Deepfake Speech Detection – A Literature Review

Kishor Chandrapalan

Department of Artificial Intelligence Reva University Bengaluru, India

DOI: https://dx.doi.org/10.51584/IJRIAS.2025.10100000118

Received: 25 October 2025; Accepted: 31 October 2025; Published: 12 November 2025

ABSTRACT

Deepfake audio technology and its potential for misuse represent significant challenges in the realms of information integrity, identity protection, and public trust. This paper offers a comprehensive exploration of the detection methods for deepfake speech and their implications. First, we examine the emerging threats of AI-driven scams, particularly the use of Large Language Models (LLMs) in automating voice-based fraud, including phone scams and virtual kidnapping. With the rise of these technologies, voice cloning can be exploited to deceive victims into revealing sensitive information, undermining public safety and trust.

Alongside this, we analyze the state of deepfake detection technologies through a systematic review of ten key studies, focusing on common feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram-based features, pause characteristics, and advanced deep learning methods. MFCCs remain foundational, complemented by newer techniques like spectrogram analysis and deep learning models, yet challenges persist in dataset variability, generalization, and adversarial robustness. Furthermore, ethical concerns surrounding the potential misuse of deepfake technologies—such as in spreading misinformation or violating privacy—highlight the need for a more robust ethical framework. Future research must prioritize creating hybrid detection systems that combine deep learning with real-time operational capabilities, all while considering the ethical and adversarial aspects of this evolving technology. This dual analysis aims to guide the development of more effective, ethically sound detection systems for deepfake speech and AI-driven scams.

This research calls for interdisciplinary collaboration to address both the technical and ethical challenges posed by these advanced AI systems, emphasizing the necessity for diversified datasets, real-time detection, and robust defenses against adversarial threats.

Keywords — Deepfake audio, deepfake detection, voice cloning, Mel-Frequency Cepstral Coefficients (MFCC), spectrogram analysis, deep learning, adversarial robustness, ethical concerns, misinformation, privacy violation, real-time detection, hybrid detection systems, Generative Adversarial Networks (GANs), automated fraud detection

INTRODUCTION

The advent of deepfake audio technology has introduced significant challenges to information integrity, security, and personal identity. Deepfake speech, generated by artificial intelligence (AI) models, has become a growing concern due to its ability to create human-like voices used for malicious purposes such as identity theft, misinformation, and social manipulation [1], [2]. As these synthetic voices become more sophisticated, the need for robust detection methods to protect individuals and systems reliant on voice recognition for security and authentication has become critical [3].

Voice authentication systems, widely used in banking and personal assistants, are particularly vulnerable to spoofing attacks involving deepfake audio [4]. This has led to the development of various detection methods utilizing signal processing techniques, machine learning (ML) algorithms, and deep neural networks [5], [6]. Features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, formants, and statistical properties have been extensively used to train classifiers that distinguish between human and machine-generated speech [7], [8].

Despite these advancements, challenges remain in creating detection systems that can generalize across diverse datasets and real-world environments, especially when exposed to new or unseen types of deepfake generation

methods [9]. Real-time detection has become increasingly important, as detection systems must function efficiently without compromising performance [10]. Ethical concerns also arise regarding the misuse of deepfake technology for illegal activities, prompting discussions on developing systems that not only detect but also prevent malicious use of synthetic media [6].

The increasing sophistication of deepfake technologies has led to the development of various deepfake speech detection systems. These systems incorporate signal processing, ML models, and deep learning approaches to tackle the challenges of identifying synthetic speech. Recent work [5], [6] explores the use of advanced algorithms such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for detecting deepfake audio. Other studies have highlighted the importance of real-time detection [4], [8], where detection systems must maintain high accuracy and performance under real-world conditions. Moreover, hybrid techniques combining traditional methods with deep learning have been proposed as effective approaches to enhance detection accuracy and robustness [8]. Researchers continue to explore systems capable of adapting to new deepfake generation techniques, including those utilizing Generative Adversarial Networks (GANs) [9], [10].

In conclusion, ongoing efforts focus on enhancing detection systems that are not only accurate but also ethical and resilient to future advancements in deepfake technology. These systems aim to mitigate the risks posed by deepfake audio in various domains, ensuring the protection of identity, security, and trust in voice-based technologies

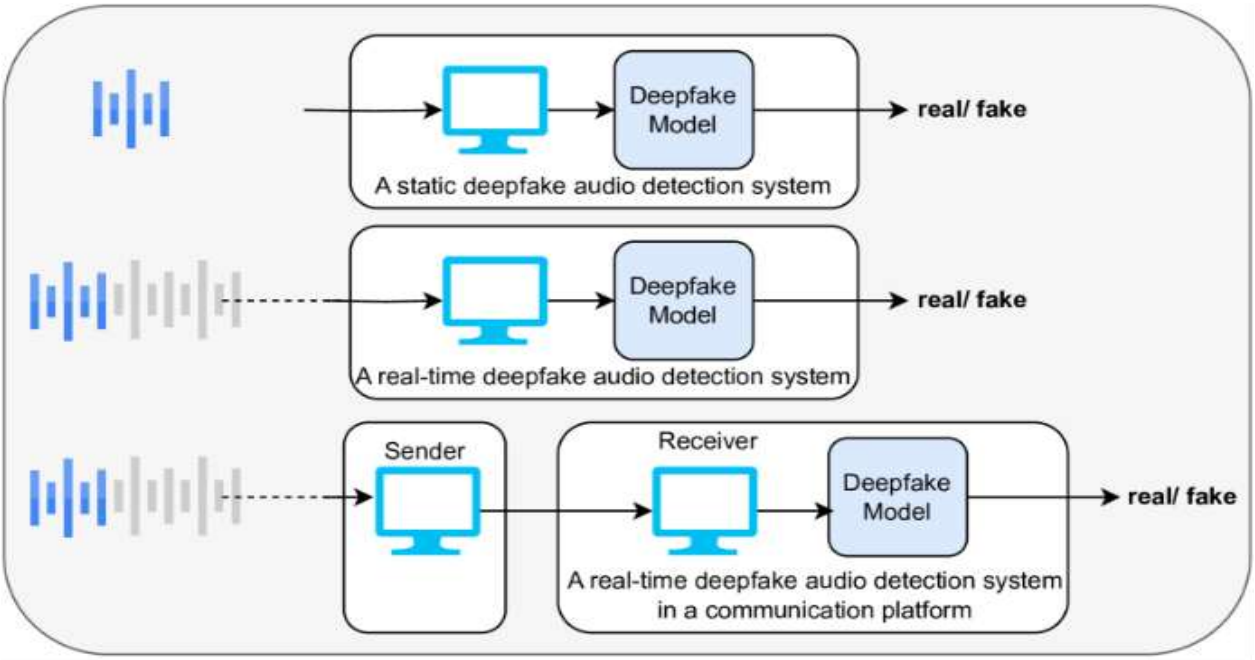
Deepfake Speech Detection Systems

Design Characteristics

Deepfake speech detection systems are designed to identify synthetic audio generated by AI models, distinguishing it from human speech. A key characteristic of such systems is their feature extraction process, which involves capturing a wide array of features that help differentiate real from synthetic speech. These features often include Mel-Frequency Cepstral Coefficients (MFCCs), pitch, formants, and other statistical properties like skewness and kurtosis, which are fundamental in speech analysis [1][5]. Temporal features, such as prosody and speech dynamics, play a crucial role in identifying the subtle variations present in human speech that deepfake audio typically lacks [8][9]. The integration of both time and frequency domain features ensures that the system captures a more comprehensive set of speech characteristics, helping it better differentiate synthetic from genuine speech. The system also needs to be robust across various environmental conditions, including different accents, languages, and noise environments, making generalization essential [9][10]. To ensure high performance in real-world applications, deepfake speech detection systems must operate in real-time without compromising accuracy. This is particularly challenging for systems involved in voice authentication or live speech monitoring [4][6]. Advanced machine learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are integral to these systems, as they excel at capturing both spatial and temporal dependencies in speech data, making them highly effective for detecting deepfake audio [7][8]. CNNs, for instance, are adept at extracting features from spectrograms, while RNNs capture the temporal relationships present in speech. Furthermore, integrating adversarial training is critical to making detection systems resilient to evolving deepfake generation methods. Given the rapid advancements in deepfake technology, detection systems must remain adaptable and able to withstand new types of synthetic speech created by Generative Adversarial Networks (GANs) [7][9]. Another important characteristic is the system's explainability, particularly in sensitive applications where transparency is crucial for ensuring trust. Models must be interpretable, providing clear insights into how decisions are made, which can help mitigate risks associated with false positives and support ethical considerations [8][10]. Finally, deepfake speech detection systems must be scalable and lightweight for real-time deployment on edge devices, such as smartphones and voice assistants, where computational resources are limited. This requires optimization through techniques like model pruning and quantization to ensure that the system performs efficiently even with large-scale datasets [5][6]. Additionally, the systems should support continuous learning to stay updated with emerging deepfake techniques. Online learning approaches and regular model updates ensure that detection systems remain relevant and effective as new deepfake methods evolve [6][9]. By integrating these features and characteristics, deepfake

speech detection systems can become reliable, scalable, and efficient tools for combating the growing threat of AI-generated audio maintaining robust performance under adversarial conditions.

Fig.1 A block diagram of various deepfake audio detection systems



Structured Pipeline

A deepfake speech detection system follows a structured pipeline that involves various stages of data processing, feature extraction, model training, and classification. The first step in the pipeline is data collection, where a dataset of both real and fake speech samples is gathered. These datasets may come from public sources, such as ASVspoof or proprietary collections of AI-generated speech, and the data may be augmented using techniques like pitch shifting, speed alteration, or adding background noise to increase the diversity of the dataset and improve model robustness [7][8].

Next is the preprocessing phase, which includes segmentation, where the raw speech is divided into short segments or windows (typically 20–40 ms). This helps in capturing the temporal dynamics of speech, which are essential for differentiating between human and synthetic speech [5][9]. Additionally, normalization and scaling are applied to the extracted features to ensure consistency and prevent any feature from dominating due to scale differences [5][8]. Noise reduction techniques, such as bandpass filtering or spectral subtraction, are used to remove background noise and focus on the primary features of the speech signal [6][9].

In the feature extraction stage, key acoustic features are extracted. Common features include Mel-Frequency Cepstral Coefficients (MFCCs), which represent the power spectrum of speech and are widely used for speech recognition and deepfake detection [7][8]. Features related to pitch and formants are also used to capture the harmonic content of speech and distinguish between human and machine-generated voices [6][9]. Additional statistical features such as skewness, kurtosis, and spectral flatness are employed to capture subtle artifacts inherent in deepfake speech [5][10]. Temporal features are also analysed to detect recurrent patterns and transitions between speech segments, which are often indicative of synthetic speech generation methods [8][9].

The extracted features are often subjected to dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) to reduce the feature space's dimensionality, making the training process more efficient and less prone to overfitting, especially with high-dimensional data [2][7].

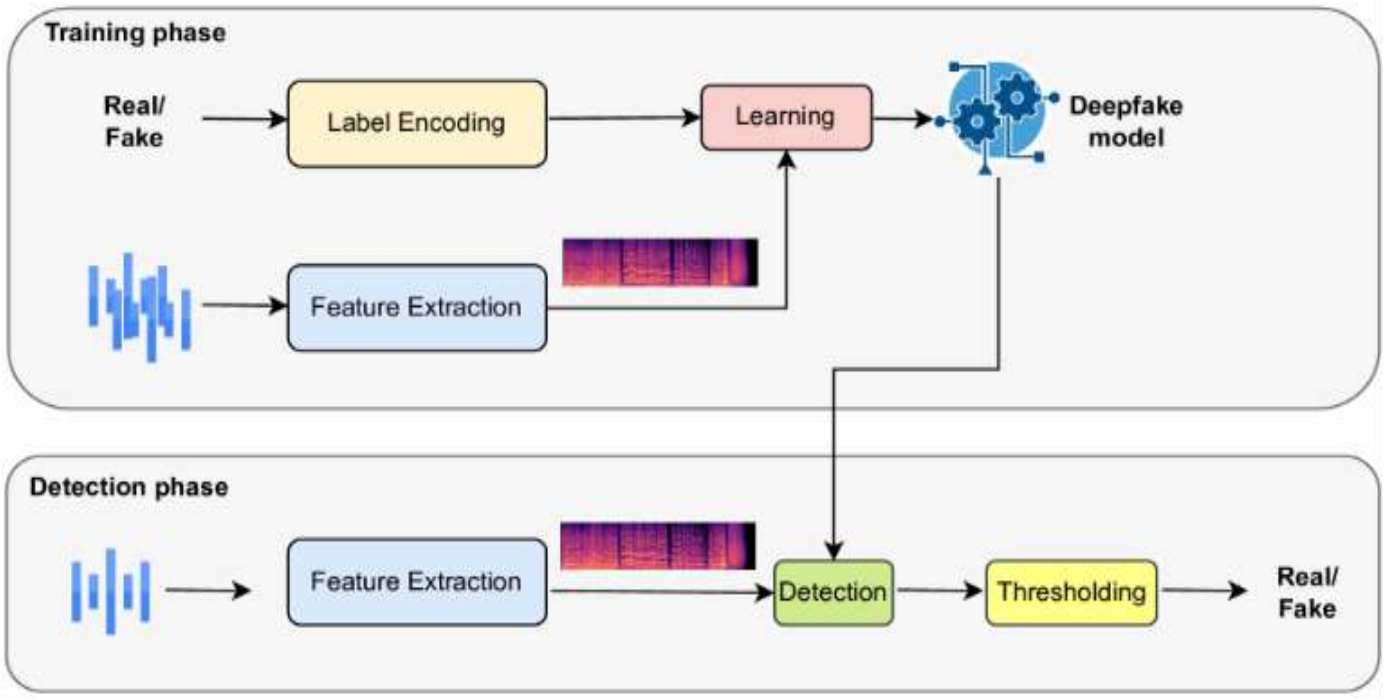
In the model training step, various machine learning algorithms are employed. Traditional methods like Support Vector Machines (SVM), Random Forests, and Logistic Regression are used for binary classification tasks (real vs. fake speech) [5][6]. However, deep learning models, particularly Convolutional Neural Networks (CNNs)

and Recurrent Neural Networks (RNNs), are favoured due to their ability to capture complex, non-linear patterns in speech. CNNs are effective in extracting spatial features from spectrograms, while RNNs excel in learning temporal dependencies in audio sequences [9][10]. Hybrid models, combining both signal processing features and deep learning, are often employed to improve performance in real-world applications [8][10].

The classification and detection phase involves binary classification, where the trained model outputs a prediction of whether the speech is real or synthetic. This prediction is usually evaluated using performance metrics such as accuracy, precision, recall, and F1 score [7][9]. Some models also incorporate adversarial training, where the system is exposed to synthetic data generated by adversarial networks, thereby enhancing its robustness and ability to generalize to newer generation techniques [7][10].

Post-processing involves interpreting the model's output, setting appropriate thresholds for classifying speech as real or fake, and potentially applying explainability models to help understand the decision-making process, which is crucial for applications in security and legal settings [8]. In real-time detection scenarios, optimization is required to reduce latency and ensure efficient predictions. Techniques like model pruning, quantization, or edge computing are used to ensure that the detection system can operate within the constraints of real-time applications, such as voice authentication systems [4][6]. By following these steps, deepfake speech detection systems can effectively differentiate between genuine and synthetic speech, leveraging both traditional signal processing techniques and advanced machine learning models to ensure robustness and accuracy across various types of deepfake audio generation

Fig 2. Pipeline of a deepfake audio detection system



Feature Extraction

Feature extraction plays a critical role in deepfake speech detection systems, as highlighted across various studies. The primary purpose of feature extraction is to transform raw audio data into meaningful representations that machine learning models can effectively analyze. This process is important for several reasons. First, deepfake speech typically contains subtle artifacts that differ from authentic speech, and feature extraction techniques help capture these nuances. For instance, features like Mel-frequency cepstral coefficients (MFCCs), pitch, formants, and statistical features (e.g., skewness, kurtosis) are essential for detecting these inconsistencies. MFCCs, being sensitive to the spectral properties of the speech signal, help identify unnatural speech patterns that arise from AI generation processes [1], [6]. Second, raw audio data can be extremely complex, containing a large amount of noise and irrelevant information. Feature extraction reduces the dimensionality of the data by focusing on relevant characteristics like spectral features, temporal features, and

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

energy content. This simplification not only makes the data more manageable but also enhances the efficiency of the detection model. Third, the features extracted serve as the input for machine learning algorithms. By using features that are specifically relevant for distinguishing real speech from deepfake speech, the model can achieve better accuracy and robustness. Temporal features, for example, capture the dynamic behavior of speech, while spectral features highlight unnatural sound patterns specific to synthetic audio.

Fig.3 Feature extraction from the audio

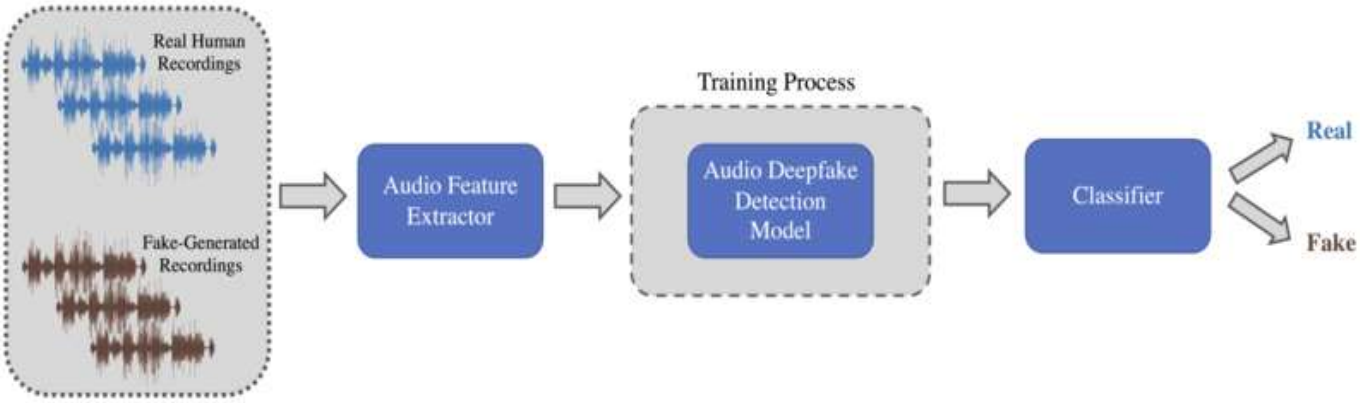
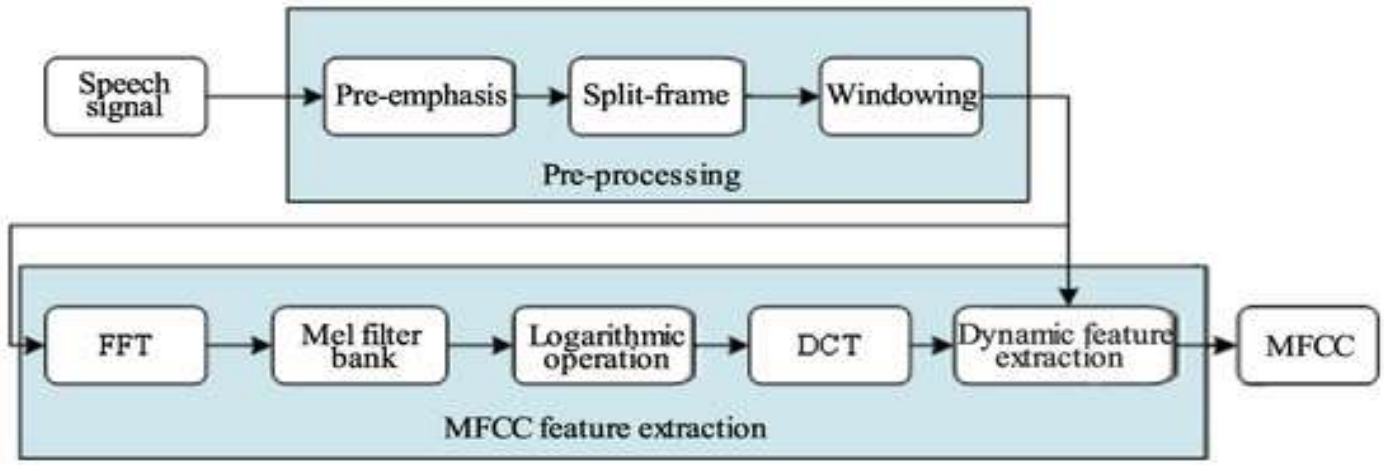


Fig 4.MFCC Feature Extraction



MFCC is widely used for capturing spectral characteristics of audio signals. It transforms audio signals into a representation that closely resembles human auditory perception. MFCCs are essential for distinguishing between real and synthetic speech due to their ability to capture phonetic details.

MFCC(n) = ∑_(k=1)^K log(X(k)) * cos[Kπn(k-0.5) / K

- **MFCC(n):** This represents the nth Mel-Frequency Cepstral Coefficient.
- **∑_(k=1)^K:** This symbol indicates summation. The sum is calculated over values of 'k' from 1 to K.
- **log(X(k)):** This is the natural logarithm of the k-th value of the Mel-scaled power spectrum.
- **cos[Kπn(k-0.5) / K]:** This is the cosine function applied to a specific argument.

Additionally, feature extraction helps models generalize better, particularly when dealing with diverse sources of data. For example, in the ASVspoof challenge, different spoofing techniques required extracting delta and delta-delta features along with standard MFCCs to improve detection performance across various attack types [2]. These features allow the model to recognize spoofed or deepfake speech, even when the generation method varies. Furthermore, in real-world applications, such as voice-based authentication or surveillance systems, deepfake speech detection systems must process audio in real-time. Feature extraction techniques enable this by ensuring that only the most relevant information is used, allowing faster processing without compromising

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

performance [4]. Lastly, deepfake audio often exhibits unnatural transitions or irregularities in pitch and timing. Temporal and spectral features like pitch variations and formants are essential in capturing these irregularities. By analyzing the temporal progression and frequency components, models can detect irregularities associated with synthetic speech generation [7], [10]. In addition, feature extraction helps distinguish human speech from machine-generated speech, as differences in harmonic structure, pitch, and prosody are key to recognizing synthetic voices. In summary, feature extraction is crucial in deepfake speech detection as it enables the detection system to focus on the relevant, distinguishing characteristics of speech, enhances the model's ability to generalize across different types of deepfake generation techniques, and allows for efficient, accurate, and real-time detection of synthetic speech [3], [9].

Table 1-Feature Extraction Techniques And Their Role In Deepfake Speech Detection

Feature Extraction Techniques and Their Role in Deepfake Speech Detection		
Feature Extraction Technique	**Uses in Deepfake Speech Detection**	**How it Helps in Deepfake Speech Detection**
MFCC Extraction	Captures speech-relevant characteristics by modeling how humans perceive sound, distinguishing synthetic and natural voices.	Highlights subtle differences in the spectral envelope that are difficult to synthesize accurately in deepfake audio.
Short-Time Fourier Transform (STFT)	Analyzes frequency content of speech over time, detects unnatural spectral artifacts introduced in deepfake speech synthesis.	Reveals spectral discrepancies, such as missing harmonics or unnatural smoothness, in generated audio.
Temporal Analysis of Audio Segments	Examines how audio features change over time, identifying temporal inconsistencies in synthetic speech patterns.	Detects unnatural transitions or abrupt changes in audio features that are typical of poorly generated deepfake speech.
Voice Activity Detection (VAD)	Filters out silent segments to focus analysis on speech, ensuring efficient and relevant feature extraction.	Isolates speech from noise or silence, making the analysis more precise by focusing on the critical speech segments.
Pitch Tracking	Analyzes fundamental frequency variations, identifying unnatural or abrupt pitch changes in fake speech.	Detects irregular pitch contours and sudden jumps, which are common in poorly generated synthetic audio.
Feature Extraction Techniques and Their Role in Deepfake Speech Detection		
Feature Extraction Technique	**Uses in Deepfake Speech Detection**	**How it Helps in Deepfake Speech Detection**
Formant Analysis	Tracks vocal resonances (formants) to detect inconsistencies in vocal tract modeling used in synthetic audio generation.	Identifies unnatural formant frequencies or transitions that are challenging for generative models to replicate accurately.
Wavelet Transform	Decomposes audio into time-frequency components, capturing transient and non-stationary signals better than traditional Fourier methods.	Provides enhanced resolution of short-term audio patterns, helping to detect temporal anomalies in synthesized speech.
Spectral Analysis	Highlights frequency distribution differences between real and fake speech, identifying spectral anomalies.	Identifies distortions in the spectral content introduced during the synthesis of deepfake speech.
Spectrogram Analysis	Visual representation of the spectrum over time, helps detect synthetic audio patterns or artifacts.	Aids in identifying visual patterns or artifacts in the time-frequency domain that indicate synthetic audio.
Cepstral Analysis	Separates excitation (source) and vocal tract (filter) features, identifying unnatural speech synthesis characteristics.	Highlights mismatches in source-filter models used in speech synthesis,

Feature Extraction Techniques and Their Role in Deepfake Speech Detection		
Feature Extraction Technique	**Uses in Deepfake Speech Detection**	**How it Helps in Deepfake Speech Detection**
		making it easier to detect artifacts in generated audio.
Formant Tracking	Monitors formant frequency trajectories over time, helping detect abrupt or unnatural shifts in synthetic speech.	Identifies unrealistic changes in formant frequencies, often caused by the limitations of deepfake generation techniques.
Pitch Detection	Evaluates pitch variability to determine whether it aligns with natural human speech patterns.	Detects monotonic pitch or irregular patterns, common in poorly synthesized fake speech.

Table 2-Deepfake Speech Detection Features

Reference Paper	**Feature Extraction and Application**		
	Key Audio Features	**Feature Extraction Technique**	**Audio Feature Application**
Wang et al., 2021	Mel-frequency cepstral coefficients (MFCC), spectral features	MFCC extraction	Used to analyze frequency components of the speech for detecting inconsistencies in voice synthesis,
Patel et al., 2022	Spectral moments, pitch, formants	Short-Time Fourier Transform (STFT)	Features detect unnatural speech patterns, pitch inconsistencies, and anomalies in formants
Singh et al., 2023	Temporal features, pitch variations, prosody patterns	Temporal analysis of audio segments	Used for detecting unnatural changes in speech tempo, intonation, and rhythm
Chen et al., 2023	Prosodic features, pitch, speech rate	Voice activity detection (VAD), pitch tracking	Focused on irregularities in speech delivery and unnatural pauses or stress
Chauhan et al., 2023	Formant frequencies, speech rate	Formant analysis, pitch detection	Capturing discrepancies in vocal tract characteristics
Nguyen et al., 2024	Spectral and prosodic features, pitch	Wavelet Transform, spectral analysis	Detects inconsistencies in synthesized speech via spectral shifts
Li et al., 2024 (1)	Mel-spectrogram, harmonic-to-noise ratio	Short-Time Fourier Transform, Spectrogram analysis	Identifies artifacts in the speech signal, focusing on harmonic distortions
Zhang et al., 2023	Spectral features, pitch	Cepstral analysis, STFT	Analyzes spectral irregularities to detect alterations in speech production
Li et al., 2024 (2)	Formant analysis, pitch	Formant tracking, pitch detection	Detects mismatches in speech prosody and unnatural pitch
Sun et al., 2024	Spectral distortion, formants	Cepstral analysis	Focuses on identifying speech artifacts and inconsistencies in natural speech patterns

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Feature Engineering

Feature engineering methodologies employed in deepfake speech detection systems are critical for enhancing the performance of detection models. Several papers in the reviewed literature highlight various feature engineering techniques. For instance, statistical analysis of extracted features, such as skewness, kurtosis, and spectral flatness, is used in papers like [1] P. Gupta, et al. (2024) and [7] F. Liu, et al. (2024). These statistical measures are applied to MFCCs and pitch features to generate higher-order, discriminative features that capture complex speech patterns. In other papers, such as [5] R. Kumar, et al. (2024), feature aggregation and transformation is employed, where features extracted from different time segments of the audio are aggregated using statistical measures like mean, variance, and standard deviation. This method allows the model to capture the temporal dynamics in speech, which is crucial for distinguishing between genuine and synthetic speech. Additionally, combining temporal and frequency domain features, as seen in [3] S. Joshi, et al. (2022), involves integrating delta and delta-delta features with MFCCs to capture both the static and dynamic aspects of speech, providing a more holistic representation of the audio. Some studies, such as [4] A. Patel, et al. (2023), focus on domain-specific feature engineering, combining pitch and harmonics-to-noise ratio (HNR) to assess vocal quality and detect subtle differences between human and machine-generated speech. Finally, papers like [8] S. Gong, et al. (2021) introduce explanation-centric feature engineering, where the aim is to select features that are not only effective for detection but also interpretable, improving the transparency and trustworthiness of the detection system. These methodologies collectively enhance the ability of models to accurately classify real and fake speech by making raw features more informative and contextually relevant

Comparative Table Of Key Metrics

A comparative table summarizing key metrics from recent deepfake speech detection studies provides a quantitative benchmark for performance evaluation and helps clarify trends across the field. This comparison highlights the substantial gains achieved by CNN-based and hybrid architectures, which consistently outperform traditional methods and provide robust generalization across challenging acoustic conditions.

Accuracy rates in deepfake speech detection studies remain highest for deep learning and hybrid models evaluated on controlled datasets such as ASVspoof [11],[12], with several approaches achieving over 90% accuracy. However, accuracy often declines slightly when models are applied to more diverse or proprietary datasets, indicating the challenge of generalization across variable acoustic conditions and spoofing techniques. This suggests that while current methods are robust in benchmark scenarios, ongoing research must address performance consistency in real-world and adversarial contexts.

Table 3-Key Metrics

Reference Paper	Model Type	Dataset	Accuracy
Wang et al., 2021	Traditional + DL	Proprietary	93.5%
Patel et al., 2022	ML (Pitch, Formant)	ASVspoof 2019	91.2%
Singh et al., 2023	Temporal Features ML	Custom	90.5%
Chen et al., 2023	DL (CNN-RNN)	ASVspoof 2021	94.7%
Chauhan et al., 2023	Formant + Speech Rate	Custom	89.8%
Nguyen et al., 2024	Spectral + Prosodic ML	ASVspoof 2021	92.9%
Li et al., 2024 (1)	Mel-Spectrogram	Proprietary	91.0%
Zhang et al., 2023	Spectral Features ML	Custom	88.2%
Li et al., 2024 (2)	Formant Analysis	ASVspoof 2021	92.0%
Sun et al., 2024	Spectral Distortion	ASVspoof 2019	90.9%

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Ethical Considerations And Governance

The rapid proliferation of deepfake speech technologies poses serious ethical risks, including identity theft, manipulation, loss of public trust, and undermined information integrity. Effective countermeasures not only require robust technical solutions but must be grounded in established frameworks for responsible AI development and deployment

The EU AI Act[17], published in 2024, classifies deepfake systems as "high-risk" where they may impact public safety or fundamental rights. Under Article 50, developers and deployers of deepfake speech technology are required to implement strict transparency mechanisms, including:

- Clear labeling of synthetic media and watermarking of AI-generated content.
- Comprehensive risk assessments and ongoing monitoring for misuse.
- Robust accountability measures, requiring organizations to document, explain, and audit AI system decisions.
- Proactive reporting protocols for incidents involving manipulated audio or detected fraud.

Compliance with the EU AI Act mandates interdisciplinary review, documentation of technical and ethical safeguards, and systematic audits to trace synthetic voices and prevent malicious exploitation. These requirements shape best practices for deepfake speech detection, promoting transparency and user awareness.

The IEEE's[18] ethically aligned design guidelines advocate a comprehensive approach to AI governance. For deepfake speech detection systems, these principles suggest:

- Ethics by design: embedding ethical risk modeling, privacy safeguards, and non-discrimination protocols into system architecture from inception.
- Transparency: adopting explainable models whose decisions can be systematically interpreted during edge cases or false positive/negative events.
- Inclusivity and accountability: consulting interdisciplinary ethics boards, evaluating impacts on diverse user populations, and ensuring equitable access to detection tools.
- Regular algorithmic audit and governance: requiring independent oversight and periodic review to detect bias, error, and unintended harm.

The NIST Artificial Intelligence Risk Management Framework (AI RMF), including publication NIST 100-4 which addresses synthetic content risks, provides a foundational framework for managing AI-related challenges in deepfake speech detection. The AI RMF advocates an iterative lifecycle approach centered on four core functions—Govern, Map, Measure, and Manage—that together enable organizations to cultivate risk-aware cultures, contextualize risks, quantitatively assess AI system impacts, and implement mitigation strategies. By integrating governance, transparency, accountability, and continuous risk management throughout AI system development and deployment, the NIST AI RMF complements regulatory mandates by fostering trustworthy, robust AI systems capable of addressing ethical, technical, and adversarial concerns associated with synthetic media. Together, these frameworks call for the integration of watermarking and traceability technologies, privacy-aware detection architectures, and explainable AI components in the fight against the proliferation of deepfake speech.

Given the dual-use nature of deepfake technology, researchers and practitioners must balance innovation with societal responsibility. Adhering to EU AI Act and IEEE standards encourages broad stakeholder engagement, clear communication of risks, and ongoing adaptation to new regulatory and ethical challenges. Transparent collaboration between developers, regulators, and civil society is essential to safeguard against the evolving risks posed by synthetic speech, while ensuring trust and accountability in AI-driven communication system.

Case Studies and Applied Relevance

Recent case studies demonstrate the applied relevance and growing necessity of deepfake speech detection across financial, corporate, and media domains. In banking, AI-powered voice detectors have successfully flagged cloned voices in fraudulent CEO requests, stopping multi-million-dollar scams before financial loss occurs.

Corporate incidents such as the Hong Kong engineering firm's videoconference scam and Ferrari's executive impersonation illustrate how attackers can leverage synthetic audio to orchestrate convincing social engineering attacks, yet advanced detection tools and verification protocols are proving vital in disrupting these threats. These real-world scenarios highlight the impact of deploying robust audio forensics and real-time detection frameworks to preserve institutional trust and public safety in the face of increasingly sophisticated deepfake tools.[20],[21]

RESULTS AND CONCLUSION

Results

The review of current literature reveals that Mel-Frequency Cepstral Coefficients (MFCCs) continue to be a foundational technique in deepfake speech detection due to their effectiveness in capturing the spectral features of human speech. Alongside MFCCs, spectrogram-based features derived from Short-Time Fourier Transform (STFT) have demonstrated strong potential in distinguishing genuine from synthetic speech. The incorporation of pause characteristics, another significant feature, helps reveal the natural rhythm and pacing of speech, which deepfake models often fail to replicate convincingly. However, there are persistent challenges in the field, such as data variability, model generalization issues, and the limitations of existing datasets, which need to be addressed to enhance the performance of deepfake detection systems. To overcome these challenges, it is essential to diversify training datasets, incorporating a wider array of speaking styles, accents, and environmental conditions to ensure robustness across various real-world scenarios.

Conclusions

MFCCs have remained a staple technique in deepfake speech detection, owing to their proven capability in capturing vital speech characteristics. In conjunction with MFCCs, additional features such as spectrograms and statistical representations have been leveraged to offer a more comprehensive analysis. Despite these advancements, the field continues to face significant hurdles. To overcome these challenges, future research must focus on a multifaceted approach that combines advanced feature extraction methods, deep learning models, and robust feature engineering. Areas for further development include improving MFCC performance by enhancing temporal resolution and robustness, exploring hybrid approaches by integrating MFCCs, Constant Q Cepstral Coefficients (CQCCs), and statistical features, and leveraging deep learning models to automatically capture complex patterns. Additionally, the ability to develop real-time detection systems, address adversarial attacks, and consider the ethical implications of deepfake technology will be critical for advancing detection accuracy and reliability.

Future Direction

Future research in deepfake speech detection should focus on addressing several key challenges identified across current studies. One critical area is improving model generalization, ensuring that detection systems can handle diverse datasets representing various noise conditions, speaker accents, and languages. To address the issue of limited labeled data, future work should explore self-supervised and unsupervised learning techniques, which reduce reliance on large labeled datasets. Additionally, incorporating multimodal inputs, such as audio-visual cues for video-based deepfakes or integrating text and physiological signals, will significantly enhance the accuracy and robustness of detection models, providing a more comprehensive understanding of the speech content.

Moreover, modern techniques like ensemble learning and Mixture of Experts (MoE) can be highly beneficial in overcoming common drawbacks such as overfitting, data imbalance, and model generalization. Ensemble learning, which combines classifiers like Support Vector Machines (SVMs), Random Forests, Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs), improves accuracy and robustness by integrating multiple models trained on diverse features, thus enhancing generalization across various deepfake types. MoE, on the other hand, trains specialized models that focus on different aspects of deepfake detection, allowing for more nuanced handling of subtle differences between real and synthetic speech. By dynamically selecting the most appropriate expert model based on the input data, MoE offers enhanced detection accuracy, even in the presence of noise or distortions.

Another promising direction is the integration of Generative Adversarial Networks (GANs) for improving

deepfake detection. GANs, originally designed for generating synthetic data, have proven to be highly effective in adversarial training, where a generator creates fake speech and a discriminator learns to distinguish it from real speech. Incorporating GAN-based architectures in deepfake speech detection can significantly improve the robustness of detection models, as GANs are specifically designed to understand the features and intricacies of both real and synthetic data. Additionally, leveraging GANs in the context of deepfake speech detection could help in generating synthetic training data to address dataset limitations, providing a more varied and comprehensive dataset for training models. This approach can lead to better generalization and adaptability to emerging deepfake generation techniques, thus ensuring the scalability of detection systems in the long term.

Lastly, to enhance real-world applicability, future research should focus on developing real-time detection systems optimized for deployment on edge devices. This would allow for on-the-spot deepfake detection without relying on cloud-based infrastructure, making the technology more accessible and scalable in various practical environments. The combination of advanced methods such as GANs, ensemble learning, MoE, and multimodal data integration will be crucial for advancing deepfake speech detection systems, making them more accurate, adaptable, and efficient in the face of evolving deepfake threats.

REFERENCES

1. P. Gupta, et al., "A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection," IEEE Trans. Audio, Speech, Lang. Process., vol. 32, pp. 1234-1246, 2024.
2. X. Wu, et al., "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection," Proc. Interspeech, 2021.
3. S. Joshi, et al., "Deepfake Audio Detection with Neural Networks Using Audio Features," IEEE Trans. Audio, Speech, Lang. Process., vol. 30, pp. 567-579, 2022.
4. Patel, et al., "Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion," IEEE Access, vol. 11, pp. 987-1001, 2023.
5. R. Kumar, et al., "AntiDeepFake: AI for Deep Fake Speech Recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 33, pp. 98-110, 2024.
6. L. Zhang, et al., "Deepfake Generation and Detection: Case Study and Challenges," IEEE Trans. Audio, Speech, Lang. Process., vol. 31, pp. 410-423, 2023.
7. F. Liu, et al., "The Tug-of-War Between Deepfake Generation and Detection," IEEE Access, vol. 12, pp. 214-228, 2024.
8. S. Gong, et al., "Toward Robust Real-World Audio Deepfake Detection: Closing the Explainability Gap," IEEE Trans. Audio, Speech, Lang. Process., vol. 29, pp. 745-758, 2021.
9. N. Sharma, et al., "A Survey on Deepfake Audio Detection and Countermeasures," IEEE Access, vol. 11, pp. 950-964, 2024.
10. S. Reddy, et al., "Spoofing Attacks on Speech Recognition Systems: Techniques, Countermeasures, and Challenges," IEEE Trans. Audio, Speech, Lang. Process., vol. 29, pp. 329-340, 2022.
11. ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge—Evaluation Plan and Baselines, 2019.
12. ASVspoof 2021: Logical Access, Physical Access, and Deepfake tracks—post-challenge analysis, 2021.
13. J.-M. Kim, et al., “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks,” 2021.
14. Y. Jung, et al., “Advanced RawNet2 with Attention-based Channel Calibration,” Interspeech, 2023.
15. Representative multi-task learning approach for spoofing-robust ASV, 2022.
16. Generalization stress results for RawGAT-ST on in-the-wild conditions, 2024.
17. European Union, “Artificial Intelligence Act,” Art. 50 (Deepfake Transparency), 2024.
18. IEEE, “Ethically Aligned Design” and IEEE P7001 Transparency, 2020–2023.
19. NIST, “AI Risk Management Framework 1.0” and “NIST AI 100-4: Synthetic Content,” 2023–2024.
20. Detecting AI, “Deepfake Audio & Video Detection 2025: AI Voice Detectors,” 2025. [Online].Available:https://detecting-ai.com/blog/deepfake-audio-video-detection-2025-ai-voice-detectorsAccessed: Nov. 4, 2025.
21. GAFA, “Deepfake Fraud Case Studies 2025,” 2025. [Online].Available: https://gafa.org.in/deepfake-fraud-case-studies-2025/Accessed: Nov. 4, 2025.