IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# CE-14 Deepfake Detection: Assessment of Speech and Emotion-Based Forensic Analysis

**P. A. Shinde[1] M. D. Laddha[2] H. R. Gaikwad[3]**
**S. S. Gandhi [4] Iram R. A. Jhetam[5]**
Department of Computer Science and Engineering,
Student [1] Faculty[2-5] Dr. Babasaheb Ambedkar Technological University, Lonere, Maharashtra, India.

**Abstract**— Deepfake technologies driven by generative adversarial networks (GANs), neural vocoders, and end-to-end speech synthesis architectures have significantly advanced the generation of highly realistic synthetic audio. Modern voice cloning systems are capable of replicating speaker identity, linguistic style, and vocal timbre with minimal training data, making synthetic speech increasingly indistinguishable from authentic human recordings. While these developments offer substantial benefits in assistive technologies, entertainment, virtual agents, and content creation, they simultaneously introduce serious security, ethical, and societal risks. Malicious applications include financial fraud, identity impersonation, political misinformation, social engineering attacks, and erosion of public trust in digital media.

This paper presents a comprehensive review of deepfake detection techniques with a primary focus on speech-based and emotion-aware forensic analysis. It systematically examines acoustic artifact detection methods, including spectral irregularities, phase inconsistencies, and harmonic distortions introduced during waveform synthesis. Furthermore, the study explores prosodic modeling and affective computing approaches that analyze emotional coherence, temporal emotion trajectories, and inconsistencies between semantic content and vocal expression. Multimodal fusion strategies integrating audio and visual cues are also discussed to enhance robustness against adversarial attacks and unseen synthesis models. Finally, key research gaps, cross-dataset generalization challenges, and future directions for resilient, explainable, and real-time detection systems are identified.

## I. INTRODUCTION

Deep generative models such as GANs [1] and advanced neural vocoders like WaveNet [2] and Tacotron [3] have revolutionized synthetic speech generation. Modern voice cloning systems can replicate speaker identity with minimal data [4]. However, subtle artifacts remain embedded in generated signals, enabling forensic detection [5].

Emotion plays a critical role in human speech communication. Affective computing research [6] demonstrates that natural speech contains dynamic emotional trajectories. Synthetic speech often fails to maintain realistic emotional coherence, making emotion-aware detection a promising strategy.

Deep generative models such as GANs [1] and advanced neural vocoders like WaveNet [2] and Tacotron [3] have revolutionized synthetic speech generation. Modern voice cloning systems can replicate speaker identity with minimal data [4]. However, subtle artifacts remain embedded in generated signals, enabling forensic detection [5].

Emotion plays a critical role in human speech communication. Affective computing research [6] demonstrates that natural speech contains dynamic emotional trajectories. Synthetic speech often fails to maintain realistic emotional coherence, making emotion-aware detection a promising strategy.

Deep generative models such as GANs [1] and advanced neural vocoders like WaveNet [2] and Tacotron [3] have revolutionized synthetic speech generation. Modern voice cloning systems can replicate speaker identity with minimal data [4]. However, subtle artifacts remain embedded in generated signals, enabling forensic detection [5].

Emotion plays a critical role in human speech communication. Affective computing research [6] demonstrates that natural speech contains dynamic emotional trajectories. Synthetic speech often fails to maintain realistic emotional coherence, making emotion-aware detection a promising strategy.

Deep generative models such as GANs [1] and advanced neural vocoders like WaveNet [2] and Tacotron [3] have revolutionized synthetic speech generation. Modern voice cloning systems can replicate speaker identity with minimal data [4]. However, subtle artifacts remain embedded in generated signals, enabling forensic detection [5].

Emotion plays a critical role in human speech communication. Affective computing research [6] demonstrates that natural speech contains dynamic emotional trajectories. Synthetic speech often fails to maintain realistic emotional coherence, making emotion-aware detection a promising strategy.

Deep generative models such as GANs [1] and advanced neural vocoders like WaveNet [2] and Tacotron [3] have revolutionized synthetic speech generation. Modern voice cloning systems can replicate speaker identity with minimal data [4]. However, subtle artifacts remain embedded in generated signals, enabling forensic detection [5].

Emotion plays a critical role in human speech communication. Affective computing research [6] demonstrates that natural speech contains dynamic emotional trajectories. Synthetic speech often fails to maintain realistic emotional coherence, making emotion-aware detection a promising strategy.

## II. RELATED WORK

Early spoof detection relied on cepstral features such as MFCC and CQCC [7], [8]. Deep CNN-based models improved feature learning from spectrogram representations [9]. Recurrent neural networks captured temporal dependencies [10].

Transformer architectures introduced self-attention mechanisms for contextual modeling [11]. Datasets such as ASVspoof 2019 [12] and FakeAVCeleb [13] enabled standardized benchmarking. Multimodal approaches combining audio and visual cues improved robustness [14], [15].

Early spoof detection relied on cepstral features such as MFCC and CQCC [7], [8]. Deep CNN-based models improved feature learning from spectrogram representations [9]. Recurrent neural networks captured temporal dependencies [10].

Transformer architectures introduced self-attention mechanisms for contextual modeling [11]. Datasets such as ASVspoof 2019 [12] and FakeAVCeleb [13] enabled standardized benchmarking. Multimodal approaches combining audio and visual cues improved robustness [14], [15].

Early spoof detection relied on cepstral features such as MFCC and CQCC [7], [8]. Deep CNN-based models improved feature learning from spectrogram representations [9]. Recurrent neural networks captured temporal dependencies [10].

Transformer architectures introduced self-attention mechanisms for contextual modeling [11]. Datasets such as ASVspoof 2019 [12] and FakeAVCeleb [13] enabled standardized benchmarking. Multimodal approaches combining audio and visual cues improved robustness [14], [15].

Early spoof detection relied on cepstral features such as MFCC and CQCC [7], [8]. Deep CNN-based models improved feature learning from spectrogram representations [9]. Recurrent neural networks captured temporal dependencies [10].

Transformer architectures introduced self-attention mechanisms for contextual modeling [11]. Datasets such as ASVspoof 2019 [12] and FakeAVCeleb [13] enabled standardized benchmarking. Multimodal approaches combining audio and visual cues improved robustness [14], [15].

Early spoof detection relied on cepstral features such as MFCC and CQCC [7], [8]. Deep CNN-based models improved feature learning from spectrogram representations [9]. Recurrent neural networks captured temporal dependencies [10].

Transformer architectures introduced self-attention mechanisms for contextual modeling [11]. Datasets such as ASVspoof 2019 [12] and FakeAVCeleb [13] enabled standardized benchmarking. Multimodal approaches combining audio and visual cues improved robustness [14], [15].

Early spoof detection relied on cepstral features such as MFCC and CQCC [7], [8]. Deep CNN-based models improved feature learning from spectrogram representations [9]. Recurrent neural networks captured temporal dependencies [10].

Transformer architectures introduced self-attention mechanisms for contextual modeling [11]. Datasets such as ASVspoof 2019 [12] and FakeAVCeleb [13] enabled standardized benchmarking. Multimodal approaches combining audio and visual cues improved robustness [14], [15].

## III. SPEECH-BASED DETECTION TECHNIQUES

Speech-based deepfake detection primarily focuses on identifying low-level acoustic anomalies introduced during synthetic waveform generation. These anomalies include spectral distortions, phase inconsistencies, and unnatural harmonic structures that deviate from the statistical properties of authentic human speech [16]. Generative adversarial network (GAN)-based systems, although highly advanced, often produce over-smoothed frequency bands and reduced micro-level stochastic variations, which can serve as detectable indicators of manipulation [17]. To strengthen forensic capability, neural fingerprinting techniques have been developed to identify generator-specific traces embedded within synthesized audio, enabling attribution to particular model architectures or training pipelines [18]. In terms of feature representation, Constant-Q cepstral coefficients (CQCC) and log-Mel spectrograms remain dominant due to their effectiveness in capturing frequency-domain characteristics and subtle spectral variations [7]. However, as adversarial training strategies continue to enhance the realism of synthetic speech, generators are increasingly capable of minimizing detectable artifacts. This evolving threat landscape necessitates the development of more robust countermeasures that combine spectral analysis, temporal modeling, and adversarial resilience mechanisms to ensure reliable detection performance under sophisticated attack conditions [19].

# IV. EMOTION-AWARE DETECTION

Emotion-aware deepfake detection leverages advanced emotion recognition models trained on benchmark datasets such as RAVDESS [20] and IEMOCAP [21] to extract key affective features, including arousal and valence dimensions, for comprehensive affect analysis. These models enable the identification of emotional states embedded within speech signals and provide quantitative representations of vocal expressiveness. Building upon this, temporal emotion trajectory modeling examines the progression of emotional states across an utterance, comparing the expected natural emotional flow of authentic human speech with the observed patterns in potentially synthetic samples. Natural speech typically demonstrates gradual, context-driven emotional transitions, whereas synthetic speech may exhibit abrupt changes, inconsistent modulation, or limited emotional variability. Additionally, prosodic irregularities serve as strong discriminative indicators of manipulation. Features such as abnormal pitch variation, irregular stress patterns, inconsistent energy distribution, and flattened affective contours often reveal artifacts introduced during synthetic generation [6], [22]. By integrating arousal-valence modeling, temporal emotion consistency analysis, and prosodic feature evaluation, emotion-aware detection frameworks enhance robustness and provide a higher-level semantic layer of forensic verification beyond purely acoustic artifact detection.

# V. GAP IDENTIFICATION

- **Limited Cross-Dataset Generalization:**
Despite achieving high accuracy on benchmark datasets, many deepfake detection models fail to maintain performance when tested on unseen datasets or new synthesis techniques [12].

- **Lack of Emotion-Based Core Features:**
Emotional inconsistency scoring is rarely integrated as a primary detection variable, limiting the potential of affect-aware forensic analysis.

- **Real-Time Deployment Challenges:**
High computational complexity and resource-intensive architectures restrict practical real-time implementation.

- **Limited Explainability:**
Deep learning-based detection models often function as black boxes, reducing transparency and limiting legal and forensic acceptance [23].

# VI. PROPOSED FRAMEWORK

## Acoustic Forensic Analysis

Acoustic forensic analysis remains the cornerstone of deepfake speech detection. Synthetic audio generated by GANs [1], neural vocoders such as WaveNet [2], and end-to-end speech synthesis models like Tacotron [3] often introduces subtle yet detectable anomalies. These include spectral distortions, phase inconsistencies, and harmonic irregularities that deviate from the statistical properties of natural human speech. By applying spectral divergence measures, researchers can quantify abnormal frequency distribution patterns that are typically absent in authentic recordings [16]. Similarly, phase coherence metrics provide insight into temporal alignment, revealing inconsistencies introduced during waveform synthesis.

Beyond traditional signal processing, neural fingerprinting techniques have emerged as a powerful tool. These methods identify generator-specific traces embedded within synthesized audio, enabling attribution to particular model architectures or training pipelines [18]. Such forensic attribution is critical in legal and investigative contexts, where determining the origin of manipulated content can strengthen evidence.

Moreover, acoustic analysis benefits from feature representations such as Constant-Q cepstral coefficients (CQCC) [7] and log-Mel spectrograms, which capture fine-grained frequency-domain variations. These features remain dominant in benchmark challenges like ASVspoof [12], where they consistently outperform

simpler cepstral features. However, adversarial training strategies continue to minimize detectable artifacts, making acoustic-only detection increasingly vulnerable [19].

Therefore, acoustic forensic analysis must evolve toward hybrid approaches that combine traditional signal-level detection with advanced machine learning models. By integrating handcrafted spectral features with deep neural representations, detection systems can achieve greater resilience against sophisticated synthesis techniques. This layered approach ensures that even as generative models improve, forensic systems retain the ability to uncover hidden manipulations.

## Emotion Consistency Modeling

Emotion plays a central role in human communication, shaping both linguistic meaning and social interpretation. Authentic speech exhibits dynamic emotional trajectories, with gradual transitions driven by context and speaker intent [6]. Synthetic speech, however, often fails to replicate this natural emotional coherence. Emotion consistency modeling leverages affective computing techniques to identify these discrepancies.

Using benchmark datasets such as RAVDESS [20] and IEMOCAP [21], emotion recognition models extract affective dimensions like arousal and valence. These quantitative measures provide a structured representation of vocal expressiveness. By analyzing temporal emotion trajectories, researchers can compare the expected natural emotional flow of authentic speech with the observed patterns in synthetic samples. Synthetic audio frequently demonstrates abrupt emotional shifts, inconsistent modulation, or flattened affective contours — all strong indicators of manipulation [22].

Prosodic features further enhance emotion-aware detection. Abnormal pitch variation, irregular stress patterns, and inconsistent energy distribution often reveal artifacts introduced during synthesis [22]. For example, while human speech naturally varies in pitch and rhythm to convey emphasis, synthetic speech may exhibit monotonous delivery or unnatural prosodic irregularities. These deviations serve as discriminative cues for forensic analysis.

By integrating emotion trajectory modeling with prosodic feature evaluation, emotion-aware detection frameworks provide a higher-level semantic layer of verification. Unlike purely acoustic methods, which focus on low-level signal anomalies, emotion-based approaches capture inconsistencies in expressive intent. This makes them particularly valuable in forensic contexts, where understanding the authenticity of emotional communication can be as important as detecting technical artifacts.

## Attention-Based Fusion

While acoustic and emotion-based modules each provide valuable insights, their integration through attention-based fusion significantly enhances detection robustness. Attention mechanisms dynamically weight feature contributions based on reliability, allowing the system to prioritize the most discriminative cues in a given context [11]. For instance, when acoustic anomalies are subtle but emotional inconsistencies are pronounced, the fusion layer can assign greater importance to emotion-based features.

This adaptive integration improves classification accuracy and resilience against adversarial manipulation. Attackers may attempt to bypass detection by minimizing acoustic artifacts or artificially smoothing emotional trajectories. However, by combining multiple feature domains, attention-based fusion ensures that manipulations in one dimension are counterbalanced by detection in another [19].

Furthermore, attention-based fusion supports explainability. By highlighting which features contributed most to a classification decision, the system provides interpretable reasoning for forensic experts [23]. This transparency strengthens trust in automated detection systems and enhances their legal admissibility.

Ultimately, attention-based fusion represents a holistic approach to deepfake detection. By unifying signal-level and semantic-level analyses, it creates a multilayered defense against evolving generative models. This ensures that detection systems remain effective not only against current synthesis techniques but also against future advancements in voice cloning and emotional modeling.

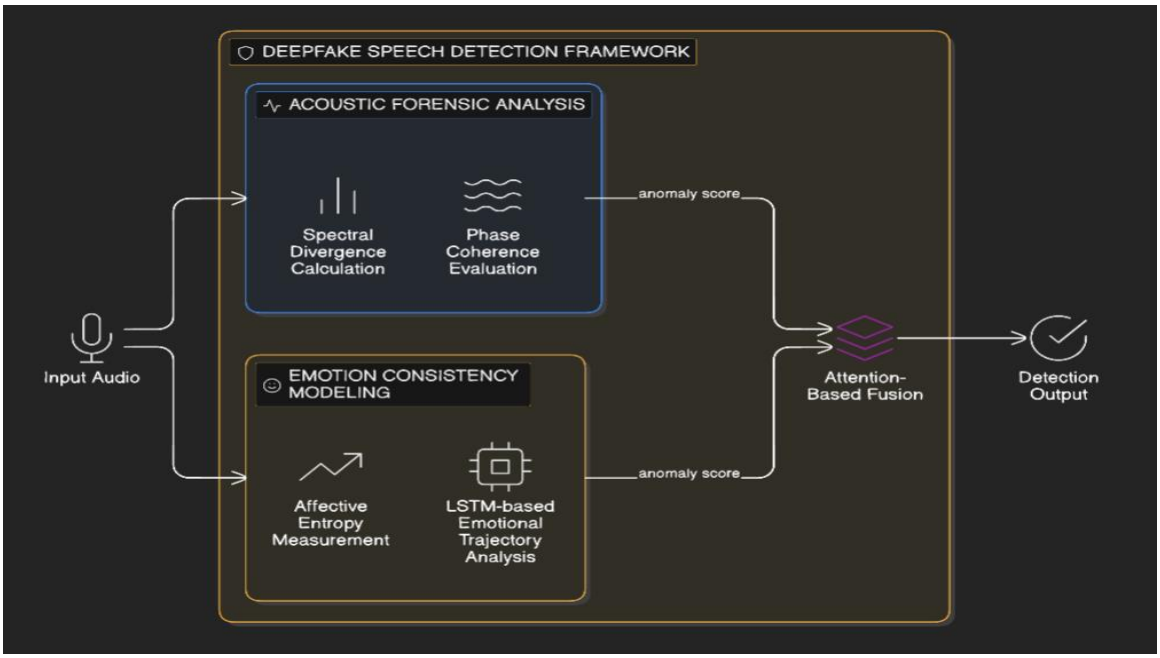


Fig. Proposed System Architecture

## VII. PERFORMANCE EVALUATION STRATEGY

Performance evaluation in deepfake detection has traditionally relied on classification metrics such as accuracy, precision, recall, F1-score, and equal error rate (EER) [12]. Accuracy provides a general measure of detection effectiveness, while precision and recall offer deeper insights into reliability. The F1-score balances sensitivity and specificity, ensuring that models are not biased toward either false positives or false negatives. EER, widely used in spoof detection challenges, reflects the operating point where false acceptance and false rejection rates are equal, making it particularly significant in forensic contexts [5].

Early spoof detection models based on cepstral features like MFCC and CQCC primarily reported accuracy and EER on benchmark datasets such as ASVspoof [7][12]. These metrics highlighted the effectiveness of handcrafted features in capturing spectral anomalies. However, performance often degraded when tested on unseen datasets, revealing limitations in generalization [8].

With the advent of CNN and RNN architectures, evaluation strategies expanded to include precision, recall, and F1-score. These models leveraged spectrogram representations and temporal dependencies, achieving higher accuracy on benchmark datasets [9][10]. Transformer-based systems introduced self-attention mechanisms [11], enabling contextual modeling and improved generalization. Their evaluation emphasized cross-dataset validation, ensuring robustness against unseen synthesis techniques.

Multimodal detection frameworks combining audio and visual cues employed robustness testing against adversarial attacks. Datasets like FakeAVCeleb [13] enabled standardized benchmarking of multimodal systems. Evaluation focused not only on accuracy but also on resilience against perturbation-based attacks designed to bypass detection mechanisms [14][15][19].

## VIII. FUTURE RESEARCH DIRECTIONS

Future advancements in deepfake detection should prioritize the development of explainable AI-based forensic systems to enhance transparency, accountability, and legal admissibility of automated decisions. Explainable models provide interpretable reasoning for classification outcomes, enabling forensic experts and judicial authorities to understand why a particular audio sample was flagged as synthetic, thereby strengthening trust and regulatory acceptance [23]. In addition to interpretability, federated learning frameworks offer a promising direction for collaborative detection across institutions without sharing sensitive raw data. By enabling decentralized model training, federated approaches preserve privacy while improving generalization across diverse datasets and attack scenarios. Furthermore, the design of lightweight transformer

architectures can significantly improve computational efficiency, allowing deployment on edge devices such as smartphones, embedded security systems, and real-time monitoring platforms. This is crucial for scalable and low-latency detection in practical environments.

Finally, integrating blockchain-based authentication mechanisms into media verification pipelines can establish immutable records of content origin and modification history. Such decentralized verification infrastructures can complement AI-based detection systems by providing provenance tracking, thereby reinforcing trust in digital communication ecosystems and strengthening resistance against large-scale misinformation campaigns.

## IX. CONCLUSION

This review paper has examined the evolution of deepfake detection techniques, highlighting the transition from cepstral feature-based spoofing countermeasures [7][8] to advanced deep learning architectures such as CNNs [9], RNNs [10], and transformers [11]. Acoustic artifact detection remains foundational, capturing low-level signal anomalies introduced during synthesis [16][17]. Emotion-aware modeling introduces a higher semantic layer of forensic analysis, identifying inconsistencies in affective trajectories and prosodic features [6][20][21][22]. Multimodal fusion strategies further enhance robustness, integrating audio and visual cues to counter adversarial manipulation [13][14][15].

Despite these advances, significant challenges persist. Many models struggle with cross-dataset generalization [12], achieving high accuracy on benchmark datasets but failing to maintain performance on unseen synthesis techniques. Explainability remains limited, with deep learning models often functioning as black boxes [23]. Real-time deployment is hindered by high computational complexity, restricting scalability in practical environments [19].

The proposed framework addresses these gaps by integrating acoustic forensic analysis, emotion consistency modeling, and attention-based fusion. This multilayered architecture enhances detection reliability, resilience, and transparency. By dynamically weighting feature contributions, it adapts to diverse attack scenarios and provides interpretable reasoning for forensic experts [23].

Future research should prioritize explainable AI [23], federated learning, and lightweight transformer architectures [11]. Explainable models will strengthen trust and legal admissibility, while federated approaches enable collaborative detection without compromising privacy. Lightweight architectures will facilitate deployment on edge devices, ensuring scalable and low-latency detection. Integrating blockchain-based authentication mechanisms can further reinforce trust by providing immutable records of content origin and modification history.

Ultimately, the future of deepfake detection lies in interdisciplinary collaboration. By merging signal processing, affective computing, and trustworthy AI, researchers can build detection systems that safeguard digital communication ecosystems against evolving synthetic media threats. Continued innovation in explainability, scalability, and adversarial resilience will be essential to counter the rapid advancement of generative models and preserve public trust in digital media.

## REFERENCES

[1] I. Goodfellow et al., 'Generative adversarial nets,' NeurIPS, 2014.

[2] A. van den Oord et al., 'WaveNet: A generative model for raw audio,' 2016.

[3] Y. Wang et al., 'Tacotron: End-to-end speech synthesis,' 2017.

[4] J. Jia et al., 'Transfer learning from speaker verification to multispeaker TTS,' NeurIPS, 2018.

[5] N. Evans et al., 'ASVspoof 2019 challenge,' Interspeech, 2019.

[6] R. Picard, Affective Computing, MIT Press, 1997.

[7] M. Todisco et al., 'CQCC features for spoof detection,' Odyssey, 2016.

[8] Z. Wu et al., 'ASV spoofing and countermeasures,' Speech Communication, 2015.

[9] H. Tak et al., 'End-to-end anti-spoofing system,' ICASSP, 2020.

[10] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' Neural Computation, 1997.

[11] A. Vaswani et al., 'Attention is all you need,' NeurIPS, 2017.

[12] ASVspoof 2019 Dataset.

[13] K. Khalid et al., 'FakeAVCeleb: A novel audio-video dataset,' 2021.

[14] Y. Li et al., 'Exposing deepfake videos by detecting face warping artifacts,' CVPRW, 2018.

[15] P. Korshunov and S. Marcel, 'Deepfakes: A new threat to face recognition,' 2019.

[16] J. Frank et al., 'Leveraging frequency analysis for deepfake detection,' ICMLW, 2020.

[17] T. Karras et al., 'StyleGAN2,' CVPR, 2020.

[18] T. Jung et al., 'Neural fingerprinting for deepfake detection,' 2022.

[19] N. Carlini and D. Wagner, 'Adversarial examples are not easily detected,' 2017.

[20] S. Livingstone and F. Russo, 'RAVDESS dataset,' PLoS ONE, 2018.

[21] C. Busso et al., 'IEMOCAP dataset,' Language Resources and Evaluation, 2008.

[22] S. Jelil et al., 'Prosodic features for spoof detection,' 2021.

[23] W. Samek et al., 'Explainable artificial intelligence,' IEEE Signal Processing Magazine, 2017.