Mrs. M. Devika, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Deep Fake Audio Detection Using MFCC And NLP

Mrs. M. Devika [1], UG Student [2]

[1] Assistant Professor Department of Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram, Chennai, India.

2 Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Abstract- The recent developments in restrictions on AI have resulted in the ability to create almost realistic-sounding deep fake audio content which has created new threats to online safety, security and privacy with many kinds of identity crimes or fraud, misleading or erroneous information, impersonation. The vast majority of audio recordings made by real people sound different than those produced synthetically with subtle variations in voice characteristics; therefore, it is becoming increasingly difficult for traditional methods to accurately identify any audio as being 'real' or 'fake.' Consequently, the design of an AI-powered method for detecting Deepfake Audio will utilize Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction (from both samples of voice recordings) using Natural Language Processing (NLP) incorporating an analysis at the level of speech to determine whether or not a sound is legitimate based on their similarities and differences to other sounds and how the two audio files relate to one another within the context of their use. Artificial Intelligence will aid in achieving this objective through the extraction and classification of the spectral and linguistic features contained within the audio files by two separate models: Machine Learning (ML) and Deep Learning (DL). The solution has been demonstrated to improve detection rates and to exploit multiple types of media that can be employed for producing deep fake audio attacks; hencely making it extremely resilient against attacks of all types. Therefore, the technology has the potential for augmenting the safety, reliability and confidence level that individuals have when using voice-based digital communications.

Keywords: Deepfake Audio Detection, MFCC, Natural Language Processing, Speech Analysis, AI Security, Audio Classification.

I. INTRODUCTION

Due to rapid developments in both AI and voice synthesis technology, we now face a new security threat with deepfake audio. These utility-based audio files are produced in such a way as to mimic the human voice to the point that it is increasingly difficult to tell whether a recording is real or altered. This capability has generated considerable concern in fields such as fraud, misinformation, identity theft, and voice recognition systems. With an increase in the use of voice interfaces and digital communication, verifying the authenticity of audio content will become critical.

Control methods for detecting and verifying deepfake audio have traditionally relied on manual assessment and/or elementary techniques that only assess the base value of the unity sound which is insufficient when assessing high-quality deepfake sounds. Therefore, existing detection methods do not account for minute differences in an utterance's characteristics and are not applicable for all types of falsifications as audio generated by Text-to-Speech (TTS), Voice Cloning (VC) or Replay attacks. Because of this, there is a need for a comprehensive solution that can automatically and reliably identify artificial audio.

Recent developments in machine learning and audio processing methods have advanced rapidly and allow for the representation of speech signals as feature sets that describe the spectral characteristics

of the signal via feature extraction methods. An example of this is Mel-Frequency Cepstral Coefficients (MFCCs). Natural Language Processing (NLP) can then be used to characterize the linguistic patterns and characteristics of spoken content. The NLP approach may also enhance deepfake audio detection systems because it should decrease both the number of false positive and false negatives by integrating acoustic and semantic analysis.

This study presents an innovative AI-enabled Deepfake Audio Detection System that utilizes an MFCC-based feature extraction method, coupled with an NLP method, to perform analysis on audio signals. The system processes the incoming data to derive acoustic features, which are then classified using machine learning to determine whether the audio is authentic or fake. The goal of this new technique is to improve detection capabilities, provide security for voice-driven platforms, and help establish trust within the digital communication ecosystem.

II. RELATED WORK

[3] As voice synthesis technology is being misused more frequently as a way to create fraudulent deepfake audio, there has been a surge in research into this area in order to develop better techniques for deepfake audio detection. Historical methods used to differentiate between real and fake audio contained a variety of standard signal processing methods such as Mel Frequency Cepstral Coefficients, pitch and spectral properties. Although these methodologies are effective for distinguishing between real and fake audio, they have limited capabilities when it comes to being resilient to sophisticated spoofing techniques that involve multiple levels of sophistication.

[6] Support Vector Machines, Gaussian Mixture Models and Random Forests represent some of the machine learning algorithm classification methods that have been widely employed in relation to audio classification problems. However, a significant disadvantage of these techniques is that they rely on the researcher for manual engineering of the features that would allow the researcher to distinguish between fake audio produced by deepfake audio and genuine audio.

[5] Recently, advances in deep learning have resulted in the adoption of neural networks such as Convolutional Neural Networks and Recurrent Neural Networks as the basis to develop models that automatically learn meaningful representations of audio signals and audio spectrograms. While the models have been able to achieve much higher levels of deepfake detection than traditional signal processing methods, they do so at the cost of significant amounts of training data required for the model to learn the correct audio representation.

[4] Also, many natural language processing techniques have been built to look at the transcripts of audio recordings in order to find any inconsistencies with what was said by that person based on their context. Unfortunately, there are many techniques out there, but they focus solely on how the sound was produced (acoustic features) and do not examine the meaning of what was said (semantic analysis).

[1] As a solution to this shortcoming, the current method combines features based on MFFC (Mel-frequency cepstral coefficients) taken from the audio and uses natural language processing methods in order to provide a better and more accurate method of identifying if an audio recording is deepfaked.

III. METHODOLOGY

This system called Deepfake Audio Detection Using MFCC & Natural Language Processing uses two types of analysis: acoustic and semantic, to identify if an audio clip was originally real or fake. The system improves detection accuracy and system stability by using an array of processing-type techniques along with an array of artificial intelligence-type methods to enhance performance of the entire system. There are three basic steps of operation for the complete system. Audio input is the first step in the operation of the complete system, second step is Preprocessing and Feature Extraction (mfcc), and the third step is to perform Speech-to-Text Conversion,

run NLP on those text results, classify those results, then generate output for that classification.

A. Audio Processing and Feature Extraction

Input audio files can be submitted in another format such as WAV and/or MP3 file types. Once received, the input audio is preprocessed by removing unwanted noise from the sample and ensuring that all sections of the audio are at a similar sound level. Once this task is completed, the audio is processed using MFCC in order to extract important spectral features from the audio, as these types of features contain attributes that can be used to distinguish between sounds. Some examples of these attributes located within the MFCC feature set include pitch and frequency; therefore, a listener will be able to distinguish between authentic audio and deepfake audio based upon the presence of these attributes within the collected data. The MFCC feature set will be converted into a feature vector, which will be used in future analysis.

B. Semantic Analysis and Classification

The system utilizes speech recognition technology to translate spoken words into written text, which aids in improving the capability of detecting speech. By employing various NLP techniques, the system analyzes word combinations for the occurrence of errors within the generated audio file based on the audio file's timestamps. The system uses the combination of MFCC feature information along with the textual information submitted to both CNN and SVM based machine learning/deep learning algorithms to classify the samples. The use of a hybrid method in this capacity will increase the level of accuracy in the analysis of spoken words because both acoustic and semantic data will be evaluated during the analysis stage.

C. Detection and Output Module

In addition to determining if an audio sample is authentic or deepfake, the system will assign a confidence score to the audio sample and display the results through a set of simplified user interfaces, so that users will not have any problem interpreting the results. This procedure will provide reliable results that will make it more difficult to compromise the integrity of audio files used within voice recognition based systems.

IV. PERFORMANCE EVALUATION

Evaluation of the Deepfake Audio Detection System by means of MFCC and NLP measurements reveals its ability to detect, resist attacks, and classify accurately. Since the system measures audio content in real time without depending on an already set data set, the evaluation will utilize both metric-based assessments and experimental verification.

A. Accuracy-Oriented Validation

This system tests and evaluates its ability to detect audio content through accuracy-based validation. It measures how well the system works when recognizing both real audio recordings and fake or "deepfake" audio recordings. In order to analyze the overall performance of the system, researchers use the classification results of how well the system can determine the differences between real (genuine/authentic) audio and fake (artificial) audio as the parameters with which they measure its performance. The performance metrics used are accuracy, precision, recall, and F1-score. To conduct this analysis, the system has been designed to minimize the number of false positive errors (i.e., occurrences in which real audio is incorrectly identified as fake), as well as the number of false negative errors (i.e., occurrences in which fake audio is incorrectly identified as real).

B. Comparative Feature Analysis

Researchers compare two traditional methods of detecting deepfake audio. Traditionally, deepfake audio can only be identified through basic spectral characteristics. The new method of detecting deepfake audio, created through a combination of acoustic analysis created from Mel-frequency cepstral coefficients (MFCCs) and Natural Language Processing (NLP)-based semantic analysis, is able to identify small deviations (changes) in the way a person speaks and inconsistencies in the context of what they say. As a result, the performance of this new detection system can be classified as being higher than both traditional methods of detecting deepfake audio.

C. Robustness Against Deepfake Attacks

Researchers assess the detection capability of the new method using various data collection techniques, including:

1) text-to-speech (TTS)
2) voice conversioN (VC) and
3) replay attacks.

Results obtained using the detection system demonstrate that this new detection system can achieve a higher overall generalization and reliability in performance throughout a variety of methods of generating deepfake audio than traditional methods.

V. FUTURE SCOPE

Although the suggested Deepfake Audio Detection system using MFCCs and NLP has more accurate and continued integrity than its ability to identify them, there remain several opportunities for enhancements that will further enhance the systems capacity to be applied in the real world:

1. State-of-the-Art Deep Learning Models:

The addition of more advanced architectures such as transformers; self-supervised models (like wav2vec and Hubert); and fine-tuned versions of deep learning algorithms will all improve the accuracy of detecting deepfakes across multi-lingual datasets.

2. Multilingual and Cross-Domain Support:

In the future, the inclusion of support for multiple languages and accents will enable the system to detect deep fakes in global settings for audio communication.

3. Real-Time Detection Systems:

Models can be further optimized to allow real-time audio processing and to support deployment in use cases such as verification of voice-to-voice calls and voice command on mobile devices.

4. Security Systems integration:

Combining biometric authentication and cybersecurity systems with the system can aid in the prevention of certain types of voice-based fraud and identity theft.

5. Strength and Generalization:

Further research may focus on the enhancement of detection capability against invisible deepfakes and adversarial attacks, thereby enhancing the model's generalization performance.

6. Explainable AI (XAI):

To promote trust and transparency, other methods could be considered for providing explainability to users as to why audio is classified as a deepfake.

7. Scalable Cloud Deployment:

Deployment on cloud-computing resources will enable processing and integration on a large scale with enterprise level security systems.

VI. CONCLUSION

The present study proposes a method for Deepfake Audio Detection, which combines MFCC (Mel Frequency Cepstral Coefficients) and Natural Language Processing (NLP) techniques. The intention is to increase the reliability and safety of vocal communication systems. Existing methods for detecting Deepfake Audio generally employ simple sound characteristics (e.g., timbral qualities) of a recording, making it difficult to distinguish between genuine and duplicated recordings. The proposed method combines MFCC and NLP to increase the reliability and safety of audio detection methods.

The system combines both audio and linguistic features to improve the accuracy of identifying real and fraudulent audio samples. This means that the proposed system has superior reliability and security over existing techniques that only reference basic audio characteristics when determining if an audio sample is fraudulent. Different methodologies are used for generating deep fakes using audio, such as text-to-speech synthesis, voice conversion, and replay attacks, within the proposed system.

This proposed system provides a more dependable and effective way to uncover counterfeit audio files, and it is useful for a variety of areas, including digital forensics and cyber security. Overall, it enhances both the stability and security of audio communication systems, and is better in establishing confidence in digital communication systems.

REFERENCE

1. Audio Deepfake Detection: A Survey – a comprehensive overview of features, classifiers, datasets, challenges, and evaluations in deepfake audio detection. Audio Deepfake Detection: A Survey (arXiv)
2. Audio Deepfake Detection Using Deep Learning (Siamese CNN) – uses Siamese CNN architecture and MFCC features for robust detection across diverse datasets. Audio Deepfake Detection Using Deep Learning (ResearchGate)
3. Deepfake Audio Detection via MFCC Features Using ML – combines MFCC feature extraction with SVM, Gradient Boosting, and VGG-16 models. Deepfake Audio Detection via MFCC Features (ResearchGate)
4. Deepfake Audio Detection Using Feature-Based and Deep Learning – hybrid approach using wavelet transforms, ANN, and ResNet50 classification. Deepfake Audio Detection Using Feature-Based and Deep Learning (TheSIA)
5. Efficient Deepfake Audio Detection with Spectro-Temporal Deep Learning – CNN + LSTM hybrid model leveraging spectral and temporal features. Efficient Deepfake Audio Detection (ESRJournal)
6. Improving Deepfake Audio Detection: SVM + MFCCs – a classic ML approach showing strong performance using spectral features. Improving Deepfake Audio Detection with SVM (IJISAE)
7. Audio Deepfake Detection in Voice Authentication – critical review of spectral feature-based techniques (spectrograms, MFCC, CQT) and CNN integration for verification. Deepfake Audio Detection in Voice Authentication (ETASR)