IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

VoiceSecure: AI Voice Detection System

Anish Nimbal, Umesh Nemane, Prathmesh Nikam, Muddasar Sayyad
UG Student, UG Student, UG Student, UG Student
Department of Information Technology
Zeal College of Engineering and Research, Pune, India

Prof. Shyamsundar Magar
Assistant Professor
Department Of Information Technology
Zeal College of Engineering and Research, Pune, India

Abstract: The rise of AI-generated voices has blurred the line between synthetic and human speech, raising concerns in areas such as voice authentication, media integrity, and human-AI interaction. This project aims to develop a deep learning-based system capable of distinguishing between AI- generated and real human voices from uploaded MP3 files. Leveraging advanced text-to-speech models like Tortoise TTS to generate AI voice samples, these are compared to real human voice recordings collected from public datasets. Audio features such as Mel-frequency cepstral coefficients (MFCCs), pitch, and spectral contrasts are extracted using the Librosa library, and a Random Forest classifier is trained to recognize patterns that differentiate AI from human voices. The system will be deployed through a user-friendly web interface where users can upload audio files and receive a classification indicating whether the voice is AI-generated or human. Future expansions include real- time detection, support for multiple languages, and the integration of voice emotion analysis, offering practical applications in securing voice-based systems and maintaining the authenticity of audio content.

Index Terms - AI- generated, Deep Learning, Real-time detection.

I. Introduction

This project aims to develop a system that can distinguish between AI-generated and real human voices in audio files. With advancements in AI voice technology, synthetic voices have become so realistic that they are often indistinguishable from human speech. This creates concerns in areas like voice-based security systems, media authenticity, and interactions between humans and AI. To address this, the system will analyze audio files by extracting relevant voice characteristics and training a machine learning model to recognize patterns that differentiate human voices from AI-generated ones. Users will be able to upload MP3 files to a web-based platform, where the system will process the audio and determine whether the voice is AI-generated or human. The project has practical applications in improving security for voice authentication systems, ensuring the integrity of media, and enhancing our understanding of human-AI interactions. Future developments could include real- time detection, multilingual support, and further analysis of voice emotions to make the system more robust and versatile across various applications.

II. Motivation

The exponential growth of AI-driven speech synthesis has created a pressing need for robust solutions to detect synthetic or deepfake voices. With advancements in deep learning and text-to-speech systems, distinguishing between genuine and AI- generated audio has become increasingly challenging. Recent cybersecurity incidents, such as a case involving an AI-cloned voice used to deceive a mother into transferring funds, underscore the critical threat posed by this technology. According to global cybersecurity reports, audio-based fraud using synthetic voices has led to significant financial losses, undermining trust in voice- based systems and communication platforms. This project seeks to address these vulnerabilities by leveraging deep learning techniques to build a real-time synthetic voice detection system, enhancing the security and authenticity of voice interactions. By providing a reliable detection mechanism, our research aims to mitigate the impact of deepfake audio and safeguard against fraudulent activities, thereby contributing to the broader field of cybersecurity and AI ethics.

III. Relative Work

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

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

Previous deepfake detection methods primarily focus on pre-trained models like wav2vec2.0 and Masked Auto Encoder, leveraging reconstruction or mask and prediction techniques for fake audio detection. These methods improve model performance but suffer from poor interpretability. This paper introduces a novel feature extraction method using color quantization to constrain reconstruction by limiting spectral image colors. This approach enhances the ability to intuitively observe differences between real and fake audio in spectral reconstruction. Experiments on the ASVspoof2019 dataset show that the proposed method improves classification performance and that pretraining the recolor network benefits fake audio detection.

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

Author: Vinaya Sree Katamneni ;Ajita Rattani

This paper addresses the challenge of detecting multi-modal deepfakes, particularly those involving audio-visual manipulation, which pose significant risks to societal and political integrity. The authors propose a novel multi- modal attention framework using recurrent neural networks (RNNs) to handle the distributional modality gap between audio and visual data. By applying attention to multi-modal, multi-sequence representations, the method effectively detects and localizes deepfakes. Experimental results on several deepfake datasets (FakeAVCeleb, AV- Deepfake1M, TVIL, and LAV-DF) show improved accuracy and precision by 3.47% and 2.05%, respectively, outperforming existing approaches.

C.*AI-Synthesized Voice Detection Using Neural Vocoder Artifacts*

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

This study addresses the growing threat of synthetic human voices used in impersonation and disinformation by proposing a method to detect artifacts from vocoders in audio signals. Neural vocoders, commonly used in DeepFake audio synthesis, generate waveforms from Mel-spectrograms. The authors introduce a multi-task learning framework for the RawNet2 model, sharing a feature extractor with a vocoder identification module. By treating vocoder identification as a pretext task, the model focuses on vocoder artifacts, improving the binary classification of synthetic voices. Experimental results show that the improved RawNet2 achieves high classification performance in detecting synthetic human voices.

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

Author: Jordan J. Bird; Ahmad Lotfi

This study addresses the ethical threats posed by AI-generated speech, such as voice cloning and real-time voice conversion, by introducing the DEEP- VOICE dataset. It contains real speech from eight public figures and AI-converted speech using Retrieval-based Voice Conversion. Through statistical t-testing of temporal audio features, significant distribution differences between real and AI-generated speech is identified. Hyperparameter optimization of machine learning models enables accurate detection, with Extreme Gradient Boosting achieving a 99.3% classification accuracy and real-time classification at 0.004 milliseconds per second of speech. The dataset is publicly available to advance research on AI speech detection.

E.*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.

VI. Literature Survey

Sr. No.	Name of Paper	Author	Year	Objective	Methodology	Limitation
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 limitations of the study include the lack of generalization to new audio types, high false negative rates, and the need for enriching the codec with a variety of audio types.
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 consideration for real-time processing • lack of consideration for multilingual scenarios
3.	Statistics-aware Audio-visual Deepfake Detector	Marcella Astrid; Enjie Ghorbel; Djamila Aouada	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	• Converting the waveform to a Mel spectrogram may introduce limitations from the conversion process.

4.	Utilizing Speaker Profiles for Impersonation Audio Detection	Hao Gu; Jiangyan Yi; Chenglong 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.	Efficient Speech Detection in Environmental Audio Using Acoustic Recognition and Knowledge Distillation	Drew Priebe; Burooj Ghani; Dan Stowell	2024	This study is to design and execute experiments to optimize deep neural networks for real-time speech detection, focusing on applying knowledge distillation to create streamlined student models that parallel the larger Eco VAD teacher model's performance.	• MobileNet V3- Small-Pi architecture	• Limited range of distillation techniques • Nondeterministic experiments
8.	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
9.	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
10.	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	•
11.	Towards the Development of a Real-Time Deepfake Audio Detection System in Communication Platforms	Jonat John Mathew; Rakin Ahsan; Sae Furukawa et al.	2024	Assess the viability of employing static deepfake audio detection models in real-time communication platforms and to propose strategies and frameworks for enhancing these models.	• Resnet • LCNN architectures	• static deepfake audio models may not consistently exhibit robust performance on real-time scenarios

12.	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 understandin g • lack of verificati on
13.	PITCH: AI-assisted Tagging of Deepfake Audio Calls using Challenge-Response	Govind Mittal; Arthur Jakobsson; Kelly O. Marshall et al.	2024	Develop a challenge- response method to detect and tag interactive deepfake audio calls, and to explore the benefits of challenge- response mechanisms in mitigating the risks associated with realtime voice conversions.	• NISQA	• lack diverse high- pitch voice samples

VII. PROPOSE SYSTEM

The proposed AI voice detection system involves a layered architecture based on deep learning for audio processing. It

begins with an input layer that accepts audio files, followed by a convolutional layer to extract features from the audio signal.

A maxpooling layer is then applied to reduce data dimensionality while preserving key features. This process is repeated with

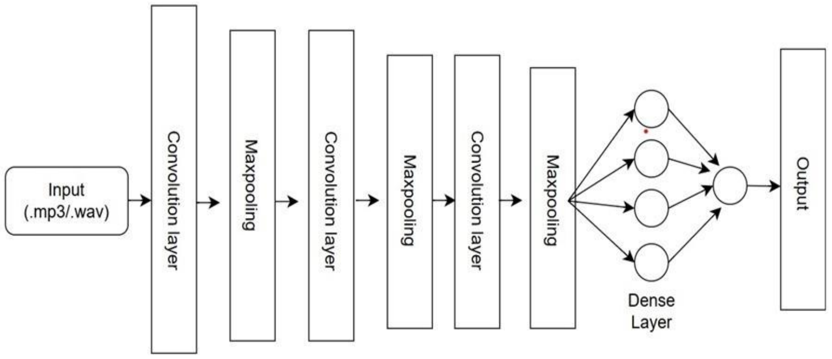
subsequent convolution and maxpooling layers, enhancing the feature extraction at each stage. After the final maxpooling

layer, the data is passed to a dense layer, which acts as a fully connected neural network to process and combine extracted

features for the final output. The design ensures efficient feature extraction and robust audio processing, making it suitable

for detecting AI-generated voices in real-time environments. This architecture leverages the strength of Convolutional Neural

Networks (CNNs) to handle complex audio features, enhancing accuracy and scalability in voice detection systems.



High level model architecture

VIII. EXPECTED OUTCOME

The proposed AI voice detection system is expected to deliver robust and accurate classification of real and AI-generated voices, including deepfake audio and voice impersonation, with high detection accuracy. It will enable seamless real-time detection through low latency processing, making it suitable for communication platforms and fraud prevention systems. By incorporating spectral color quantization techniques, the system will provide enhanced interpretability, allowing intuitive differentiation between real and synthetic audio signals. The integration of multi-modal attention frameworks will effectively detect and localize audio-visual deepfakes. Additionally, the system is designed to generalize across diverse datasets, audio codecs, and emerging generative AI methods while maintaining computational efficiency through lightweight architectures like MagicNet, ensuring deployment in resource-constrained environments. These advancements will establish the system as a scalable and adaptable solution for addressing the challenges posed by AI generated voice manipulation.

IX. METHODOLOGY

This approach presents a systematic framework for detecting synthetic voices through deep learning techniques, ensuring accurate and real- time classification of audio inputs. It combines advanced feature extraction, model training, and deployment into a robust system that operates efficiently on standard hardware without cloud dependencies.

9.1 Data Collection and Preprocessing

Collect a comprehensive dataset of both genuine human speech and synthetic audio generated by various text-to-speech (TTS) systems and deep learning models. Apply preprocessing techniques such as noise reduction, normalization, and feature extraction using tools like Librosa to enhance audio quality and consistency.

9.2 Feature Extraction

Extract relevant acoustic features, including Mel-frequency cepstral coefficients (MFCCs), spectrograms, and chroma features, to differentiate synthetic voices from natural human speech. Use TensorFlow and Keras libraries to process these features for training deep learning models.

9.3 Model Selection and Training

Design and implement deep learning models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), tailored for audio classification tasks. Train the model on labeled data with supervised learning techniques, optimizing parameters to achieve high accuracy and minimal false-positive rates.

9.4 Real-Time Detection System Development

Integrate the trained model into a real-time detection system using Python and the MERN (MongoDB Express.js, React.js, Node.js) stack for seamless front-end and back-end communication. Implement an interface that accepts audio input,performs real-time feature extraction, and classifies the input as synthetic or authentic.

9.5 Evaluation and Performance Tuning

Evaluate the system's performance using standard metrics such as accuracy, precision, recall, and F1 score on test datasets. Tune the model by adjusting hyperparameters and retraining with augmented datasets to improve robustness against sophisticated deepfake audio.

9.6 Deployment and User Interface

Develop a user-friendly web application that displays detection results in real-time with clear visual feedback and actionable alerts. Ensure the system operates efficiently on commodity hardware without cloud dependencies for offline and secure usage.

This methodological framework ensures a comprehensive and scalable solution, addressing the growing threat of synthetic voice fraud and enhancing trust in voice-based systems.

X. 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.

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] Yuankun 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 AUDIO- 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 He. Attngan: Fine-grained text to image generation with attentional generative adversarial network.