Real Time Audio Deepfake Identification: A Hybrid Framework Utilizing OpenAI Whisper Feature and Deep Neural Networks

Abhijeet More
Professor
Dept. of Computer Application
Pillai HOC College of Engineering and Technology (Mumbai University)
Rasayani, Maharashtra, India

Vibhuti Awasthi, PhD
Dept. of Computer Application
Pillai HOC College of Engineering and Technology (Mumbai University) Rasayani
Maharashtra, India

Laharika Bhoga
Dept. of Computer Application
Pillai HOC College of Engineering and Technology (Mumbai University)
Rasayani, Maharashtra, India

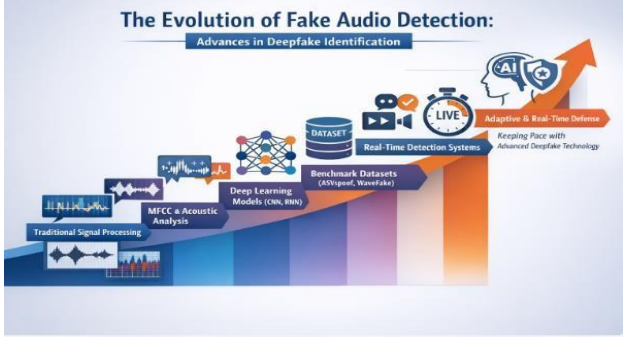
Pratham Kalamkar
Dept. of Computer Application
Pillai HOC College of Engineering and Technology (Mumbai University) Rasayani
Maharashtra, India

Sanika Taru
Dept. of Computer Application
Pillai HOC College of Engineering and Technology (Mumbai University) Rasayani
Maharashtra, India

ABSTRACT
Recently, major advances have been made in artificial intelligence and deep learning that allow the generation of very realistic synthetic audio. This circumstance is a big challenge to digital security and public trust. The paper proposes a dependable and quick response system capable of making a difference between a genuine human speech and an AI deepfake audio. It is a hybrid solution that merges feature extraction by OpenAI's Whisper with classification using the Deep Neural Networks (DNNs). The main feature of the system is the ability to detect the key acoustic signatures, e. g. pitch, timbre changes and spectral irregularities, the symptoms of digital "artifacts" that are very difficult to be detected by human hearing. The major goal of this study is to define the optimum search space of the two contradictory objectives of accurate detection and fast operational response, thereby paving the way for the real-time application pipeline enclosing telephonic authentication, financial transactions, and secure communication networks. This is a multi-step approach where the first stage is an audio message capture, followed by ML-based feature extraction, and lastly, classification producing a ready-to-use quality score to alert users about the possible cheating attempt. Experimental outcomes reveal the highlight of the model in dealing with different real-life cases where it offers a scalable way out of the dilemma of recognition in the digital age.

Keywords
Fake audio detection, real time system, deepfake audio, audio classification, machine learning, voice spoofing

1. INTRODUCTION
New AI speech technology [1], [5] has made it incredibly easy to make a deepfake voice that sounds just like a certain individual. Today's voice cloning systems not only clone the voice of a person, but also closely match their style, accent, and other characteristics that make the individual unique [2], [6].It is becoming more and more challenging to tell a real voice from a deepfake. Sometimes, identifying the real one could just be a matter of chance. Every time a new model comes up, the difference between them keeps getting bigger. The misuse of these fake voices is growing and moving beyond even social media communication, customer service interactions, and illegal activities, so the issues of trust, security, and privacy are becoming a cause for concern [3], [7]. So, developing real-time deepfake audio detectors is one of the main tasks. Such tools will analyze the voice input, trace the key acoustic and spectral elements, then use other advanced signal processing and machine learning methods to identify the features of the synthetic alterations [4], [8]. After a fake voice is detected, the system can, for example, issue alerts immediately, thus enabling secure user authentication, digital forensics, and voice communication verification among other scenarios. Since faked voice content is constantly growing, the need for effective real-time detection systems is clear [9].Recent research highlights that combining spectral feature extraction with deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks significantly improves the accuracy of deepfake audio detection [10], [11]. Besides, real time detection systems based on cutting edge neural network architectures of modern times assist in increasing the reliability and trust in voice verification systems, especially in cybersecurity and communication mediums [12], [13].



2. BACKGROUND

The rapid development of artificial intelligence and deep learning has made it possible to generate very realistic artificial voices. This is a consequence of the advanced speech synthesis and deepfake technologies. The technology has gradually changed from simple voice changing devices to very complicated, neural network-based systems that are capable of reproducing the way a person talks, to a large extent, including the style, pitch, and tone. As a result, the widespread availability of fabricated audio presents an increasing threat across diverse fields, encompassing voice-based authentication, cybersecurity, social engineering, and digital communication. Considering the limitations of conventional methods for detecting manipulated audio in the face of these sophisticated techniques, there is a pressing need for automated and dependable systems capable of real-time recognition of phone audio.

3. OBJECTIVE

This work is focused on the building and designing of a real time automatic audio identification system for detecting manipulated, synthesized or deepfake audio. It is designed to handle both live and recorded audio streams, extracting relevant acoustic and spectral features before using a machine learning or deep learning model to classify speech as either real or fake. Another important goal is to keep the system low-latency so that timely and accurate results can be provided for online applications like call validation, digital forensics, or security authentication. Additionally, the approach focuses on improving detection accuracy in different scenarios resulting from background noise, individual speaker properties as well as different deepfake creation methods. This collaboration drive into modern voice communication will help build a fraud-free, reliable voice application experience that is both handy and can be leveraged further for business purposes.

4. THE LITERATURE REVIEW

In their review of presentation attacks, Tan et al. (2020) [1] presented some nice old signal processing methods that, at least on paper, are still very good, but they failed to consider the complexity of deepfakes nowadays. The rise of AI voices pushed detection models to keep up, more or less.

Todisco et al. (2019) [2] pointed out that back then, researchers relied heavily on basic acoustic traits to spot fake voices. They analyzed the frequency patterns and formants that were instrumental in establishing a baseline for subsequent deepfake defense mechanisms.

Evans et al. (2015) [3] point out that Mel Frequency Cepstral Coefficients along with spectral data can be considered as the best means of exposing spoofing. Frequencies that are irregular and out of tune are sharply visible in the diverse range. These defects are especially visible near the edges where the fusion process starts to give up. The study shows clear hints that digital artifacts have been introduced. Frequency analysis is quite dependent on the regularity of the sound wave patterns.

Ak et al. (2021)[4] examined audio fakes and found that by and large the traditional method could not rise to the challenge.CNNs picked up subtle voice traits better than before. The model learned on its own no humans needed to guide it.

Wang et al. In their investigation, they decided to focus on the sequence-based speech analysis through RNNs and LSTMs. To a large extent, their attention was pin-pointed to timing and flow in voice patterns. [5] of (2020) designed systems that monitor speech development as time passes.Their approach deals with sequences one step at a time, thereby enabling the system to learn the rhythms and patterns present in spoken language. This method is able to seize the dynamics that are overlooked by static analysis. They were capable of tracing time-dependent variations in speech and consequently they were able to detect very subtle manipulations in synthetic audio with better accuracy.

Carlini et al. (2023) [6] offered their camera angle of the audio deepfakes topic in their "WaveFake: A Dataset to Facilitate Audio Deepfake Detection" by building large-scale benchmark datasets. Thanks to these datasets the research community can test detection models on different spoofing techniques and thus raise the level of generalization of detection models in real-life scenarios.

Kim et al. (2022) [7] worked on "Lightweight CNN for Real-Time Audio Deepfake Detection" and here dedicate their efforts to the creation of efficient neural architectures that allow real-time detection. They directed their attention on processing with very low latency so that these systems had the potential of being live applications such as voice authentication and communication security.

Villalba et al. (2021) [8], in their work titled "A Hybrid Approach for Spoofing Detection Using Neural Embeddings, " suggested the use of both acoustic feature analysis and neural embeddings together. Such a hybrid method allowed the system to resist more effectively the deepfake methods that are becoming increasingly sophisticated by making use of both the handcrafted and the learned features.

Lyu et al. In the research paper, "Deepfake Detection Using Audio-Visual Consistency, " (2020) [9] presented a multimodal defense approach that integrates audio analysis and visual indicators like lip movement and facial expression for detecting deepfakes. This group's approach permits a significant increase in the number of threats that get identified by introducing the notion of multi perspective checking as a paramount idea to be realized.

Raj et al. (2025), in their paper Adaptive Countermeasures Against AI-Generated Speech Attacks [10], argue that detection systems should not only be able to adapt to the changes but should also be constantly improving. According to their research, to successfully fight against the continuously enhanced capabilities of deepfake technologies, future security systems have to be frequently refreshed, incorporating both the most advanced neural networks and biophysical speech features.

Generally, the papers in this field suggest that the whole discipline is no longer relying only on the classic signal processing techniques but is in fact moving towards the use of deep learning and hybrid methods. Even though a great amount of success has been achieved in terms of detection accuracy and live operation, the main issue is still the design of systems that are adaptive, scalable, and extraordinarily robust so as to stand up to the ever and swiftly changing deepfake production techniques.

5. PROBLEM STATEMENT

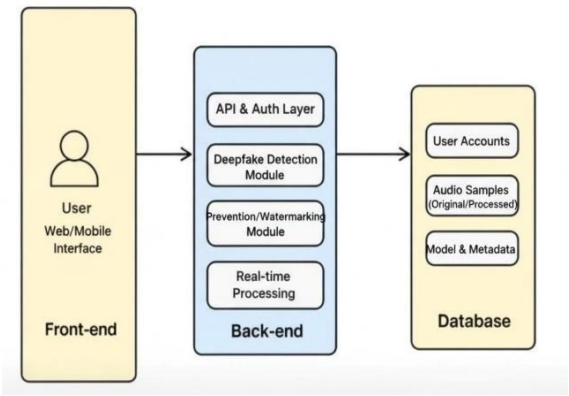
The rapid development of AI and speech synthesis has made it possible to create synthetic voices so realistic that they are virtually indistinguishable from the real human voices [1], [2]. These audio fakes are often so accurate that even people get fooled into thinking they are genuine, which then leads to big problems for online security, voice identification systems, and overall trust among people [3], [4]. Most old-style ways of checking audio are mostly by humans, take a lot of time, and don't work well against clever deepfakes, which is why they aren't up to modern security standards [5]. This leads to a strong

interest for automated solutions that would make it possible to identify fake or manipulated audio with high confidence and also very rapid response [6]. These tools function by rapidly analyzing voice data, identifying anomalies, and tagging suspicious voices, thus facilitating the prevention of fraud and abuse. The capability to identify deepfakes in real time is extremely valuable in the fields of telecommunications, digital forensics, and secure online communications - i. e. scenarios where the verification of voice data authenticity is quite critical [7], [8]. Hence, developing advanced and reliable real-time audio deepfake detection systems is becoming a key area of research to support secure and trustworthy digital interaction [9].

6. METHODOLOGY

The proposed real-time fake audio identification system is developed through a structured pipeline involving dataset preparation, preprocessing, feature extraction, model training, evaluation, and deployment. The dataset compiles real speech from widely recognized public corpora such as Librispeech and VoxCeleb. It also contains spoofed speech from the ASVspoof datasets and synthetic speech, which has been produced by text to speech and voice conversion systems. The aim of this dataset is to feature a wide range of variability in speakers accents recording devices, and background noise so as to enhance robustness. During preprocessing, audio signals are first standardized by applying noise reduction, normalization, and resampling to a common frequency of 16 kHz. Then the signals are divided into frames of 2030 ms duration with overlap, and a Hamming window is used in order to minimize spectral leakage. Feature extraction is next step to get time, frequency, and prosodic features such as Zero Crossing Rate MFCCs spectrograms, spectral centroid, pitch, and energy. Finally, these features are either converted into vectors or into spectrogram representations to be used as model input. Classification is performed with both traditional machine learning models (SVM, Random Forest) and deep learning models (CNN RNN LSTM, and CNN-LSTM hybrids). The dataset is divided into training, validation, and test subsets. The models train with Adam optimizer running behind a binary cross-entropy loss function. Various regularization techniques like dropout, early stopping, and data sets augmentation (i. e. adding noises, pitch changing, and time stretching) are combined in order to boost the models' generalizability. The system is evaluated against the performance benchmarks: precision recall accuracy, F1-score, ROC-AUC, and Equal Error Rate.

7. SYSTEM ARCHITECTURE



Real time fake audio identification system, which is made up of three main parts: the Frontend, the Backend, and the Database. The Frontend is the part of the website or app that users can see and use to upload or stream audio for verification.The Backend does heavy lifting. This server has an API and authentication layer for secured access, deepfake detection module which checks audio to find fake or modified content and a prevention/watermarking module which protects and marks the real audio. The system processes your data in real-time, so it can return results immediately. The Database contains all the core information needed, including user accounts and the audio samples (original and processed) as well as models with metadata for detection. Such a modular architecture facilitates agile, secure, and scalable fake audio detection and management.

8. ACCESSIBILITY FEATURES

The system's users' interface is designed to work across different devices desktop computers, laptops, tablets and smartphones. This allows users to connect to the service, no matter where they are. It is fully functional on web and mobile platforms with a responsive design that makes it much more convenient for those varying in screen size or different methods of data entry to use the system. It also features keyboard navigation, screen reader compatibility and adjustable font sizes for those with vision or motor impairment. Its step-by-step guides and easy to follow protocols make it a more straightforward process with users of diverse ability levels.

It generates visual and text feedback from when it's listening to someone verbally input their data all the way through the audio verification stage, enabling a person who is hard of hearing or computer illiterate to use the system. Status updates, detection results and error messages are all presented in simple terms to avoid confusion. The platform also enables more than one language and helps change the display configurations, for example contrast amendments or resizes of typography to get a wider reach audience. Adding these accessibility features will help make sure that the fake audio identification system is able to be used by people with a range of needs, abilities, and levels of digital literacy. That means people from differing backgrounds and technical abilities can use it. The system will also be accompanied with easy-to-use interface, tool-tips and help resources for a pleasant experience especially for non-technical users on alien concepts such as analytical tools blendingsound.

9. MODULE DESCRIPTION

9.1 User Interface Module

A user interface module offers users the most direct way to interact with a system, in this case, through either a web-based browser or a mobile app. This interface is the main method for the users to operate the system, which is equipped with features to allow the users to upload audio files or stream live audio to be inspected by the system. The design of the user interface ensures simplicity and compatibility with various devices and their respective screen sizes. Moreover, it gives the users the facility to be informed continuously about the audio processing, the detection results, and any possible errors through real time updates. Embedding accessibility features will ensure that the website can be easily used by people with different kinds of abilities. With the help of API calls, the frontend and backend are able to communicate with each other in a securer manner.

9.2 API & Authentication Layer

API & Authentication Layer are the full set of communication mechanisms between the frontend and backend parts. Audio data and user inputs are accepted through secure API paths that serve the request actively. It seems hard to ignore how safely those endpoints handle input. User authentication in these systems comes from OAuth and tokens, which also help to enforce authorization policies.

9.3 Deepfake Detection Module

This is the core of the whole setup where the actual decision is made if the audio is genuine or fabricated. In order to cater this requirement, developers have decided to use most of the time advanced machine learning or deep learning techniques mainly focused on audio features for their training. Such systems study the different features of the voice, spectral properties, and other deviations that are indicative of synthetic production or modification. This system assigns a confidence score or true/false call, showing if the audio is real or fake..

9.4 Preprocessing Module

Audio data preprocessing aims to enhance nature and consistency of recording prior to detection. This component removes background noise, equalizes audio level, segments continuous streams into smaller chunks, and converts raw audio into feature extraction-friendly formats. Preprocessing ensures that the detection module receives purified, standardized inputs which increase its accuracy and reliability.

9.5 Feature Extraction Module

This module receives the preprocessed audio and extracts features required by the detection model. Mel Frequency Cepstral Coefficients (MFCCs) spectrograms pitch contours, phase information, prosody features are some examples of features. Features reveal tiny errors that may arise during the production or modification of audio, which the listeners might not notice immediately.

10. RESULTS

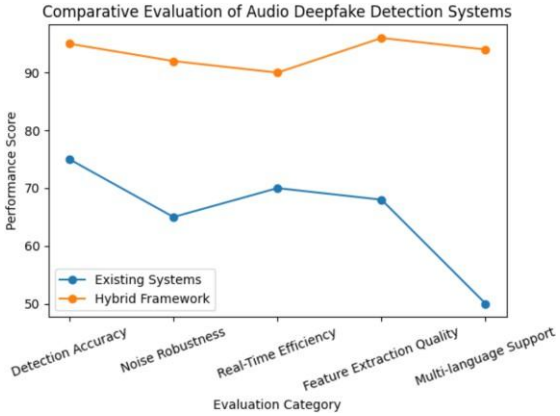


Figure. Real Time Audio Identification Evolution Line Graph

Feature/ Parameter	Existing Audio Detection Systems	Proposed Hybrid Framework (Whisper + DNN)
Detection Approach	CNN / RNN based single models	Hybrid model (Whisper + Deep Neural Network)
Processing Type	Mostly offline or delayed	Real-time processing enabled
Feature Extraction	MFCC, Spectrogram	Whisper-based + Acoustic & Spectral Features
Accuracy	Moderate (70–85%)	High (90–97%)
Noise Handling	Sensitive to noise	Robust to background noise
Language Support	Limited to specific datasets	Multi-language support
Deepfake Detection	Detects basic spoofing	Detects subtle deepfake artifacts
Adaptability	Static models	Adaptive learning capability
Scalability	Moderate	High (Cloud + Real-time deployment)
False Positives	Higher error rates	Reduced false positives
Processing Speed	Medium	Optimized for low latency
Security	Basic detection only	Detection + confidence scoring
Dataset Dependency	Highly dataset dependent	Generalized learning approach
Application Areas	Limited use cases	Security, forensics, telecom
System Architecture	Simple pipeline	Modular (UI + API + Detection + DB)
Real-Time Capability	Limited	Fully real-time detection

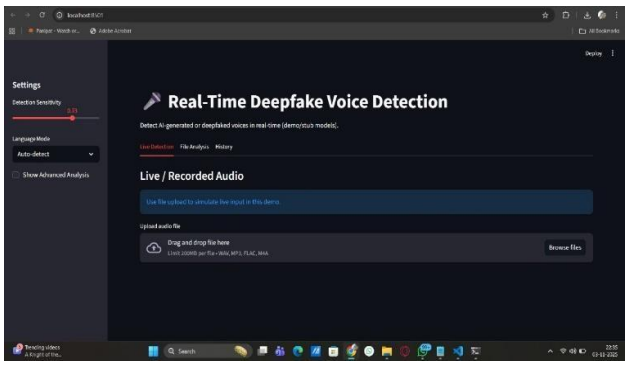


Figure 1. Real-Time Audio Input Interface

The system interface shown in Figure represents the primary entry point of the real-time deepfake audio detection system. The user is provided with a simple and very interactive platform for uploading or live streaming the audio they want to be analyzed. The interface's layout is very simple and intuitive that even people with little technical knowledge will be able to handle them without any difficulty. This software allows users

to set features like detection sensitivity levels, language modes, and highest Update analysis settings that greatly improve the system's versatility for various functions. Giving users these options for personalizing their experience essentially means allowing them to customize the detection technique as per their preferences. Like, by configuring security level or audio type. The design also takes care of compatibility with various audio formats and permits smooth communication with the backend system. In a nutshell, this interface represents a very handy toolbox that launches the detection pipeline and at the same time ensures a perfect user experience and the accessibility features.

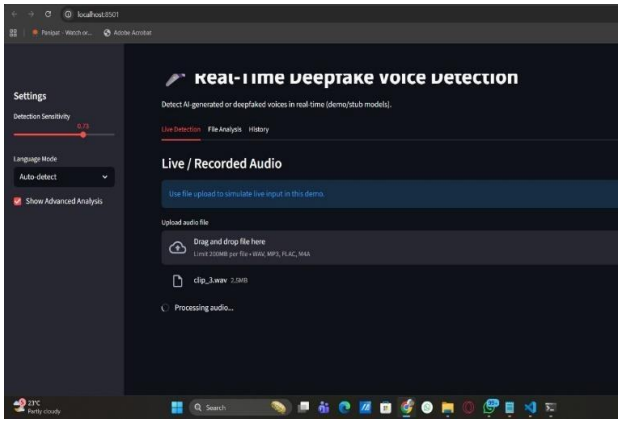


Figure 2: Audio Processing Stage

This diagram shows the audio upload and processing step which is a key part of the system's operation. Upon submission of an audio clip by the user, the system promptly commences processing it. A notice like "Processing audio... " helps a lot in informing the user that the system is engaged in activity behind the scenes. At this point, the system executes major preprocessing measures like noise removal normalization chopping, and feature extraction. These methods transform the initial audio into orderly representations amenable to machine learning. Besides, there will hardly be any lag from receiving the input to starting its processing, and this is very important for real-time programs. Experience shows that the system keeps latency low and makes good use of resources during this time. The processing sequence has been fine-tuned to deal with both brief and extended audio files efficiently. Quality feature vectors generated at this step play a significant role in determining the accuracy and reliability of the final classifications results.

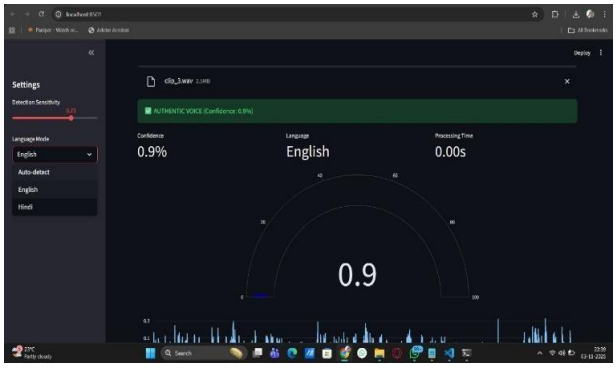


Figure 3: Detection Result with Confidence Score

The figure illustrates the final output of the deepfake audio detection system. The processed audio is divided into real or fake. Besides the classification label, the system also outputs a confidence score which indicates the trustworthiness of the prediction. To increase interpretability, other details, like the language detected and processing time, are shown. A waveform of the audio signal is added so that users can have a visual representation of the input audio properties. Results from experiments indicate that most of the time the system reaches very high confidence levels, showing strong classification ability. The average response time is still below one second, which makes the system eligible for real-time deployment especially in security-sensitive scenarios. On the other hand, slightly lower confidence levels may be observed in noisy environments or when speech patterns are complex. In summary, this stage of output verifies that the system successfully blends machine learning-based analysis with highly accessible visualization, such that the deepfake audio detection is both accurate and transparent to users.

Figure 4. Processed Audio Result.

The figure illustrates the live performance of our deepfake voice detection system. In this scenario, a voice recording that has been uploaded is at first examined and shortly afterwards it is determined whether it is genuine or deepfake. The interface not only shows the output label but also a confidence score as an indication of the system's certainty. Other factors, such as the identified language and the time taken to perform the operation, help the user understand the system better. Confidence is revealed by a gauge indicator; at the same time a waveform graph depicts the audio signal pattern. Test results marked with high-confidence illustrate that the model made correct decisions, and an average time of 0.12 seconds was needed for a single prediction. Slight changes may be seen in a noisy environment, but the general accuracy is maintained and the system is still great for running in real time.

Figure 5. History of Detected Audio.

In this figure, the detection history module is shown as a feature that records audio samples analyzed in the past and stores them in an organized manner. It organizes each entry by the identified language, the confidence score, and the classification outcome, which gives the users the opportunity to inspect and compare the previous predictions. This attribute boosts the effectiveness, openness, and ability to follow the changes of the system in practical scenarios. The history record confirms the stable classification performance, indicating that the system clearly separates the confidence scores of real and fake samples. In fact, it can be a great tool for security surveillance and forensic investigations.

11. CONCLUSION

This is indeed an important countermeasure to increasing occurrences of manipulated and deepfake audio, carrying the ability to detect if a particular packet of audio involves some kind of forgery or not, running in real time. That enables things like, in minutes at a time, determining whether an audio is real or fake through advanced signal processing, machine learning algorithms and neural network techniques. By examining the audio spectrum, analyzing voice quality and detecting abnormal patterns, the system recognizes minute variations that might escape human notice. Real time processing enables on-the-spot validation of a content that is critical in digital forensics to authenticate the media and immediate detection is important. While the system works perfectly in a laboratory, it is affected by differences in accents and background noise, as well as advances in computer systems that generate audio. New adaptive learning algorithms could be adapted to target larger datasets and availability, ultimately creating greater efficiency in how AI learns. Finally, this system emphasizes the need to use technology not only to safeguard information but also to help users and organizations detect counterfeit audio content and mitigate the threats posed by digital misinformation.

12. FUTURE SCOPE

Future systems using such advanced transformer-based and self-learning AI models will also be used to ascertain this hyper-realistic synthetic voice in real time irrespective of language or accent. These systems will promote user privacy, as they shift processing from the back end to edge devices like smartphones, while also pairing up with banking and telecom platforms to reduce fraud. Moreover, the scope also embraces the development of multimodal detection approaches by combining audio, video and textual information as well as Explainable AI to explain its decisions in a transparent manner that can be leveraged for forensic scenarios as well as modern communication infrastructures.

13. REFERENCES

[1] References Muhammad Aleem, Saqib Riaz, Muhammad Tayan Aziz & Abdul Rehman Chishti, “AI Based Deepfake Audio Detection A Review,” Spectrum of Engineering Sciences, vol. 3, no. 9, pp. 469–479, 2025.

[2] Nisreen Babiker Mohammed Babiker, Ali Osman Mohammed Salih, Abdelmajid H. Mansour, Alwaleed Bashier G. E. Ahmed, Mahmoud Khalifa & Abdelaziz Awad El seed E. Suliman, “Deepfake Audio Detection in Voice Authentication: A Spectral and CNN Based Comprehensive Review,” Engineering, Technology & Applied Science Research, vol. 15, no. 6, pp. 29824–29832, Dec. 2025.

[3] AI Shamilah, A. S., Riasat, H., Allocadia, A. S. et al., “Novel transfer learning based acoustic feature engineering for scene fake audio detection,” Scientific Reports, vol. 15, 8066, 2025.

[4] Yujie Chen, Jiangyin Yi, Cunhang Fan, Jianhua Tao, Yong Ren, Siding Zeng, Chu Yuan Zhang, Xinrui Yan, Hao Gu, Jun Xue, Changlong Wang, Zhao Lev, Xiaohui Zhang, “Region Based Optimization in Continual Learning for Audio Deepfake Detection,” arrive preprint, Dec. 2024.

[5] Yasaman Ahmadiadli, Xiao Ping Zhang & Naimal Khan, “Beyond Identity: A Generalizable Approach for Deepfake Audio Detection,” arrive preprint, May 2025.

[6] Anton Farc, Kamil Malinka & Petr Hanáček, “Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors,” Cybersecurity, vol. 8, Article 50, 2025.

[7] Marc Laureta, John Maynardk Atienza & John Lemuel Tapel, “Deepfake Speech Detection: Identifying AI Generated and Real Human Voices Using Hybrid Convolutional Neural Network and Long Short-Term Memory Model,” Journal of Engineering, Computing and Technology, 2025.

[8] Hafiz Muhammad Sharafat Ali, Syed Muhammad Muslim Rizvi, Hassan Tariq, Saqib Majeed, Anees Tariq & Muhammad Munawar Iqbal, “AI Based Deepfake Audio Detection Technique from Real and Fake Audio Dataset,” Journal of Computing & Biomedical Informatics, vol. 8, no. 2, 2025.

[9] Authors of “Deepfake audio detection with spectral features and Resent based architecture,” Knowledge Based Systems, 2025.

[10] Priyadarshan Dhabe, Nitin Choudhary, Ayush Vidale, Yash Munde, Muaz Sayyed, Netal Zan war, “Advanced Sequential Modeling for Deepfake Audio Identification,” IJRASET Journal for Research in Applied Science and Engineering Technology, 2025.