IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI GENERATED AND HUMAN AUDIO DETECTION

[1]Prof. K. S. Warke, [2]Siddhi Choughule, [3]Ketki Dandgavale , [4]Anjali Mundhe
[1]Professor, [2]Student, [3]Student, [4]Student
[1]Computer Department
[1]BVCOEW, Pune, India

Abstract: The existing fake audio detection systems often rely on expert experience to design the acoustic features or manually design the hyperparameters of the network structure. However, artificial adjustment of the parameters can have a relatively obvious influence on the results. It is almost impossible to manually set the best set of parameters. The pervasive presence of manipulated, fabricated, and counterfeit audio recordings has raised serious concerns regarding misinformation, identity theft, and privacy violations. In response to the growing threat of manipulated and counterfeit audio recordings, our project, "AI-Generated and Manipulated Audio Detection," aims to develop a fully automated end-to-end solution. This system addresses the limitations of existing fake audio detection methods, which often rely on manual parameter adjustments, by proposing an efficient and automated approach. Our approach leverages a convolution neural network (CNN) framework, utilizing speech waveforms and acoustic features such as MFCCs to extract high-level representations while considering prosody differences between genuine and fake speech. The project begins with comprehensive data collection, assembling a diverse dataset of audio samples, and annotating them to distinguish between authentic and manipulated content. To enhance the system's efficiency and adaptability, we introduce feature selection and extraction techniques, with a focus on MFCCs for robust feature representation. The machine learning model employed is an SVM (Support Vector Machine), offering effective classification capabilities. In addition to the technical aspects, our project prioritizes user-friendliness and accessibility. We provide an interactive user interface that allows users to input audio in various formats. The system seamlessly converts these inputs into the required WAV format for further processing, simplifying the user experience. This comprehensive approach, combining automated feature selection, MFCC based feature extraction, and an intuitive user interface, empowers our system to accurately detect AI-generated and manipulated audio. This not only contributes to safeguarding against misinformation and privacy violations but also ensures that the detection process is accessible and user-friendly for a wider audience.

.

Keywords: Fake audio detection, Acoustic features, Manipulated audio, CNN framework, MFCCs, CNN classification, User-friendly interface, data annotation, Accessibility.

1. INTRODUCTION

The increase in fake audio recordings has caused big problems in spotting lies, protecting identities, and keeping things private. Detecting fake audio usually involves adjusting settings and relying on experts, which isn't always reliable. So, it's crucial to create automatic ways to find AI-made and human audio.

Using fancy techniques like MFCCs helps extract important audio details, making detection better. Scientists have improved speech representation and made smarter neural networks. Most methods for finding deepfake audio rely on deep learning, especially Convolutional Neural Networks (CNNs). Research shows that deep learning, especially Convolutional Neural Networks (CNNs), is best at spotting fake content. Using big datasets like FF++ and focusing on accuracy has helped improve detection. But, it's still hard to find deepfake audio, so we need new and strong ways to detect it.

Our project, "AI-Generated and Manipulated Audio Detection," aims to contribute to the evolving landscape of audio forensics by developing end-to-end solution. We combine CNNs and MFCCs, our system endeavors to accurately detect and distinguish between ai generated and human audios.

2. LITERATURE REVIEW

Developed by Ameer Hamza et al., the system introduces an innovative method for deepfake audio detection utilizing MFCC features and machine learning algorithms. Through experiments on the Fake-or-Real dataset, they demonstrate SVM's superiority in most subsets and VGG-16's exceptional performance on the for-original dataset. Their contributions include proposing a transfer learning-based approach, leading to significant accuracy improvements, with the VGG-16 model achieving an impressive 93% accuracy. This study marks a significant advancement in deepfake audio detection methodology and experimentation.

Wang et al. pioneered a fully automated end-to-end fake audio detection system, circumventing manual parameter tuning and feature design pitfalls. They harnessed wav2vec for speech representation and innovatively introduced light-DARTS for neural network optimization. Achieving a remarkable 1.08% EER on ASV spoof 2019 LA, their method surpassed existing single systems. Notably, their approach exhibited robustness across ASVspoof2021 DF and ADD 2022 Track 1 datasets, showcasing its versatility and efficacy in diverse settings.

Rana et al.'s systematic review on deepfake detection methods (2018-2020) underscores deep learning's prominence, notably CNN models, and advocates for amalgamating diverse strategies to enhance accuracy. Their contributions feature novel approaches like recurrent convolutional models and Deepfake Stack, an ensemble learning method. Despite metric standardization challenges, the review offers pivotal insights, affirming deep learning's efficacy in deepfake detection and charting pathways for future research.

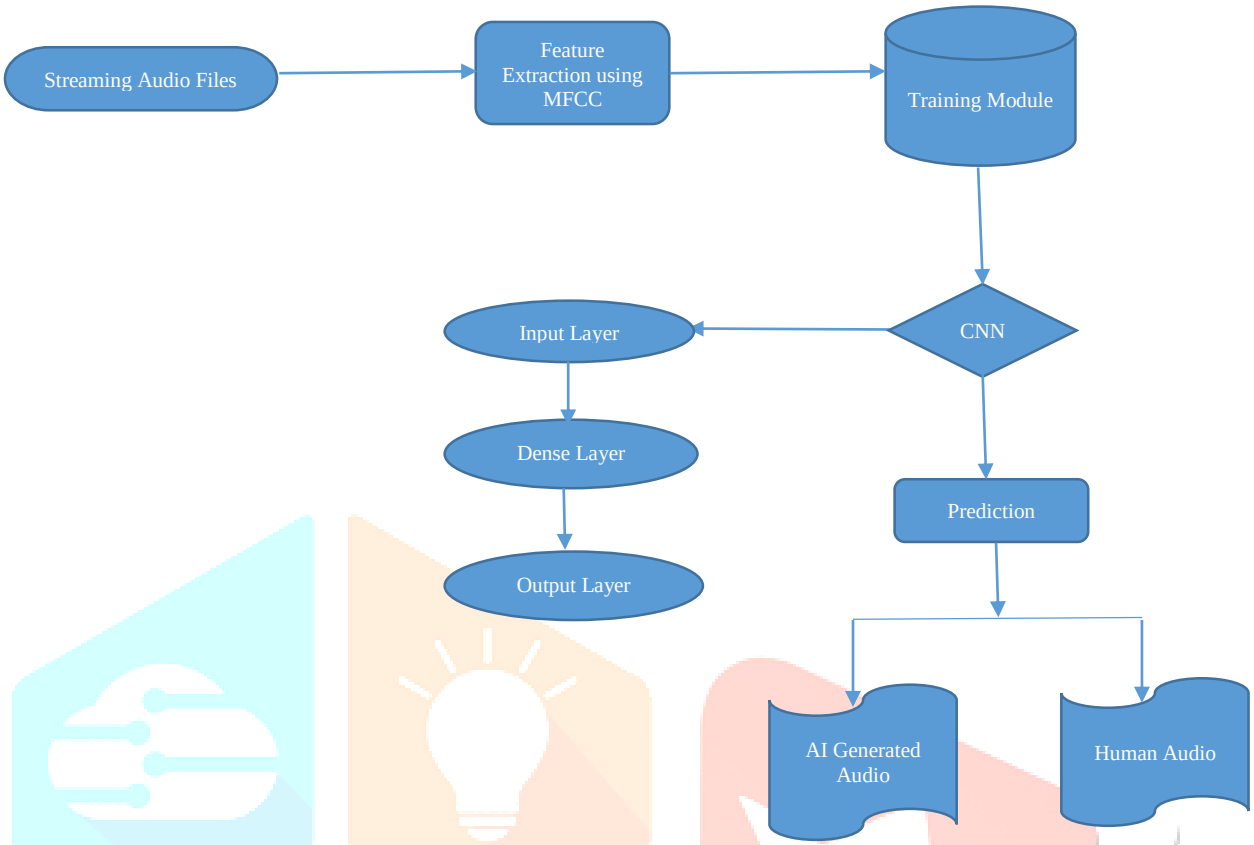
In their cutting-edge study, "Reimagining Fake Audio Detection: The Jitter-Shimmer Synthesis," Kai Li, Xugang Lu, Masato Akagi, and Masashi Unoki introduce a groundbreaking fusion of jitter and shimmer attributes within a Mel-spectrogram framework, propelled by a novel LCNN architecture with bi-directional recurrent layers. Their research, validated on the challenging ADD2022 and ADD2023 datasets, yields remarkable advancements in fake audio detection efficacy, surpassing conventional methods. Moreover, their exploration into the nuanced effects of different F0 estimation algorithms on shimmer feature performance offers invaluable insights, leading to optimized feature combinations and significant reductions in EERs. This pioneering work not only revolutionizes multimedia security but also establishes a new benchmark for authentication methodologies.

In their groundbreaking study titled "Unveiling Deceptive Sound: A Novel Approach to Fake Audio Detection," Shilpa Lunagaria and Mr. Chandresh Parekh, from Raksha Shakti University, pioneer an innovative method leveraging deep learning, including wav2vec and light-DARTS algorithms. By amalgamating Tacotron2 and DeepVoice3 models and preprocessing audio into mel-frequency spectrograms, they engineer a robust detection framework. Their pioneering contributions revolutionize the field of audio forensics, providing unparalleled insights into uncovering manipulated audio content.

In their innovative pursuit , "AI-Synthesized Voice Detection Using Neural Vocoder Artifacts" Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu introduce a novel multi-task learning framework aimed at identifying synthetic human voices by discerning neural vocoder artifacts. Their pioneering system, leveraging the unprecedented LibriSeVoc dataset, achieves an impressive Equal Error Rate (EER) of merely 0.13%, highlighting the crucial role of vocoder artifacts in voice authentication. Intra-dataset and cross-dataset evaluations demonstrate the effectiveness of the proposed approach in detecting synthetic human voices. The method shows robustness against common post-processing operations like resampling and background noise. This seminal work not only advances synthetic voice detection but also establishes a new performance benchmark in the field. These results highlight the significance of vocoder artifacts in detecting AI-synthesized voices and provide a promising direction for future research in this area.

In their exploration of audio deepfake perceptions among college students, Gabrielle Watson, Zahra Khanjani, and Vandana P. Janeja leverage the MelGAN framework to develop a robust survey and synthetic audio clips. Their study demonstrates that students exhibit better accuracy in identifying complex sentences and shorter clips, with prior knowledge of deepfakes yielding mixed results. Moreover, they find that academic backgrounds influence discernment amidst societal uncertainties, with non-computing majors displaying higher accuracy (79%) compared to computing majors (76%). Their research delves into various parameters such as grammar complexity, clip length, and political connotations, shedding light on how prior knowledge of deepfakes and academic backgrounds influence perception. Despite challenges like sample imbalance, their work provides valuable insights into the implications of deepfake technology, particularly among college populations, amidst societal uncertainties like political upheaval and the COVID-19 pandemic.

3. SYSTEM ARCHITECTURE



Designing a system architecture for AI-generated and human audio detection involves several components working together to analyze and classify audio signals accurately. Here's a high-level overview of the architecture :

1. **Data Creation:** Utilized a dataset such as the one available on Kaggle containing 26,000 human-generated and 26,000 AI Generated audio files. These files will serve as the primary training data for the model. Ensure the dataset is properly labeled to indicate which files are human-generated.

2. **Preprocessing**: Applied preprocessing techniques to the raw audio data, including Mel-frequency cepstral coefficient (MFCC) extraction with 12 coefficients. Extraction of 4 additional relevant features from the audio files. Store the preprocessed MFCC features along with the other 4 features on the system for further processing.

3. **Feature Extraction using MFCC**: After preprocessing, we've applied MFCC feature extraction to the preprocessed audio data.Segment the audio into short frames and apply a series of processing steps, including Fourier transforms, Mel filterbanks, and cepstral analysis, to compute the MFCCs. Store the extracted MFCC features on the system for use in model training.

4. **Training Module** : In the Training Module, data is split into training, validation, and test sets, and a CNN model is designed, compiled, and trained using the training set. The model's performance is validated with the validation set, and its generalization capability is evaluated using the test set.
5. **CNN Model**: Designed a Convolutional Neural Network (CNN) model with a total of 10 layers, including 1 input and 1 output layers for improved performance.

a. Input Layer: The input layer was configured to accept the preprocessed MFCC features along with the other 4 extracted features. This layer served as the entry point for the audio data into the neural network.
b. Dense Layer: Dense layers were employed to learn high-level representations from the extracted features. These layers were fully connected, meaning each neuron in a dense layer received input from every neuron in the previous layer, facilitating complex non-linear transformations.
c. Output Layer: The output layer was designed with appropriate activation functions to classify the input audio into AI-generated or human-generated categories. It produced the final predictions based on the learned features from the preceding layers.

6. **Prediction**: Deployed the trained CNN model for real-time prediction by streaming and preprocessing audio samples, then interpreted the model's output probabilities to classify them as either AI-generated or human-generated, and outputted the prediction result accordingly.

4. CONCLUSION

Employing a sophisticated deep neural network (DNN) architecture and prioritizing Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction, the system showcases a nuanced understanding of acoustic characteristics, thereby enhancing its capability to differentiate between genuine and manipulated audio recordings. By integrating a CNN for classification, the model achieves heightened accuracy in identification. Moreover, the user-friendly interface not only ensures accessibility but also facilitates seamless interaction, catering to a diverse range of users. The iterative approach adopted in data refinement and model optimization underscores a commitment to precision and adaptability, addressing the evolving challenges posed by AI-generated and manipulated audio. This comprehensive solution not only contributes significantly to the advancement of audio forensics but also serves as a robust tool in combating misinformation and safeguarding privacy in the digital age.

5. FUTURE SCOPE

1. **Advanced Deep Learning Techniques**: Explore and integrate advanced deep learning techniques such as recurrent neural networks (RNNs) and attention mechanisms to improve the system's ability to capture subtle nuances in AI-generated and manipulated audio.
2. **Real-time Detection Capabilities**: Enhance the system to perform real-time detection of fake audio during live audio streams or recordings, providing immediate feedback and intervention in scenarios where rapid detection is critical.
3. **Multimodal Approach:** Incorporate multimodal analysis by combining audio features with visual cues from spectrograms or lip movements, leveraging the synergy between different modalities to enhance detection accuracy and robustness.
4. **Continual Learning and Adaptability**: Implement continual learning algorithms to enable the system to adapt and evolve over time, learning from new data and emerging audio manipulation techniques to stay ahead of evolving threats.
5. **Cross-Domain Application**: Explore applications of the detection system beyond fake audio detection, such as identifying deepfakes in videos or detecting audio forgeries in forensic investigations. Adapting the system's algorithms and methodologies for cross-domain use cases can broaden its impact and relevance.

6. REFERENCES

1. Chengzhe Sun, Shan Jia, Shuwei Hou, Siwei Lyu; "AI-Synthesized Voice Detection Using Neural Vocoder Artifacts" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 904-912.
2. Hamza et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning," in IEEE Access, vol. 10, pp. 134018-134028, 2022, doi: 10.1109/ACCESS.2022.3231480.
3. Kai, Li & lu, Xugang & Akagi, Masato & Unoki, Masashi. (2023). Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3301616.
4. Rana, Md & Nobi, Mohammad & Murali, Beddhu & Sung, Andrew. (2022). Deepfake Detection: A Systematic Literature Review. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3154404.
5. Wang, J. Yi, J. Tao, X. Chen, Z. Tian, H. Sun, H. Ma, C. Fan, "Fully Automated End-to-End Fake Audio Detection," in Year 2022.
6. Watson, Gabrielle & Khanjani, Zahra & Janeja, Vandana, “Audio Deepfake Perceptions in College Going Populations” in Year 2021.
7. Bird, Jordan & Lotfi, Ahmad, “Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion” in Year 2023
8. Williams, Ross. (2024). Voice in the Machine: AI Voice Cloning in Film. 13. 129-144. 10.5281/zenodo.10443451.
9. Amirjalili, Forough & Neysani, Masoud & Nikbakht, Ahmadreza. (2024). Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English literature
10. Gong, Chen. (2023). AI voices reduce cognitive activity? A psychophysiological study of the media effect of AI and human newscasts in Chinese journalism. Frontiers in Psychology. 14. 10.3389/fpsyg.2023.1243078
11. Baris, Antonios. (2024). AI covers: legal notes on audio mining and voice cloning. Journal of Intellectual Property Law & Practice. 10.1093/jiplp/jpae029.
12. Ijiga, Onuh & Idoko, Idoko & Enyejo, L.A & Akoh, Omachile & Ugbane, Solomon & Ibokette, Akan. (2024). * Corresponding author: Onuh Matthew Ijiga Harmonizing the voices of AI: Exploring generative music models, voice cloning, and voice transfer for creative expression. 10.30574/wjaets.2024.11.1.0072.
13. Efthymiou, Fotis & Hildebrand, Christian & Bellis, Emanuel & Hampton, William. (2023). The Power of AI-Generated Voices: How Digital Vocal Tract Length Shapes Product Congruency and Ad Performance. Journal of Interactive Marketing. 59. 10.1177/10949968231194905.
14. Chen, Wenjun & Jiang, Xiaoming. (2023). Voice-Cloning Artificial-Intelligence Speakers Can Also Mimic Human-Specific Vocal Expression. 10.20944/preprints202312.0807.v1.

15. Choudhary, Surbhi & Kaushik, Neeraj & Sivathanu, Brijesh. (2023). EXPLORING THE CRITICAL SUCCESS FACTORS OF AI-BASED VOICE ASSISTANTS: A TEXT MINING AND STRUCTURAL TOPIC MODELLING APPROACH. 2456-9011. 10.31620/JCCC.09.23/09