# Detection of AI-Generated (Cloned) Voices Using Audio Forensics: A Review

*Hasmita Kaur*[1*]

[1*]*B.Sc. (Hons) Digital Forensic Science, Malla Reddy University, Hyderabad, India*

[*]*hasmitakaur@gmail.com*

## Abstract:

Artificial intelligence is a domain which is constantly evolving in various fields. The recent advances in AI have given rise to voice cloning technologies which produce voices that sound exactly like human speech. These are highly realistic and common man cannot easily distinguish between the generated(cloned) voice and a genuine voice. This technology has useful applications but can be dangerously misused in criminal activities like fraud, impersonation and generating fake audio evidence. This review paper focuses on the detection of AI-generated (cloned) voices using audio forensic tools and also addresses the challenges that the forensic investigators face while detecting fake voice and a genuine voice sample. The paper also understands the forensic tools and techniques being used currently such as audio forensic methods like acoustic analysis and spectrographic analysis and identifies the gaps and limitations and addresses the need for more robust and improved forensic tools. Additionally, it discusses the upcoming machine learning oriented detection methods which are more efficient for detecting AI generated audio like deepfake audio. This paper also includes a comparative study of the traditional forensic methods and the emerging methods. Furthermore, this review focuses on the other implications of cloned voices for criminal investigations, legal and judiciary systems with respect to admissibility and reliability in the courtroom. The aim of this paper is to understand how these AI generated voices are detected and addressing the need for better and upgraded tools to combat the challenges faced by the forensic investigators.

**Keywords:** Artificial Intelligence, Audio, Voice cloning, Audio Forensics, AI-generated Voices, Forensic tools, Forensic Voice Analysis, Deepfake Audio.


D. R. Reddy et al. (eds.), *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, Advances in Computer Science Research 127,
https://doi.org/10.2991/978-94-6239-610-4_22

## 1. Introduction:

Artificial intelligence can now replicate a person's voice and make it seem original and genuine making it possible to generate human-like voices through processes like voice cloning. [3,7,8] These voice cloning technologies can replicate a human's voice characteristics such as pitch, accent, speaking style and etc. which makes the generated audio highly realistic. Voice cloning technology has numerous useful applications such as virtual assistants, chat-bots and in entertainment and media for example in audiobooks and etc., while this technology has its advantages it raises concerns with its disadvantages as well. Voice cloning technology can be dangerously misused for illicit activities such as fraud calls, impersonation, financial fraud, identity theft and manipulating audio evidence. [5,8]

Audio forensics is a branch of digital forensics that deals with the collection, observation, enhancement, analysis and evaluation of sound or audio recordings for legal and criminal investigations.[6] Earlier forensic experts depended on traditional forensic techniques such as acoustic analysis, spectrographic analysis and the voice comparison feature [1,6]. However, these methods were simply not enough as the emergence of AI generated voice has challenged the traditional methods since voice cloning has the ability to closely mimic natural human voice along with its characteristics making it difficult to distinguish between the two voices.

Deciding whether an audio recording is genuine or AI generated is one of the biggest challenges for the forensic investigators [3,8]. This review paper examines the existing research on the detection methods. It compares the traditional and emerging methods, discusses about their strength and limitations and highlights how this fake audio can have effects on the legal proceedings. By analysing recent literature and past studies and overviews gaps are identified in existing methods, and the aim is to focus on the need for better and upgraded forensic tools.

## 2. Literature review:

**2.1** Traditional Audio Forensics Methods:

Traditionally, speaker recognition and voice comparison were the two main focuses of audio forensics originally. These were done by using methods like perceptual and acoustic analysis [1,6]. Traditional audio forensic approaches were based on features like frequency (fundamental and formant), temporal patterns like pauses and silence duration, rhythm, voice quality features like jitter and lastly linguistic and behavioural features like accent and pronunciation. The analysis of these features was widely used to help forensic investigators in determining speaker's identity.

According to my research on this topic, several studies underline that identifying human voice imitation can pose as a challenge to the traditional methods. Research on professional voice impersonators shows that speakers can deliberately manipulate features like pitch, formant frequency, spectral shapes and etc in order to resemble a targeted speaker [1]. These imitated voices may appear very original and convincing but in-depth acoustic analysis can reveal differences between genuine and fake speech which shows that acoustic features play a critical role in the identification of imitated voices.

Even though, these methods are very useful, they have certain limitations, like sometimes forensic recordings that are obtained might be affected by factors like noise, compression, limited duration which make even the acoustic analysis difficult [6]. Moreover, most of the conventional methods depend on the expert interpretation, which leads to partiality and limited resulting. These limitations make the effectiveness of the traditional methods a lot weaker.

**2.2** AI- Generated and Cloned Voice:

Recently, there have been many advances in the field or artificial intelligence one of which is the development of realistic voice cloning and speech systems. There is also modern text to speech and text to voice conversion models which are based on the concept of deep learning and they can generate synthetic speech that almost exactly mimics a target's voice characteristics [3,7,8]. AI generated voices also can depict consistent and acoustic patterns making them difficult to be distinguished from genuine speech.

Studies on audio deepfakes focus on the fact that AI generated voices can replicate accurately the exact human voice characteristics and these voice cloning technologies are getting accessible day by day which has introduced serious threats in real world scenarios like fraud, manipulation of audio evidence, impersonation and etc [3,5,8].

When AI generated voices were first introduced a gap was noticed in the use of traditional forensic methodologies since they could not produce accurate results due to which there was a need for better and improved tools and techniques for audio analysis [6].

**2.3** Detection of AI Generated Voices (Modern Methods)

Since traditional methods posed many challenges, recent research has focused on automated detection techniques used on the basis of concepts such as machine learning and deep learning approaches [3,4,7]. These methods involve extracting audio features such as the MFCC- Mel-frequency cepstral coefficients, LFCC-linear cepstral coefficients and CQCC-Q cepstral coefficients [3,7]. It also included techniques such as spectrogram analysis and etc, classification models are used to differentiate between genuine and cloned voices.

Deep learning architectures involve convolutional neural networks, and have shown great results in detecting deepfake audio under controlled experimental conditions [4,7]. Survey based studies report that there is high detection accuracy when models are trained and tested on standard datasets, showing that AI based detectors are capable of detecting such features that can't be observed by traditional acoustic analysis

However, the detection performance might reduce when models are applied to unseen datasets this shows issues like generalization and robustness [3,4]. And the deep learning models also operate as black box systems and offer limited interpretability that raises concerns about the reliability during legal proceedings [4,5].

## 3. Methodology:

This paper was approached using a review-based research methodology to observe the existing research papers articles and information on the topic of AI generated and cloned voices in the field of audio forensics. The main objective of this review paper is to analyse the traditional forensic techniques as well as comparing them to the new and emerging machine learning based approaches and techniques.

A systematic search was conducted using academic databases and resources like Google Scholar, IEEE Xplore, and SpringerLink ResearchGate and ScienceDirect. Keywords used to find the content included audio forensics, AI- generated voice, detection of cloned voices, deepfake audio detection. Conference papers, research papers, journal articles, and survey studies that were published during the span of 2018-2024 were considered for this review.

The selected research papers were observed analysed understood and categorised based on the content needed like based on detection techniques used, that included traditional as well as machine learning methods. A comparative study was performed to identify the gaps, strengths and limitations.

## 4. Detection techniques:

Detecting AI generated and cloned voices are one of the biggest challenges in the field of audio forensics since the synthetic speech is getting highly realistic with time. The existing detection techniques are classified into traditional feature-based methods and automated machine learning based methods, both have their distinct advantages and disadvantages.

### 4.1 Audio forensic feature-based techniques:

The traditional audio forensics detection is based on the in-depth analysis of acoustic features that characterize human speech. The parameters that are commonly examined are fundamental and formant frequencies, temporal characteristics, spectral energy distribution, and voice quality measurements like jitter and glottal source features [1,6]. These features are originally derived from physiological and behavioural aspects of the human voice and human speech and these features have been for speaker recognition and comparing voice samples.

Studies on voice replication and speaker recognition show that speakers (human or synthetic) can manipulate certain vocal features to resemble a particular voice, subtle differences are still seen in the acoustic patterns [1]. Analysis based on features allows the forensic experts to identify any unnatural irregularities, irregular pitch behaviour, or inconsistencies in spectral structures these factors may indicate artificial speech generation. Such techniques hold importance in forensic contexts because they interpret and provide direct link.

Though, feature based forensic methods face disadvantages when they are applied to AI generated voices. Voice cloning systems are capable of replicating even the most minute details like acoustic cues that too with high precision due to which the traditional methods become ineffective [3,7]. Additionally, real-world forensic recordings are affected by factors like noise,

compression and channel variations, which complicate feature extraction and analysis. Due to these challenges the scalability and robustness of traditional forensic approaches are limited.

### 4.2 Machine learning and deep learning approaches

Since the traditional methods had numerous limitations, recent research has adopted machine learning and deep learning techniques to overcome the limitations faced by traditional methods and to detect AI generated voices accurately. These approaches involve extracting differentiating features such as Mel-frequency cepstral coefficients (MFCC), linear and constant Q cepstral coefficients (LFCC and CQCC), phase-based features, spectrogram representations and etc.

Deep learning models, like convolutional neural networks, have shown reliable performance in identifying even the most minute and subtle details which go unnoticed by traditional methods [4,7]. These models automatically learn the complex patterns and inconsistencies that cannot be detected otherwise. Survey studies show that these deep learning bases detection techniques provides high detection accuracy under controlled experimental conditions [3,4]. Which makes them attractive for large scale and automated scan of audio evidence.

Even though they offer many advantages, they still give rise to significant forensic concerns. Many deep learning models act as black box systems which means that the internal working is unknown and there is limited transparency during decision making processes [5,8]. And this uncertainty can challenge the courtroom acceptance and the evidence may be labelled as inadmissible. Another limitation is that the detection performance decreases when models are tested on real world audio that challenge reliability.

### 4.3 Comparative Analysis:

The comparative analysis of both traditional audio forensic methods as well as machine learning methods show clear differences between the two in the applications, effectiveness and reliability. Traditional forensic methods are relied upon acoustic parameters like pitch, formant frequencies, and spectral characteristics. These methods do analysis in depth and are transparent which can be legally advantageous [1]. However, their detection is limited when they face highly realistic sounding AI generated voices that can accurately produce even the little acoustic features.

In contrast, machine learning and deep learning approach show better and more robust performance in identifying AI generated speech under controlled conditions. It has the ability

to learn complex patterns and synthesis related artifacts directly from data, these models are better equipped to detect the most minute features and subtle inconsistencies that cannot be observed by manual analysis. Due to this they are considered more effective for large scale and automation detection tasks. However, this method also has a limitation which is that they have limited explainability as the process is done in a black box environment raising concerns in court proceedings and also this approach completely depends on the training data which limits its scope.

From a forensic point of view, both the above discussed techniques are not sufficient to address the increase in threat posed by AI generated voices. Recent studies emphasize the importance of hybrid detection frameworks that combine the advantages of both traditional methods as well as machine learning techniques, thereby balancing the performance [3,6].

## 5. Challenges And Limitations:

It is challenging to detect AI-generated and cloned voices because of the rapid advancement in the voice cloning technologies [7,8]. Modern speech systems can closely and precisely replicate human speech along with accurately imitating its feature and characteristics as well. Due to which both traditional and automated detection techniques become ineffective.

There is a lack of standardized datasets, benchmarks, procedures, standard procedure of operating, evaluation protocols further limit the comparability [5,8]. Due to these challenges there are concerns about courtroom admissibility.

## 6. Legal And Forensic Implications:

The advancement of AI generated voices has serious implications for the audio evidence to be reliable in legal proceedings. When the audio recording is highly realistic it undermines by assuming that the audio recording is inherently authentic, which raise concerns about courtroom admissibility [5,8].

Both traditional as well as automated detection techniques have advantages but disadvantages similarly, while traditional forensic methods offer transparency and clarity, many AI based detection systems operate as black box models [5], that make them unreliable in court because the process is unknown. Plus, there are also no clear legal standards and frameworks that have been made as of now which also complicates the judicial process

All these issues show us the need for legally reliable detection techniques, along with expert interpretability, and there should also be proper standardized guidelines and the growing threat of cloned voice audio in forensic investigations.

## 7. Future Directions:

The research in the field of AI generated and cloned voices is still ongoing due to new advancements in technology every day. And the future research should focus on developing hybrid detection frameworks which are a combination of both traditional audio forensic analysis and the machine learning techniques, with a combined approach like this we will have the best of both worlds improving both detection performance and reliability.

Another important part is the development of explainable AI based detection models, which will enhance transparency and it will allow forensic experts to be able to produce the result in the court in a systematic manner and will enhance courtroom acceptance. Researchers must also focus on improving the robustness and generalization of detection systems to handle real world recording conditions like noise, compression and other advanced voice cloning techniques.

At last, the creation of universally acceptable standardised datasets, benchmarks and protocols which are essential for the proper working of detection methods. Collaboration between researchers, forensic investigators, and legal experts will play a key role in addressing the challenges posed by AI generated audio.

## 8. Conclusion:

The increase in the use of AI- based voice cloning technologies has introduced new challenges for audio forensics and the reliability and admissibility of audio evidence. This review has explored the existing techniques which are used to detect AI-generated and cloned voices, covering both traditional audio forensic methods and the modern machine learning methods. Traditional techniques are valuable due to their interpretability but they are becoming less effective when put against highly realistic synthetic speech. Machine learning based methods show stronger detection capabilities but they are dependent on training data and there is no transparency in the procedure that can raise legal concerns.

The findings of this review suggest that we cannot rely on a single detection approach as that becomes insufficient in the current threat landscape. To make sure that the detection of AI-generated voices is effective there is a requirement to combine the forensic acoustic expertise

and the advanced computational techniques which are backed by standardised evaluation methods. Following this will strengthen the transparency, efficiency, robustness and legal admissibility of audio evidence.

## Acknowledgement:

We would like to express our sincere gratitude to **Mr. Vinod Kaaparthi,** Department of Digital Forensic Science, Malla Reddy University, for his valuable guidance, constructive suggestions, and continuous support throughout the course of this research. His academic expertise and feedback significantly contributed to the direction, methodology, and overall quality of the study. We also acknowledge the academic environment and resources provided by Malla Reddy University, which facilitated the successful completion of this work.

## References:

1. Kitamura T. (2008). “Acoustic analysis of imitate voice produced by a professional impersonator.”. *Proceedings of Interspeech 2008*, 813-816
2. Hasan, T., & Hansen, J.H.L (2013). “Acoustic factor analysis for robust speaker verification”. *IEEE Transactions on Audio, Speech, and Language Processing, 21(4*), 842-853. https://d.docs.live.net/201E49A02D133047/Desktop/research%20papers/Acoustic_Factor_Analysis_for_Robust_Speaker_Verifi.pdf
3. Yi, J., Wang, C., Tao,J., Zhang, C. Y., Zhao,Y. (2023). “Audio deepfake detection: A survey”. *IEEE Journal of Selected Topics in Signal Processing* https://d.docs.live.net/201E49A02D133047/Desktop/research%20papers/audio%20deepfake%20detection%20survey.pdf
4. Mvelo Mcubaa, M. Singh, a., Ikuesan & Venter, H. (2023). “The effect of deep learning methods on deepfake audio detection for digital investigation”. *Procedia Computer Science*, 219, 211-219 https://d.docs.live.net/201E49A02D133047/Desktop/research%20papers/research%20paper%201.pdf
5. Galyashina, E. I., & Nikishin, V. D. (2021). “AI generated fake audio as a new threat to information security: Legal and forensic aspects”. *Proceedings of the International Conference on Computer and Information Security (INFSEC 2021),* 17-21. https://d.docs.live.net/201E49A02D133047/Desktop/research%20papers/research%20paper%202.pdf

*International Journal for Research in Applied Science & Engineering Technology,* 11(9), 680-686 https://d.docs.live.net/201E49A02D133047/Desktop/research%20papers/research%20paper%203.pdf

7. Barrington, s., Cooper, E. A., & Farid, H. (2025). "Peoples are poorly equipped to detect AI-powered voice clones". *Scientific Reports*, 15, Article 11004. https://d.docs.live.net/201E49A02D133047/Desktop/research%20papers/s41598-025-94170-3.pdf
8. A. K. Singh and P. Singh, "Detection of ai-synthesized speech using cepstral & bispectral statistics," 2021 *IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR),* pp. 412–417, 2021.