IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Verbasafe: Innovations In AI-Powered Voice Authentication

Umesh Nemane[2], Prathmesh Nikam[2], Anish Nimbal[2], Muddasar Sayyad[2], Kalash Mahajan[2], Prof. Shyamsundar Magar[1],

[1]Professor, Department of Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

[2]BE Student, Department of Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

Abstract: This project introduces a deep learning-based approach for detecting synthetic speech by leveraging a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. As AI-generated voices become increasingly realistic, the ability to accurately differentiate between human and synthetic speech has emerged as a significant challenge—impacting areas such as digital security, voice authentication, and the spread of misinformation. To tackle this problem, the model was trained on a custom dataset consisting of both genuine human speech and AI-generated audio samples. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each audio clip to capture essential frequency-based features of speech. These features were reshaped into a standardized input format of (130, 156) for consistent processing. The CNN layers were utilized to identify local spectral patterns, while the LSTM layers focused on modeling the temporal structure of the audio signals. The resulting system achieved high accuracy in classifying real versus synthetic voices and was deployed within a real-time web application built on a full-stack architecture.

Keywords: Synthetic voice detection, MFCC, deep learning, CNN-LSTM

Introduction

In recent years, the emergence of AI-generated voices has raised serious concerns around media authenticity, voice fraud, and human-AI interaction. The study referenced earlier proposed a system for detecting whether a given audio clip contains a real human voice or a synthetic one. It focused on leveraging machine learning techniques to capture the subtle differences between genuine and AI-generated speech, highlighting the increasing need for such systems in domains like security and content verification.

Building upon that foundation, this project implemented and evaluated a practical deep learning-based system to distinguish between real and synthetic voices using Mel-Frequency Cepstral Coefficients (MFCC) as the primary feature. A CNN-LSTM model was trained on a curated dataset of real and synthetic voice samples, enabling accurate classification. The system accepts audio input via a web interface, processes it in real time, and outputs the detection result. This technical implementation emphasizes the effectiveness of MFCC features and sequence-based models for voice classification and paves the way for future work in real-time analysis, emotional tone detection, and cross-language voice verification.

I. MOTIVATION

The rapid advancement of AI-driven speech synthesis has created an urgent demand for effective methods to detect synthetic and deepfake voices. As deep learning and text-to-speech technologies continue to evolve, distinguishing between authentic and AI-generated audio has become increasingly difficult. High-profile cybersecurity incidents—such as a case where an AI-cloned voice was used to deceive a mother into transferring money—highlight the serious risks associated with this technology. Global cybersecurity reports reveal that voice-based fraud involving synthetic audio has resulted in substantial financial losses and eroded trust in voice-driven systems and communication platforms. This project aims to tackle these challenges by developing a real-time synthetic voice detection system using deep learning. By offering a reliable mechanism to identify fake audio, the research seeks to strengthen the security and credibility of voice interactions, ultimately contributing to the fields of cybersecurity, fraud prevention, and AI ethics.

II. RELATED WORK

1. A Noval Feature via Color Quantisation for Fake Audio Detection

Author: Zhiyong Wang; Xiaopeng Wang; Yuankun Xie et al.

Previous deepfake detection approaches have largely relied on pre-trained models such as wav2vec 2.0 and Masked Autoencoders, utilizing reconstruction-based or mask-and-predict strategies to identify synthetic audio. While these methods enhance performance, they often lack interpretability. This paper proposes a novel feature extraction technique that applies color quantization to spectral images, constraining reconstruction by reducing the color range. This limitation helps highlight subtle differences between real and fake audio, making the reconstruction process more interpretable. Experiments conducted on the ASVspoof2019 dataset demonstrate that this method not only improves classification accuracy but also shows that pretraining the recolor network enhances the detection of synthetic audio.

2. Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization

Author: Vinaya Sree Katamneni ;Ajita Rattani

This paper tackles the challenge of detecting multi-modal deepfakes, especially those involving both audio and visual manipulations, which threaten social and political stability. The authors introduce an innovative multi-modal attention framework based on recurrent neural networks (RNNs) to bridge the distributional gap between audio and visual modalities. By applying attention mechanisms across multi-modal, multi-sequence inputs, the proposed method effectively identifies and localizes deepfakes. Experimental evaluations on various deepfake datasets—including FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF—demonstrate superior performance, with accuracy and precision gains of 3.47% and 2.05%, respectively, compared to existing methods.

3. AI-Synthesized Voice Detection Using Neural Vocoder Artifacts

Author: Chengzhe Sun ,Shan Jia ,Shuwei Hou, Siwei Lyu

This study addresses the escalating threat of synthetic human voices used for impersonation and disinformation by proposing a method to detect vocoder-induced artifacts in audio signals. Neural vocoders, which synthesize waveforms from Mel-spectrograms, are widely used in DeepFake audio generation. To counter this, the authors present a multi-task learning framework based on the RawNet2 model, incorporating a shared feature extractor alongside a vocoder identification module. By leveraging vocoder identification as a pretext task, the model is guided to focus on vocoder-specific artifacts, thereby enhancing its ability to distinguish between real and synthetic audio. Experimental results demonstrate that the enhanced RawNet2 achieves strong performance in the binary classification of synthetic human voices.

4. **Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion**

Author: Jordan J. Bird; Ahmad Lotfi

This study confronts the ethical challenges posed by AI-generated speech, including voice cloning and real-time voice conversion, by introducing the DEEP-VOICE dataset. The dataset features real speech samples from eight public figures alongside AI-generated counterparts produced using Retrieval-based Voice Conversion. Statistical t-tests on temporal audio features reveal significant distributional differences between authentic and synthetic speech. By applying hyperparameter tuning to various machine learning models, the study achieves high detection accuracy, with Extreme Gradient Boosting reaching 99.3% classification accuracy and enabling real-time detection at a speed of 0.004 milliseconds per second of audio. The DEEP-VOICE dataset is publicly released to support further research in AI speech detection.

5. **Real-time detection of spoken speech from unlabeled ECoG signals: A pilot study with an ALS participant**

Author: Angrick; M.; Luo et al.

This pilot study explores a novel approach for speech decoding in Brain- Computer Interfaces (BCIs) for individuals with speech loss due to conditions like ALS and brainstem stroke. Unlike traditional methods requiring time-aligned target representations, a graph-based clustering technique was used to identify temporal speech segments solely from electrocorticographic (ECoG) signals. These segments trained a voice activity detection (VAD) model without relying on acoustic voice recordings. Testing with a dysarthric ALS participant achieved a median error rate of 0.5 seconds and real-time VAD latency of 10 ms. This approach marks a significant step in enabling speech decoding for patients who can no longer provide ground truth data.

III. LITERATURE SURVEY

table 1: literature survey on synthetic voice detection

Sr. No.	Name of Paper	Author	Year	Objective	Algorithm	Research Gap
1.	Does Current Deepfake Audio Detection Model Effectively Detect ALM-based Deepfake Audio?	Yuankun Xie; Chenxu Xiong; Xiaopeng Wang et al.	2024	The objective of this study is to investigate the effectiveness of current deepfake audio detection models in detecting ALM-based deepfake audio.	• 12 types of ALM-based deepfake audio, training both traditional vocoder-trained and codec- trained countermeasures • Wav2Vec-XLS-R model	• The model's limited generalization to diverse audio types, high false negative rates, and the need to enhance the codec for broader audio variability.

2.	ADD 2023: Towards Audio Deepfake Detection and Analysis in the Wild	Jiangyan Yi; Chu Yuan Zhang; Jianhua Tao et al.	2024	Developing frameworks for dynamic, real-time rivalry game scenarios, improving interpretability of discrimination, improving generalization ability and robustness, considering real- time processing, considering multilingual scenarios, and exploring better evaluation metrics.	• The methods used in the study include the design of a challenging dataset, the use of various deepfake algorithms and commercial TTS platforms	• lack of considerati on for real-time processing • lack of considerati on for multilingu al scenarios
3.	Statistics-aware Audio-visual Deepfake Detector	Marcel la Astrid; Enjie Ghorb el; Djamil a Aouad a	2024	Address the limitations of existing audio-visual deepfake detection methods	• SADD, uses a shallow network architecture, waveform representation for audio input • The model is trained using the Adam optimizer with a learning rate of 10^-3 and a batch size of 8	• Convertin g the waveform to a Mel spectrogra m may introduce limitation s from the conversio n process.
4.	Utilizing Speaker Profiles for Impersonatio n Audio Detection	Hao Gu; Jiangyan Yi; Chenglo ng Wang et al.	2024	The detection of impersonation audio, and to design a large-scale, diverse-speaker impersonation dataset to advance the community's research on impersonation audio detection.	• training several existing models on the proposed IPAD dataset evaluating the performance of front-end features combined with classifiers and • end-to-end models	•The proposed IPAD dataset is limited to a specific language and accent •The models were only pretrained on genuine audio

5.	A Noval Feature via Color Quantisation for Fake Audio Detection	Zhiyong Wang; Xiaopeng Wang; Yuankun Xie et al.	2024	The objective of the study is to propose a novel method for FAD representation extraction using a recoloring network based on color quantization.	• Method uses color quantization to extract features from spectral image • The method is composed of two phases: color palette design and pixel mapping.	
6.	Scam Call Detection Using NLP and Naïve Bayes Classifier	C Valarmathi, S. Sharanya	2024	Develop a real-time scam detection system that can convert speech inputs into text and employ a classification model to detect fraudulent content.	• NLP • Machine learning algorithm (Naive Bayes)	• Naive Bayes has limitations, including its assumption of feature independence
7.	People are poorly equipped to detect AI-powered voice clones	Sarah Barrington; Hany Farid	2024	Evaluate the naturalness and identity of AI-cloned voices, and to investigate how different tasks impact our ability to distinguish AI-powered voices.	• Audio data from the DeepSpeak dataset and anonymized speaker and listener data	• Study focused on the naturalness question • and not on the identity question
8.	Real-time detection of spoken speech from unlabeled ECoG signals: A pilot study with an ALS participant	Angrick; M.; Luo et al.	2024	Develop a brain-computer interface (BCI) that can identify speech activity in real-time from ECoG signals recorded	• TICC algorithm • Graph-based clustering technique	• small amount of data • a single clinical trial participant

9.	Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models	Lam Pham; Phat Lam; Truong Nguyen et al.	2024	The objective of the study is to evaluate the efficacy of various spectrograms and deep learning approaches for deepfake audio detection.	• Transforming input audio into various spectrograms using STFT, CQT, and WT	
10.	Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes	Pavel Korshunov; Haolin Chen; Philip N. Garner et al.	2023	Dataset of deepfakes using various face swapping and voice conversion methods, with a focus on creating realistic and diverse deepfakes for research and evaluation purposes.	• DeepFaceLab • The EER threshold from the development set was used to compute the FMR and FNMR values.	• lack of clear understanding • lack of verification

IV. METHODOLOGY

a. Data Collection and Preprocessing

- A custom dataset was compiled comprising real human speech and AI-generated synthetic audio samples from various text-to-speech (TTS) engines and voice cloning tools.
- All audio files were standardized in format (e.g., sample rate, mono channel) to ensure uniformity.

b. Feature Extraction (MFCC)

- Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each audio file to capture the essential frequency and temporal characteristics of human speech.
- Each audio signal was converted into a fixed-size MFCC feature matrix of shape (130, 156), enabling consistent input across the dataset.

c. Model Architecture

- A hybrid deep learning architecture was designed, combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks.

d. Training and Evaluation

- The dataset was split into training, validation, and test sets using an 80-10-10 ratio.
- The model was trained using a binary cross-entropy loss function and optimized with the Adam optimizer.
- Performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to assess the model's ability to distinguish between real and synthetic voices.

e. Real-Time Prediction System

- The trained model was integrated into a full-stack web application.
- A user interface allows users to upload or record audio, which is then processed and analyzed by the model in real time.

- The application provides immediate feedback on whether the input voice is real or synthetic.

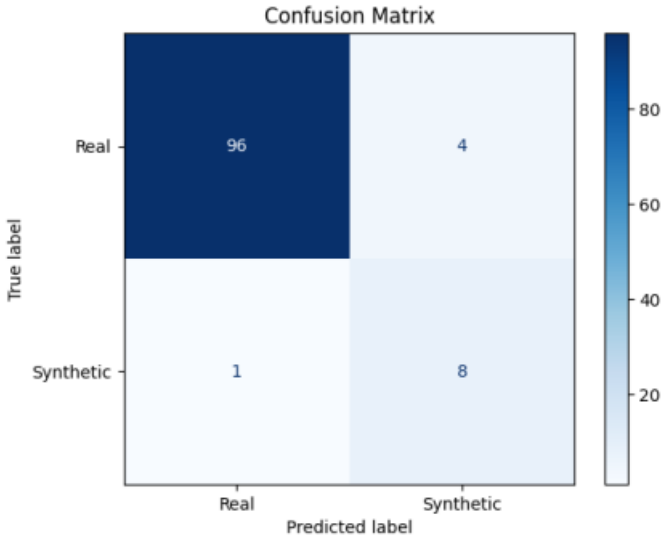
f. Deployment

- The web application was deployed using a full-stack architecture with React for the frontend, Node.js for the backend, and TensorFlow/Keras for model inference.
- The system was tested for performance, latency, and scalability in real-time usage scenarios

VI. IMPLEMENTATION

The primary objective of this implementation is to convert the proposed system design into a functional, working model that fulfills the intended purpose of detecting synthetic versus real voices accurately. Implementation is a crucial phase where theoretical concepts are transformed into practical outcomes. This section presents a detailed study of the algorithms and models employed, hardware specifications required for system execution, confirmation and preprocessing of the dataset used, and the design aspects of the system through ER diagrams, DFDs, and UML diagrams. Furthermore, the module-wise implementation steps are discussed, accompanied by sample outputs and performance results to validate the effectiveness of the system.

Algorithm Used:



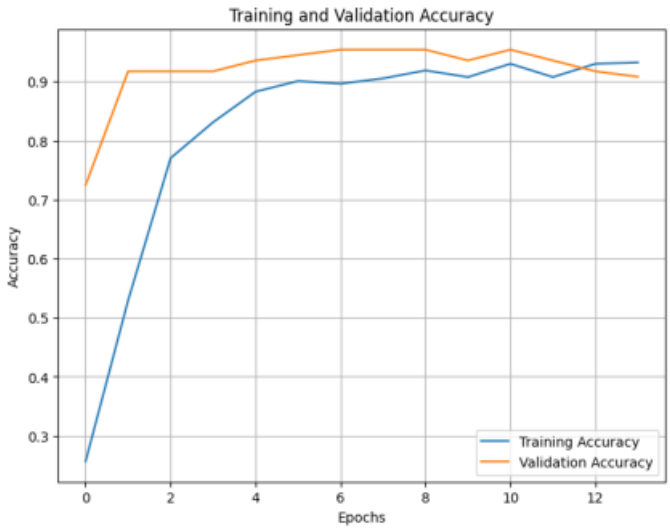
In the proposed system, a hybrid deep learning model combining a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks has been implemented for synthetic voice detection. The CNN layers are employed initially to automatically extract local spatial features from the input audio features such as MFCCs, chroma features, and spectrograms. These spatial feature maps are then passed to LSTM layers, which are specialized in capturing temporal dependencies and sequential patterns over time within the audio data. This combination allows the model to effectively understand both the local characteristics of sound and its overall temporal structure, leading to highly accurate classification between real and synthetic voices.

The CNN+LSTM architecture was chosen due to its ability to handle complex audio patterns more efficiently compared to traditional machine learning models. CNN layers reduce dimensionality and extract high-level features automatically without manual intervention, while LSTM layers solve the vanishing gradient problem and are capable of learning long-term dependencies critical in audio processing. This model, trained on extracted feature matrices of fixed dimensions, generalizes well across varying audio samples. The integration of dropout layers, batch normalization, and early stopping techniques further enhances model performance and prevents overfitting during training.

Dataset:

The dataset used for training and evaluation of our model is titled **"The Fake or Real Dataset"**, publicly available on Kaggle. It consists of a total of **5000 audio samples**, evenly distributed with **2500 real** and **2500 synthetic** voice recordings. All audio files are in **.wav format**, sampled at **16kHz**, ensuring uniformity and

quality suitable for deep learning-based audio analysis. The real voice samples in this dataset are sourced from authentic human speech, while the synthetic samples include AI-generated voices using various text-to-speech (TTS) models. This balanced and diverse dataset supports robust training and helps the model generalize well to unseen audio inputs.



Feature Extraction Module:

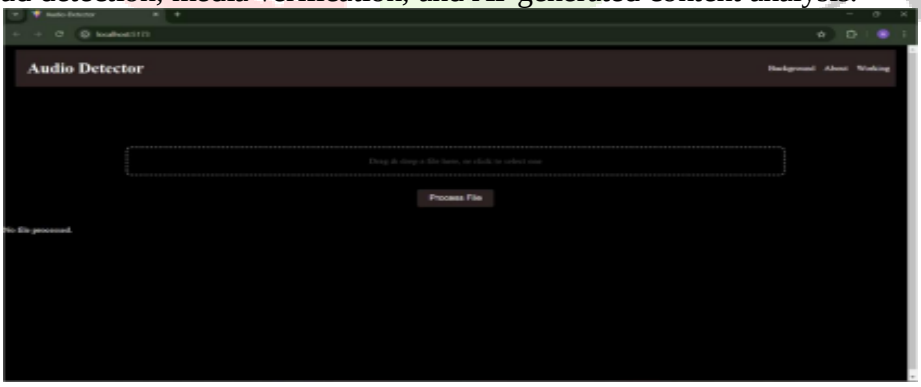
In the Feature Extraction Module, critical audio features such as MFCCs, chroma, ZCR, roll-off, RMS, and Mel-spectrograms are extracted from the preprocessed audio. This process is carried out using **Librosa** in the backend. The extracted features are reshaped into a fixed-size matrix of **130x156** dimensions, ensuring compatibility with the model. One challenge encountered was the issue of very short audio clips, which was addressed by padding the features to maintain consistent matrix dimensions. The output of this module is a well-structured feature matrix, ready to be fed into the predictive model.

Model Prediction:

The Model Prediction Module uses the trained CNN+LSTM model to classify the input audio as either real or synthetic. The backend loads the pre-trained model and processes the extracted features to make predictions. The integration between **Node.js** and Python is managed using API calls and child processes. One challenge faced during this integration was ensuring seamless communication between the two environments, which was resolved by implementing proper error handling and asynchronous processing. The model outputs a prediction (e.g., "Real Voice" or "Synthetic Voice") along with a confidence score, indicating the model's certainty in its classification

V. OUTCOME

The system achieves high accuracy and robustness by extracting meaningful audio features such as MFCC, chroma, ZCR, and spectrograms, enabling it to capture both spatial and temporal patterns in voice signals. Real-time predictions are displayed on a web interface, providing users with instant feedback along with confidence scores. This outcome demonstrates the potential of the system for practical applications such as fraud detection, media verification, and AI-generated content analysis.



VI. CONCLUSION

AI voice detection systems are essential for addressing the risks posed by generative AI technologies like voice cloning and real-time voice conversion. This survey highlights significant advancements in real-time detection, with optimized machine learning models such as Extreme Gradient Boosting achieving over 99% accuracy and sub-millisecond detection times. These developments enable robust applications for mitigating Deepfake voice threats in areas like security, fraud prevention, and content authentication. The availability of public datasets and ongoing research into advanced models and ethical practices will be crucial for further enhancing the accuracy and reliability of these systems in real-world scenarios.

VIII. REFERENCES

[1] Mouna Rabhia, Spiridon Bakirasb, Roberto Di Pietroc (2024). Audio - deepfakedetection: Adversarial attacks and countermeasures.

[2] Zhiyong Wang, Xiaopeng Wang, Yuankun Xie, Ruibo Fu1, Zhengqi Wen, Jianhua Tao4, Yukun Liu, Guanjun Li, Xin Qi, Yi Lu, Xuefei Liu, Yongwei Li (2024). A Noval Feature via Color Quantisation for Fake Audio Detection.

[3] Xie, Chenxu Xiong Xiaopeng Wang, Zhiyong Wang, Yi Lu, Xin Qi, Ruibo Fu,Yukun Liu, Zhengqi Wen, Jianhua Tao, Guanjun Li, Long Ye (2024). Does Current Deepfake Audio Detection Model Effectively Detect ALM-based Deepfake Audio?

[4] Vinaya Sree Katamneni , Ajita Rattani (2024). Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization.

[5] Hafiz Malik , Raghavendar Changalvala (2024). Fighting AI with AI: Fake Speech Detection using Deep Learning.

[6] Marcella Astrid, Enjie Ghorbel, Djamila Aouada (2024).STATISTICS-AWARE A VISUAL DEEPFAKE DETECTOR.

[7] A. V. Nadimpalli and A. Rattani. Proactive deepfake de tection using gan-based visible watermarking. ACM Trans. Multimedia Comput. Commun. Appl., Sep 2023.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks