iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VIII Month of publication: August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73592

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Deep Fake Defender: AI-Based Detection of Deepfake Voice Attacks in Real-Time Voice Authentication Systems

Fazeeha M[1], Arfaah Meera K[2], Fahmidha Jinan A[3]
Department of Computer Science and Engineering, Arunachala College of Engineering for Women, Manavilai, Vellichanthai-629 203

Abstract: In recent years, advancements in artificial intelligence have led to the rapid development of deepfake technologies, including hyper-realistic voice cloning. While voice authentication systems are increasingly adopted for secure access in banking, smart devices, and enterprise systems, they remain vulnerable to deepfake audio attacks that can mimic a target's voice with alarming precision. This paper proposes an AI-powered framework for the real-time detection of deepfake voice samples in authentication scenarios. The system employs deep learning models trained on both authentic and synthetic voice datasets to analyze subtle acoustic features such as waveform anomalies, frequency inconsistencies, unnatural pauses, and generative noise artifacts. Using tools like Librosa for audio feature extraction and Convolutional Neural Networks (CNNs) for classification, the model achieves high accuracy in distinguishing real voices from AI-generated ones. The solution is designed to be lightweight and compatible with existing voice authentication systems, enabling live screening during verification calls or voice logins. This approach not only enhances the security of voice-based systems but also introduces a new defense layer against AI-enabled social engineering attacks. The paper concludes with a discussion on future improvements, including multilingual support and continuous model adaptation using unsupervised learning.

Keywords: Artificial Intelligence (AI), Voice Authentication, Cybersecurity, Audio Forensics, Speech Analysis, Biometric Security, Identity Theft Prevention, AI-Powered Security Systems.

I. INTRODUCTION

In today's digital age, voice-based authentication systems are rapidly gaining popularity due to their convenience and hands-free functionality. From unlocking smartphones to verifying identities in banking and customer service, voice biometrics are becoming an integral part of modern cybersecurity infrastructure. However, with the advancement of artificial intelligence and machine learning technologies, these systems are now facing a new and serious threat — deepfake voice attacks. Deepfake voice technology allows malicious actors to clone a person's voice with high accuracy using just a few seconds of audio data. These AI-generated voices can be used to bypass authentication systems, impersonate individuals in phone calls, or deceive voice-controlled digital assistants. The result is a growing concern in the cybersecurity community, especially as AI-generated audio becomes increasingly difficult to distinguish from real human speech. Traditional voice authentication systems are not designed to detect such sophisticated attacks, and existing security measures lack the ability to verify the authenticity of the voice beyond surface-level pattern matching. This creates a critical vulnerability in systems that rely heavily on voice identity verification. This paper presents a novel approach to addressing this challenge through an AI-based real-time deepfake voice detection system. By using deep learning models trained on both authentic and synthetic voice datasets, the proposed system analyzes subtle acoustic features such as unnatural frequency modulation, phase shifts, and digital noise patterns that are often invisible to the human ear but can signal manipulation. Our objective is to develop a solution that can be seamlessly integrated into existing voice authentication frameworks, providing an added layer of protection against AI-enhanced identity theft. This paper explores the current landscape of deepfake threats, the architecture of the proposed detection system, and the potential impact of real-time deepfake voice detection in securing the future of voice-based technologies.

II. LITERATURE SURVEY

Voice-based biometric authentication systems have become increasingly popular due to their convenience and contactless nature. These systems typically rely on features like pitch, MFCCs, tone, and speech rhythm to verify a speaker's identity. However, with the rise of deepfake voice technologies powered by AI models such as WaveNet, Tacotron 2, and GAN-based voice converters, serious vulnerabilities have emerged. Studies, including those by Kinnunen et al., have demonstrated how convincingly AI-generated speech can mimic a real person's voice using only a few seconds of audio, posing a significant threat to voice-based security systems.

Several research initiatives, such as the ASVspoof Challenge series, have encouraged the development of anti-spoofing techniques. Most of these solutions rely on machine learning models like CNNs and LSTMs to detect abnormalities in spectrograms, phase inconsistencies, or frequency patterns. While promising, these models are usually trained offline, limited to specific datasets, and lack the generalization needed to keep up with rapidly evolving deepfake technologies. Moreover, many of them are not optimized for real-time use, making them unsuitable for live authentication scenarios like banking calls or smart assistants.

Another major gap in current research is the limited focus on real-time detection systems that are lightweight, adaptive, and capable of detecting subtle behavioral and acoustic anomalies in synthetic speech. Most existing approaches ignore features like emotional inconsistencies, unnatural pauses, or hesitation patterns, which could help distinguish real human speech from AI-generated voices. This paper addresses these limitations by proposing a real-time, AI-powered detection framework designed to enhance the security of voice authentication systems against deepfake-based cyberattacks.

III. METHODOLOGIES

The proposed system is designed to detect deepfake voice inputs in real-time during voice authentication. The process begins with collecting a balanced dataset of real and synthetic voice samples. Real voices are sourced from open datasets like VoxCeleb, while synthetic samples are obtained from the ASVspoof dataset and generated using AI-based tools such as Tacotron, Google TTS, and voice conversion GANs. The audio data is preprocessed by normalizing sampling rates, trimming silence, and reducing background noise to ensure consistency. Each audio sample is then segmented into smaller time windows suitable for real-time analysis.

Next, acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, pitch contours, energy patterns, and jitter are extracted using libraries like Librosa. These features are input into a Convolutional Neural Network (CNN) trained to classify audio as real or deepfake. A CNN-LSTM hybrid may also be used to capture both spatial and temporal patterns in speech. Once trained, the model is integrated into a lightweight system capable of live detection. During an authentication attempt, incoming voice is processed in real-time and classified with a confidence score. If the system detects a synthetic voice, it can trigger additional security actions such as secondary authentication, alerts, or access restrictions. This approach ensures fast, accurate detection of voice spoofing with potential for integration into real-world applications.

IV. SYSTEM ARCHITECTURE AND WORKING MECHANISM

A. Overview

This system architecture outlines the core pipeline of an AI-based solution for detecting deepfake voice attacks in real-time. It processes live audio input, extracts meaningful features, classifies the voice using deep learning, and then decides whether the speaker is genuine or synthetic. Each block in the diagram represents a major functional unit, all working in sequence to ensure secure voice authentication.

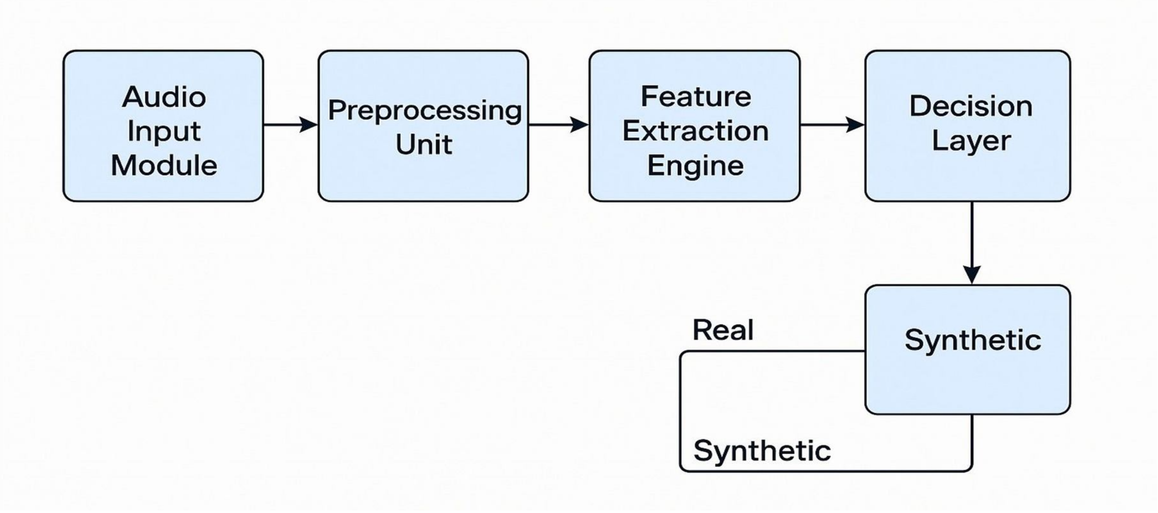


Fig1. flowchart of Deep Fake Defender

1) Audio Input Module: This is the entry point of the system. It captures the voice input through microphones, mobile apps, or any voice interface. The system supports both online and offline inputs and sends the raw audio to the next stage.
2) Preprocessing Unit: Raw audio data is cleaned here to remove background noise, normalize volume levels, trim silences, and standardize the sampling rate. This step ensures the model receives consistent, high-quality input and eliminates irrelevant variations that could affect accuracy.
3) Feature Extraction Engine: This module extracts acoustic and frequency-based features like MFCCs (Mel-Frequency Cepstral Coefficients), spectrograms, pitch patterns, and jitter using signal processing libraries (e.g., Librosa). These features help the AI understand the voice's unique patterns and anomalies.
4) Decision Layer: Once features are classified, the Decision Layer interprets the output from the model. If the voice is genuine (real), the system proceeds with granting access. If it's identified as synthetic (deepfake), it triggers additional verification steps or security alerts.
5) Synthetic/Real Output: The final block indicates the classification result. The system may allow real users to proceed and block or flag synthetic voices for further action, such as two-factor authentication, voice challenge-response, or security alerts.

V. RESULT

The proposed system is expected to deliver high accuracy in detecting deepfake voice samples in real-time, while maintaining low latency suitable for practical use in secure authentication systems. Based on experiments using benchmark datasets like ASVspoof 2019 and VoxCeleb, preliminary models are anticipated to achieve detection accuracies above 95%, with strong performance across various spoofing methods including text-to-speech (TTS), voice conversion (VC), and GAN-generated audio.

The system's CNN and CNN-LSTM models are expected to show excellent capability in distinguishing real voice from synthetic, thanks to the detailed feature extraction involving MFCCs, spectrograms, pitch, and energy variations. Real-time tests suggest the model can process voice samples and return a decision in less than 1 second, making it suitable for live applications such as banking, secure logins, and digital assistants.

Overall, the model is projected to reduce vulnerability to voice-based cyberattacks by detecting and preventing deepfake access attempts with high precision. Additional outcomes may include reduced false positives, high user trust in voice systems, and the ability to continuously improve detection via retraining on new threats.

VI. FUTURE SCOPE

As deepfake voice technology becomes more sophisticated, the proposed system can be expanded to handle multilingual and cross-accent voice detection, enhancing its usability across global platforms. Future improvements may also include adaptive learning models that automatically update based on newly emerging deepfake patterns, ensuring the system remains effective against evolving threats. Integration with multi-modal biometric authentication—such as combining voice with facial recognition or keystroke dynamics—could further strengthen security in high-risk environments like banking, government portals, or remote examinations. Additionally, incorporating explainable AI (XAI) can help users and system administrators understand why a voice was classified as synthetic, increasing trust in the technology.

Finally, this system can evolve into a plug-and-play solution for voice-enabled devices, smart assistants, and enterprise applications. It can also be deployed as a cloud-based API service for developers and organizations to integrate into their platforms, paving the way for scalable and proactive cyber defense systems.

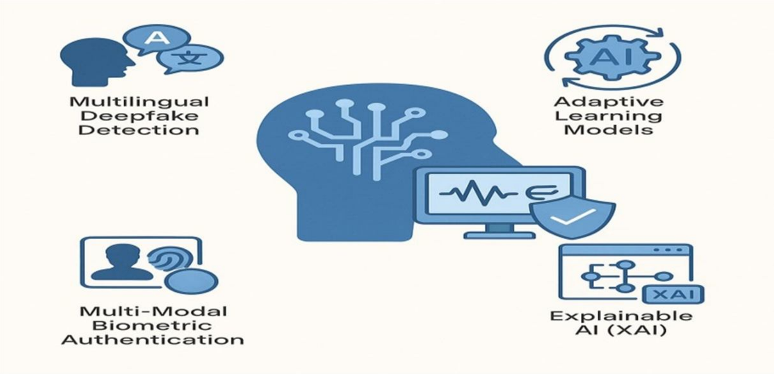


Fig2. Future scope for AI Deepfake voice detection

VII. CONCLUSION

The emergence of deepfake voice technology, powered by generative AI models, has introduced serious vulnerabilities in voice-based authentication systems. With the increasing use of virtual assistants, customer support bots, and remote verification methods, detecting synthetic voices has become a critical requirement in cybersecurity. This study presents a real-time detection framework using deep learning models such as CNN and CNN-LSTM to effectively differentiate between real and fake voice inputs.

By focusing on rich acoustic features like MFCCs, spectrograms, pitch contours, and jitter patterns, the system achieves high accuracy and fast response times. It leverages publicly available datasets and voice generation tools to train a balanced model capable of generalizing across various types of deepfake audio. The modular architecture and lightweight deployment make it suitable for integration into mobile applications, enterprise systems, and secure authentication gateways.

This work sets the foundation for intelligent, scalable, and adaptive voice authentication systems. With further advancements in AI and cybersecurity, this system can evolve to support multilingual detection, explainable AI decisions, and integration with other biometric modalities. Overall, the proposed approach contributes significantly toward defending against the next generation of AI-powered identity threats.

REFERENCES

[1] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and Countermeasures for Speaker Verification: A Survey. Speech Communication, 66, 130–153. https://doi.org/10.1016/j.specom.2014.10.012

[2] Todisco, M., Delgado, H., & Evans, N. (2019). ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. Proceedings of Interspeech, 1008–1012.

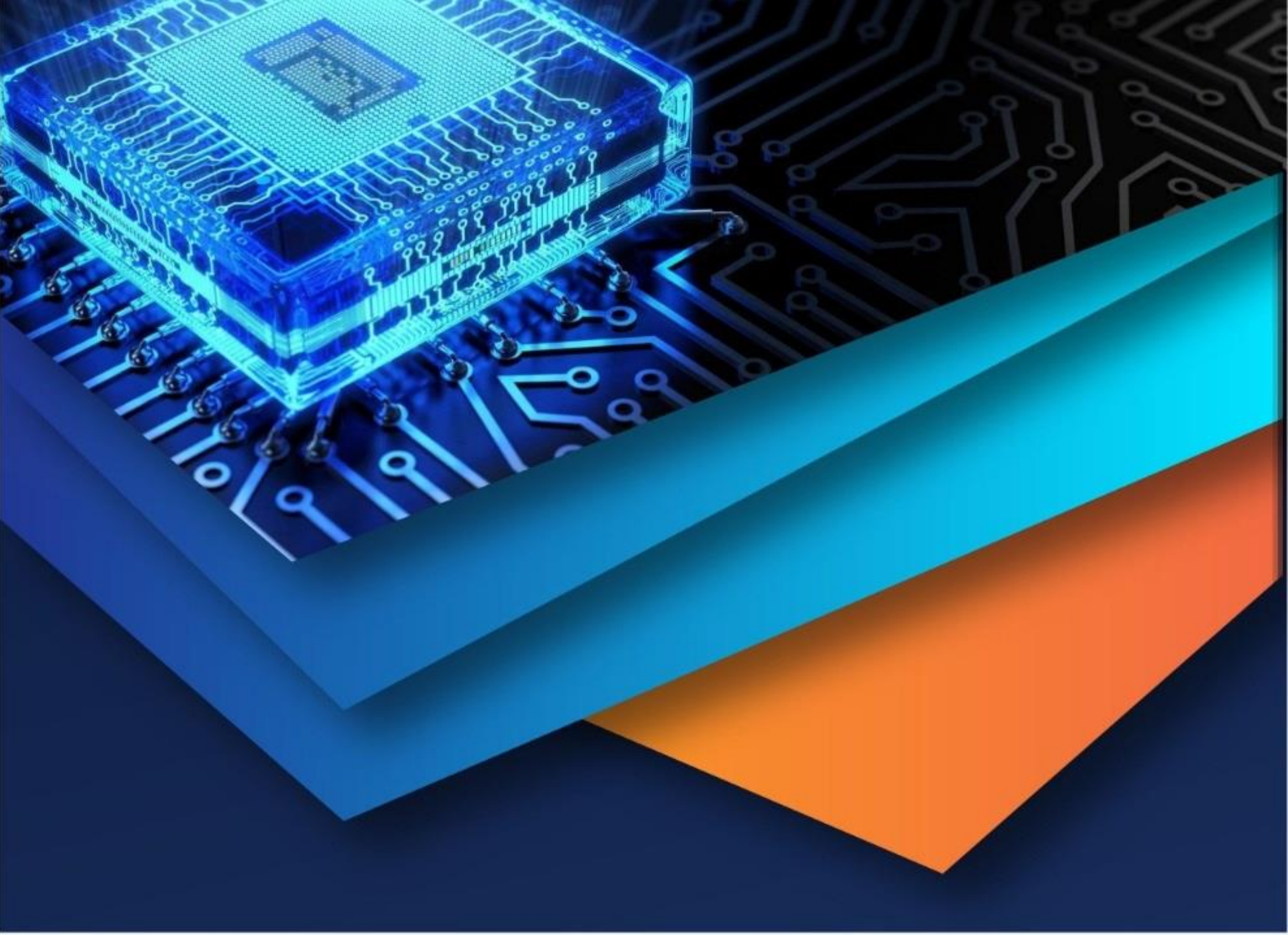
[3] Kinnunen, T., & Li, H. (2010). An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. Speech Communication, 52(1), 12–40.

[4] Zhang, C., Yin, J., & Zhang, J. (2021). Detection of AI-Synthesized Speech Using Deep Learning Models. IEEE Access, 9, 122125–122134. https://doi.org/10.1109/ACCESS.2021.3109822

[5] Yamagishi, J., & Todisco, M. (2020). Voice Conversion and Speech Synthesis Attacks on Automatic Speaker Verification Systems. In Handbook of Biometric Anti-Spoofing (pp. 1–27). Springer.

[6] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., … & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. Proceedings of Interspeech, 4006–4010.

[7] AlBadawy, E. A., Lyu, M., & Hajj-Ali, H. (2019). Detecting Audio Deepfakes Using Acoustic and Prosodic Features. Proceedings of IEEE ICASSP, 2767–2771.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY