detection have gained importance with the growth of speech synthesis, voice conversion, and audio editing. Earlier studies analyzed time- and frequency-domain characteristics for detecting audio tampering, including waveform, spectral, and phase-based features [1], [2]. The ASVspoof 2021 challenge further examined synthetic and converted speech while considering channel variation and compression [3]. Recent approaches use machine-learning and deep-learning models for spoof detection, but these require trained models, computational resources, and suitable datasets. Digital audio forensic studies also emphasize the effects of sampling, encoding, and compression on signal analysis [4]. Based on these works, the proposed system uses a lightweight DSP-based approach with **ZCR variance, Auto-Correlation variance, and RMS variance**, combined through a scoring mechanism for real-time fake and tampered audio analysis.

IV. PROPOSED SYSTEM

The proposed system is a software-based real-time audio analysis framework designed to identify characteristics associated with fake and tampered speech. It uses Digital Signal Processing (DSP) techniques to analyze the input audio and determine whether its acoustic behavior is more consistent with natural or artificially generated/manipulated speech. The processing consists of audio acquisition, frame-based analysis, feature extraction, feature evaluation, and classification.

A. Audio Acquisition and Preprocessing

The system receives an audio signal through a microphone or uploaded audio file. The input is processed by the backend and divided into continuous frames of 256 samples. Frame-based processing allows changes in the speech signal to be examined over time before feature extraction.

B. DSP-Based Feature Extraction

Three primary features are extracted: Zero Crossing Rate (ZCR) variance, Auto-Correlation (AC) variance, and Root Mean Square (RMS). These features represent signal activity, periodicity, and amplitude dynamics.

1. Zero Crossing Rate Variance:

ZCR represents the rate at which the audio waveform changes its sign. ZCR values are calculated for successive frames, and their variance measures changes in crossing behavior over time. Natural speech contains varying voiced and unvoiced regions, while some synthetic speech may exhibit more uniform temporal characteristics

2. Auto-Correlation Variance:

Auto-correlation measures the similarity between a speech signal and a delayed version of itself and is useful for examining periodicity and pitch-related characteristics. The system calculates auto-correlation across successive frames and determines its variance. Differences in periodic behavior can provide information for authenticity analysis.



Fig - DSP Feature Analysis and Detection Performance

3. RMS Variance:

Root Mean Square (RMS) represents the energy or amplitude level of an audio frame. RMS values are calculated over successive frames and their variance measures changes in speech intensity. Variations caused by emphasis, pauses, and speaking intensity can provide additional information for authenticity analysis.

C. Feature-Based Scoring and Classification

After extracting the three DSP features, their statistical behavior is evaluated using a mathematical scoring mechanism. The individual feature scores are combined into an overall detection score representing the likelihood of synthetic or manipulated speech. The combined score is compared with a predefined decision threshold, and the audio is classified as **Real** or **AI-Generated/Manipulated**. A sensitivity parameter can be adjusted to modify the decision threshold.

D. Real-Time Analysis

The framework uses lightweight mathematical DSP operations rather than computationally intensive neural-network weights as the primary detection mechanism. Frame-level processing allows progressive analysis with low processing overhead, making the approach suitable for applications requiring quick audio verification.

E. AI-Assisted Analysis

The system also provides an optional AI-assisted analysis interface for additional interpretation of characteristics such as prosody, cadence, breathing patterns, hesitation, and linguistic behavior. This serves as an auxiliary analysis layer, while the primary detection mechanism remains based on DSP feature extraction and mathematical scoring.

The proposed system combines signal-level analysis with optional AI assistance to provide a lightweight framework for real-time fake and tampered audio detection.

V. METHODOLOGY

The proposed system follows a frame-based Digital Signal Processing (DSP) methodology for analyzing the authenticity of speech audio. The methodology consists of audio

acquisition, preprocessing, frame segmentation, extraction of signal-level features, statistical analysis, feature-based scoring, and final classification. The complete processing pipeline is designed to provide a lightweight and responsive assessment of whether an input audio signal is likely to be genuine or AI-generated/manipulated.

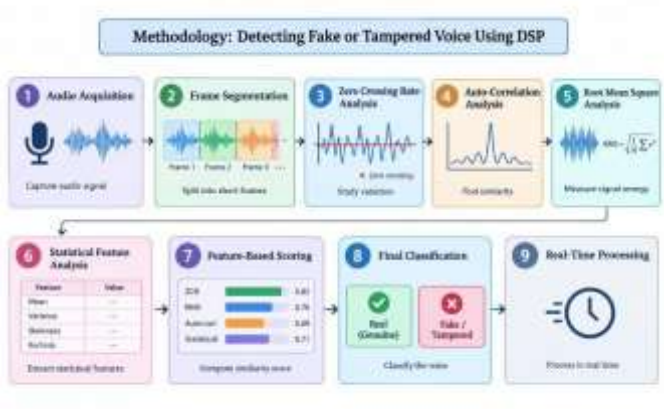


Fig – Methodology description for detecting tampered voices

A. Audio Acquisition

The system accepts an audio signal through a microphone or uploaded audio file. The received audio is processed by the backend and prepared for frame-based analysis.

B. Frame Segmentation

The audio signal is divided into consecutive frames of 256 samples. Each frame is analyzed independently, and the resulting feature values are used for statistical analysis across the complete audio signal.

C. Zero Crossing Rate Analysis

Zero Crossing Rate (ZCR) represents the number of times the audio waveform changes its sign within a frame. It provides information about the temporal characteristics and phonetic activity of speech. For a frame containing N samples, ZCR can be represented as

$$ZCR = \frac{1}{2N} \sum_{n=1}^{N-1} | sgn(x[n]) - sgn(x[n-1]) |$$

where $x[n]$ represents the audio sample and sgn represents the sign function.

ZCR is calculated for successive frames, and the variance of the resulting values is determined to measure changes in zero-crossing behavior throughout the recording.

D. Auto-Correlation Analysis

Auto-correlation measures the similarity between a signal and a delayed version of itself, providing information related to periodicity and pitch characteristics. For a frame, the auto-correlation at lag k can be expressed as

$$R_{xx}[k] = \sum_{n} x[n]x[n-k]$$

The auto-correlation characteristics are calculated across successive frames, and their variance is used as an indicator in the authenticity assessment.

E. Root Mean Square Analysis

Root Mean Square (RMS) represents the energy or amplitude level of an audio frame. For a frame containing N samples, RMS is calculated as

$$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x[n]^2}$$

RMS values are calculated across successive frames, and their variation provides an additional signal-level indicator for classification.

F. Statistical Feature Analysis

After calculating ZCR, AC, and RMS values for individual frames, their variations across the recording are statistically analyzed. The variance is given by:

$$Var(X) = \frac{1}{M} \sum_{i=1}^{M} (X_i - \bar{X})^2$$

where N is the number of analyzed frames, x_i is the feature value of a frame, and μ is the mean feature value.

These measurements provide information about temporal activity, periodicity, and amplitude dynamics.

G. Feature-Based Scoring

The extracted DSP characteristics are combined using a mathematical scoring mechanism. Each feature contributes to an overall detection score according to its observed behavior and predefined ranges. A decision threshold is then applied, which can be adjusted using the system's sensitivity setting.

H. Final Classification

Based on the combined feature score and selected threshold, the audio is classified as **Real** or **AI-Generated/Manipulated**. The classification is an estimate based on the selected DSP characteristics rather than absolute proof of authenticity.



Fig – Dashboard Presentation of Detection Performance

VI. IMPLEMENTATION

The proposed system is implemented as a software-based web application using **React with Vite** for the frontend and **Node.js with Express** for the backend. The frontend supports audio capture/upload and displays the detection score, extracted features, sensitivity, and classification results. The backend

processes the audio by dividing it into **256-sample frames** and calculating **ZCR, Auto-Correlation (AC), and RMS** values.



Fig - System Architecture and Implementation Workflow of the Proposed Audio Authenticity Detection System

Their variances are combined through a **rule-based mathematical scoring mechanism** with a configurable sensitivity threshold to classify the audio as **Real or AI-Generated/Manipulated**. An optional **Google Generative AI SDK** component provides additional analysis of prosody, cadence, breathing, hesitation, and linguistic characteristics, while the primary detection remains DSP-based. The frontend is deployed on **Vercel** and the backend on **Render**, with API communication enabled through **CORS**.

VII. CONCLUSION

The work presented a software-based framework for real-time fake and tampered audio detection using DSP techniques. The system analyzes short audio frames using ZCR variance, Auto-Correlation variance, and RMS variation to assess temporal activity, periodicity, and amplitude dynamics. The approach provides a lightweight and interpretable method for audio authenticity analysis; however, its reliability can be affected by noise, echo, compression, recording conditions, and advances in synthetic speech. Future work can improve robustness using additional features, diverse datasets, adaptive thresholds, and machine-learning or deep-learning models. Overall, the proposed system provides a foundation for real-time audio authenticity analysis.

REFERENCES

[1] R. Korycki, "Time and spectral analysis methods with machine learning for the authentication of digital audio recordings," *Forensic Science International*, vol. 230, no. 1–3, pp. 117–126, 2013.

[2] DGMA, "Artificial Intelligence Approach To Improve Lifespan Of Wireless Sensor Network : A Survey "" *International Conference on Innovation & Development of Future Smart Cities*, pp. 1–9.

[3] J. Yamagishi *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," *Proc. ASVspoof Challenge Workshop*, pp. 47–54, 2021.

[4] J.-W. Jung *et al.*, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6367–6371, 2022, doi: 10.1109/ICASSP43922.2022.9747766.

[5] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[6] H. Tak, J. Patino, D. Delgado, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End Anti-Spoofing with RawNet2," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[7] X. Wang *et al.*, "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," *Computer Speech & Language*, vol. 64, 2020.

[8] B. Ustubioglu, G. Tahaoglu, and G. Ulutas, "Detection of Audio Copy-Move-Forgery with Novel Feature Matching on Mel Spectrogram," *Expert Systems with Applications*, vol. 213, 2023, Art. no. 118963, doi: 10.1016/j.eswa.2022.118963.

[9] H.-P. Hsu *et al.*, "Detection of Audio Tampering Based on Electric Network Frequency Signal," *Sensors*, vol. 23, no. 16, Art. no. 7029, 2023, doi: 10.3390/s23167029.

[10] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," *Algorithms*, vol. 15, no. 5, Art. no. 155, 2022, doi: 10.3390/a15050155.

[11] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," *NeurIPS 2021 Datasets and Benchmarks Track*, 2021.

[12] N. Subramani and D. Rao, "Learning Efficient Representations for Fake Speech Detection," *Proc. AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 5859–5866, 2020, doi: 10.1609/aaai.v34i04.6044.

[13] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4SNet: Deep Learning for Fake Speech Classification," *Expert Systems with Applications*, vol. 184, 2021, Art. no. 115465, doi: 10.1016/j.eswa.2021.115465.

[14] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards End-to-End Synthetic Speech Detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021, doi: 10.1109/LSP.2021.3089437.

[15] A. Gómez-Alanis, A. M. Peinado, J. A. González, and A. M. Gómez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection," *Proc. Interspeech*, pp. 1068–1072, 2019, doi: 10.21437/Interspeech.2019-2212.

[16] Y. Ma, Z. Ren, and S. Xu, "RW-Resnet: A Novel Speech Anti-Spoofing Model Using Raw Waveform," *Proc. Interspeech*, 2021.

[17] C. Wang, J. Yi, J. Tao, H. Sun, X. Chen, Z. Tian, H. Ma, C. Fan, and R. Fu, "Fully Automated End-to-End Fake Audio Detection," *Proc. 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM)*, pp. 27–33, 2022, doi: 10.1145/3552466.3556530.

[18] C. Sun, S. Jia, S. Hou, and S. Lyu, "AI-Synthesized Voice Detection Using Neural Vocoder Artifacts," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[19] S.-Y. Chang, K.-C. Wu, and C.-P. Chen, "Transfer-Representation Learning for Detecting Spoofing Attacks with Converted and Synthesized Speech in Automatic Speaker Verification System," *Proc. Interspeech*, pp. 1063–1067, 2019, doi: 10.21437/Interspeech.2019-2014.

[20] Z. Xie, B. Li, X. Xu, Z. Liang, K. Yu, and M. Wu, "FakeSound: Deepfake General Audio Detection," *Proc. Interspeech*, 2024.