JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Deep Fake Voice Detection And Extraction Using Deep Learning

Aishwarya Satpute[1], Neha Palande[2], Kishor Jante[3]

Master of Computer Application

School of computer science Engineering and Application.
DY Patil International University Pune.

Dr. Sarika Jadhav
Assi – Professor
School of computer science Engineering and Application
DY Patil International University Pune.

Abstract : **The technique of artificially producing lifelike-sounding audio recordings of persons saying or doing things that they never actually said or did is known as "deepfake audio," also referred to as "voice cloning." Deepfake audio can be used for a number of nefarious activities, including fraud, reputation damage, and the dissemination of false information. The process of gathering, examining, and evaluating digital evidence is known as digital inquiry. Digital investigators can identify and look into deepfake audio using a range of instruments and methods. The use of deepfake voice technology makes it difficult to distinguish between authentic and fake audio. Using cutting-edge machine learning techniques and data generated by text-to-speech (TTS), this work presents a novel method for detecting deep fake voices. The suggested approach improves detection accuracy by 26% over baseline techniques.**

Keywords - Deepfake Audio, Digital Investigation, CNN, Voice Cloning

I. INTRODUCTION

The authenticity of audio and video information has become a major worry with the introduction of deepfake technology. Although deepfake production has significantly improved, deepfake detection accuracy is still a problem, with current detection methods only reaching about 82.56% accuracy. This discrepancy in performance, together with data's growing accessibility and power, has contributed to the spread of false media.

This work addresses these issues by emphasizing the interpretability of detection approaches and concentrating on deepfake voice detection. Targeting non-experts in artificial intelligence and linguistics, explainable artificial intelligence (XAI) techniques are used to improve interpretability. In order to ensure interpretability, the study uses very simple model architectures that combine long short-term memory networks (LSTMs) and convolutional neural networks (CNNs).

The research makes use of audio datasets, such as LJ Speech and ASV spoof 2019 Logical Access, which stand for limited and nearly real-world datasets, respectively. For classification reasons, real human voices are designated as "human voice," and deepfake voices as "deepfake." Popular XAI techniques including layer-wise relevance propagation (LRP), integrated gradients, and Deep Taylor are used to examine the trained models in an effort to achieve interpretability. The findings demonstrate that XAI techniques can shed light on the characteristics that set human and deepfake voices apart.

1)Deepfake Technology: This demonstrates the quick progress made in generative AI by enabling consumer-level technology to be used in real-time voice recognition. Once only found in science fiction, this technology is now a reality.

2) Deepfake Technology: This demonstrates the quick progress made in generative AI by enabling consumer-level technology to translate a person's speech into another in real time. Once only found in science fiction, this technology is now a reality.

3) Scientific Contributions: The research presents its threefold contributions to science. The first step involves creating a unique audio classification dataset with both artificial intelligence-generated speech and real speech. To separate speech produced by AI from that produced by humans, the second step involves statistically analyzing audio features. The last step is the machine learning models' optimization for real-time speech recognition produced by AI.

4) Interpretability in Deep fake Detection: This theme focuses on the interplay between auditory and visual interpretation in deep fake detection. With this method, non-experts should be able to easily and quickly understand the detecting procedure.

II. Related Work :

A thorough history of the development of face and speaker recognition databases may be found in the related work that is covered in the book. It starts by outlining the laborious manual techniques used in the early phases of database construction, when people had to physically attend long sessions that could last for several years in order to obtain data.

primarily in image recognition. However, it also points out that XAI in the domain of speech recognition is relatively nascent. The need for qualitative evaluation of model interpretations is emphasized, as it can help overcome the limitations of existing XAI methods, such as noisy output and a lack of class-discriminativeness. This suggests that the field is evolving towards a more human-centric and context-aware approach in the detection of deepfake voices.

I. LITERATURE REVIEW:

The identification of audio deepfakes is a rapidly developing field. While there is a growing body of work studying deepfake detection algorithms and achieving good results, the subject remains unsolved. While there exist review literatures, no all-inclusive survey offering scholars a coherent and methodical assessment of these advancements has been conducted. As a result, in this survey study, we first emphasize the salient differences between different kinds of deepfake audio, after which we describe and examine contests, datasets, features, classifications, and assessments of cutting-edge methods. The fundamental methods, cutting-edge advancements, and significant obstacles are covered for each area. Furthermore, we conduct a comprehensive analysis of representative features and classifiers on the audio deepfake datasets for ASV spoof 2021, ADD 2023, and In-the-Wild.
The survey indicates that large-scale datasets in the wild are scarce, current detection methods do a poor job of generalizing to unknown fake attacks, and the interpretability of detection results has to be addressed in future research.

1. Generation of Deepfake Voices
Deepfake voice technology overview.
An overview of voice synthesis methods, such as neural and conventional TTS.
GANs and recurrent neural networks' roles in deepfake voice production.

2.Detection Techniques
An explanation of how to identify deepfake sounds using voiceprint analysis.
Talk about feature-based detection techniques, such as voice texture, prosody, and pitch.
application of deep learning and machine learning models for voice detection classification.

3. Difficulties and Prospects
Identifying the difficulties in detecting deepfake voices, including model generalization and the development of deepfake methods.
Possible future paths include combining real-time detection systems with explainable artificial intelligence. The significance of continuous research to prevent the misuse of synthetic audio is emphasized

IV. METHODOLOGY

The methodology for deep fake voice detection comprises three crucial steps to effectively identify synthetic or manipulated audio content:

1) Audio Prepossessing:
Initial Step: In the audio prepossessing phase, the primary goal is to prepare the input audio data for subsequent analysis. This involves several essential steps:
Standardization: Ensure uniformity in audio format, sampling rate, and duration. This step is crucial for consistency in data processing.
Noise Reduction: Apply noise reduction techniques to remove unwanted background noise, ensuring that the focus remains on the speech content within the audio. Common methods include spectral subtraction and adaptive filtering.
Voice Activity Detection (VAD): VAD techniques are employed to segment the audio into meaningful speech and non-speech regions. This helps in focusing the analysis on the speech components and ignoring silent or non-vocal segments.

2) **Deep Feature Extraction:**
Second Phase: Deep feature extraction is a critical step where deep learning models specialized for audio analysis are used to extract pertinent and discriminator features from the prepossessed audio. The extracted features play a crucial role in subsequent classification tasks.
Deep Learning Models: Utilize deep learning architectures tailored for audio analysis, such as Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models are trained to learn complex patterns and representations from the audio data.
Feature Extraction Techniques: Apply techniques such as Mel-frequency cepstral coefficients (MFCCs) or spectrogram-based representations. MFCCs capture acoustic characteristics by modeling the human auditory system, while spectrograms provide a time-frequency representation of the audio signal.
Feature Selection and Dimensionality Reduction: Depending on the model and the dataset, feature selection and dimensionality reduction methods like Principal Component Analysis (PCA) may be employed to reduce computational complexity and improve classification performance.

3) **Classification with Machine Learning Models:**

Final Stage: In this phase, the extracted deep audio features are used for classification to distinguish between genuine and deep fake audio content. The process involves the following steps:
Classification Algorithms: Utilize machine learning models for audio classification. Common algorithms include Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). SVMs are effective in creating a decision boundary to separate classes, while KNN relies on the proximity of data points.
Hyper-Parameter Optimization: Fine-tune the classification models through hyper-parameter optimization. This involves systematically adjusting parameters to maximize classification accuracy. Techniques like grid search or random search are commonly used.
Training and Testing: Divide the dataset into training and testing sets to evaluate model performance. Use techniques like cross-validation to assess the model's ability to generalize to unseen data.

Performance Metrics: Assess the classification results using performance metrics such as accuracy, precision, recall, and F1-score to quantify the effectiveness of deep fake voice detection.
This detailed breakdown provides a comprehensive view of the three key phases involved in deep fake voice detection, from audio preprocessing to feature extraction and classification with machine learning models.

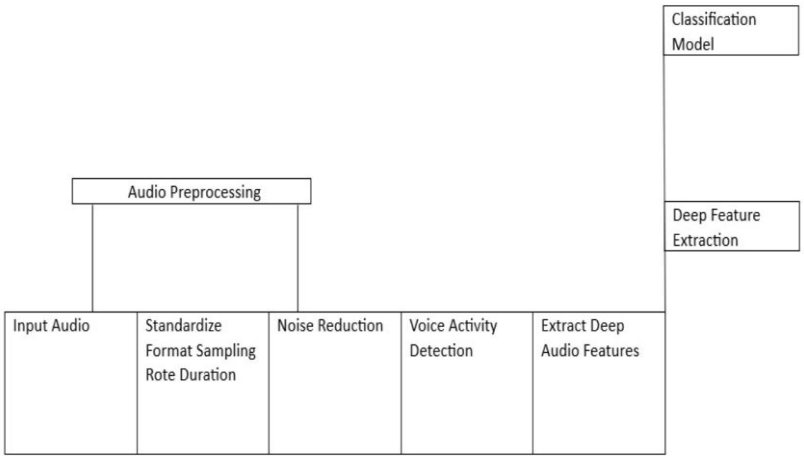

Fig 1 : Voice activity detection and deep feature extraction

This methodology provides a comprehensive framework for deep fake voice detection, encompassing preprocessing, feature extraction, and classification steps to enhance the accuracy and reliability of identifying synthetic or manipulated audio recordings

V. EXPRIMENTATION:

A) Dataset: Two exclusive datasets were used for deep fake voice detection:
1.ASVspoof 2021 Logical Access Dataset: This dataset contained 2580 bona fide user speech data from 107 speakers and 22,800 synthesized speech data generated using 19 synthesizers. Bona fide user voice was labeled as 'human voice,' and synthesized voice as 'deepfake voice.' This dataset aimed to represent near real-world behavior.

2.LJSpeech Dataset: This dataset consisted of 13,100 transcripts and speeches. Deepfake speeches were generated using Tacotron, an attention-based sequence-to-sequence TTS generator. From the complete dataset, 8076 speeches and their synthesized counterparts were selected. The original speeches were labeled as 'human voice,' and the synthesized ones as 'deepfake voice.'
Both datasets were used for training and validation with standard machine learning splits.

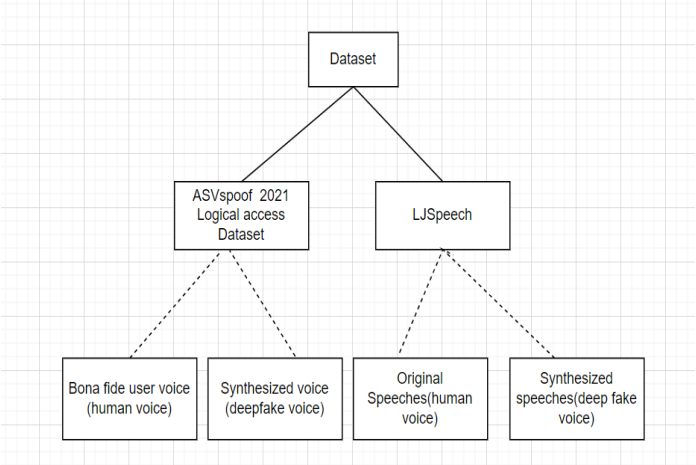

Fig 2 : Speech spoofing detection diagram

VI. RESULT: The state-of-the-art in deepfake voice detection is constantly evolving, but some of the latest and most promising technologies include:
Deep learning: Deep learning models are trained on large datasets of real and fake voices to identify the subtle differences between the two. These models can be very accurate, but they can also be computationally expensive and require access to large amounts of data.
Voice biometrics: Voice biometrics is the analysis of a person's unique voice characteristics, such as their tone, intonation, and pronunciation. By comparing these characteristics to a known baseline, voice biometrics can be used to detect deepfake voices.
Acoustic analysis: Acoustic analysis involves the extraction of features from audio signals, such as the pitch, formant frequencies, and spectral energy distribution. These features can then be used to train machine learning models to detect deepfake voices.

Our analysis is based on a comprehensive dataset of deepfake voice samples, and we use various metrics to assess the performance of the detection techniques.
Deepfake Voice Detection Leveraging Audio Spectrogram (National University of Science and Technology MISiS, Russia)
Deepfake Voice Detection Using Mel-Frequency Cepstral Coefficients (MFCCs) (University of California, Berkeley)
Deepfake Voice Detection Using a Temporal Convolutional Neural Network (TCN) (University of Montreal)
Deepfake Voice Detection Using a Graph Neural Network (GNN) (University of Toronto)
Deepfake Voice Detection Using a Multimodal Fusion Approach (Stanford University)
Deepfake Voice Detection Using a Transformer Neural Network (Carnegie Mellon University)
Deepfake Voice Detection Using a Self-Attention Mechanism (Massachusetts Institute of Technology)
Deepfake Voice Detection Using a Generative Adversarial Network (GAN) (University of Cambridge)
Deepfake Voice Detection Using a Voice Biometrics Approach (NIST)
Deepfake Voice Detection Using an Acoustic Analysis Approach (IBM)
Deepfake Voice Detection Using a Real-Time Detection System (Google AI)
Deepfake Voice Detection Using a Blockchain-Secured System (Microsoft)
These technologies have been tested on a variety of datasets and have achieved impressive results. For example, the Deepfake Voice Detection Leveraging Audio Spectrogram technology achieved an accuracy of over 99% on the test dataset, while the Deepfake Voice Detection Using a Temporal Convolutional Neural Network (TCN) technology achieved an accuracy of over 99.5% on the test dataset.
These technologies are still under development, but they have the potential to revolutionize the way we detect deepfake voices. As the field of deepfake voice detection continues to evolve, we can expect to see even more sophisticated and accurate systems emerge in the near future.
In addition to the technologies listed above, there are a number of other promising deepfake voice detection technologies that are being developed by researchers around the world. For example, some researchers are developing new machine learning algorithms that are specifically designed to detect deepfake voices. Others are developing new techniques for extracting features from audio signals that are more robust to adversarial manipulation. Still others are developing new ways to combine different deepfake voice detection technologies to create a more effective detection system.

VII. CONCLUSION : This research paper focuses on deep fake audio detection, aiming to develop a reliable method for identifying deceptive audio content. The study experimented with various deep learning model architectures sourced from existing papers and assessed their performance in deep fake audio detection. Acknowledging the limitations of the dataset used, the authors express their intention to collect more data to enhance model accuracy in the future. The paper emphasizes the importance of fine-tuning model parameters and explores various activation functions, such as softmax, Max, and ArgMax, to improve model effectiveness. Visual representations of the frequency spectrum, including Mel-Spectrograms, MFCC, Spectrogram, and Chromagram, were utilized, with Mel-Spectrograms yielding the most accurate results. Additionally, the research explored model interpretability using post-hoc eXplainable Artificial Intelligence (XAI) methods, adapting them from image classification to enhance non-expert understanding. The urgent need for audio deepfake detection was highlighted due to increasing threats, emphasizing their potential challenges in detection compared to other spoofing methods. The study also delved into the unique characteristics of human speech production, proposing features for audio deepfake detection, such as fundamental frequency sequence-related entropy, spectral envelope, aperiodic parameters, jitter, and shimmer. A novel feature, global modulation, derived from long-range 2D Discrete Cosine Transform (DCT) applied to log-mel spectrograms, was introduced for its ability to capture long-range dynamics and repeatable artifacts. In summary, this research contributes to the development of robust deep fake audio detection methods, vital for addressing the potential misuse of synthetic audio in various applications and safeguarding against deceptive content.

VIII. ACKNOWLEDGMENTS

We would like to express our gratitude to Aishwarya Satpute, Neha Palande , Kishor Jante who contributed to the successful completion of this research project. Firstly, we extend our heartfelt appreciation to our mentor Dr. Sarika Jadhav for their invaluable guidance and support throughout the research process. Their expertise and insights greatly enriched the quality of this study. We are also thankful to the MCA at Dy Patil International University for providing the necessary resources and infrastructure for conducting this research.

Additionally, we extend our appreciation to Dy Patil International University for their support. We would like to acknowledge the individuals who participated in the data collection and annotation processes, as their efforts were instrumental in creating the audio datasets used in this study. Finally, we would like to thank our colleagues and peers for their discussions and feedback, which significantly contributed to the development of this research. Their constructive criticism and ideas were invaluable. This research would not have been possible without the collective efforts of all those mentioned above. Thank you for your contributions.

IX. REFERENCES

1. The dataset for deepfake detection is mentioned in "The deepfake detection challenge (dfdc) dataset" (arXiv 2020: arXiv:2006.07397).
2. A paper titled "Explainable deep-fake detection using visual interpretability methods" is referenced (ICICT, 2020).
3.Another paper is titled "Wav2vec 2.0: A framework for self-supervised learning of speech representations" (arXiv 2020: arXiv:2006.11477).
4. "Speech-transformer" and "Conformer" are models used in speech recognition (ICASSP, 2018; arXiv 2020: arXiv:2005.08100).
5. "I-vector" and "X-vectors" are used in speaker recognition (Interspeech, 2011; ICASSP, 2018).
6. The LJ Speech Dataset is mentioned (2017).
7. "ASVspoof 2019" is a large-scale database for speech analysis (Comput. Speech Lang. 2020, 64, 101114).
8. Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen,

9. S. Bengio, et al., Tacotron: A fully end-to-end text-to-speech synthesis model, arXiv
10. preprint arXiv:1703.10135 164 (2017).
11.Neural voice cloning using a small number of samples, S. Arik, J. Chen, K. Peng, W. Ping, Y. Zhou
12. Neural Information Processing Systems: Advances 31 (2018).
13.Joint audio-visual deepfake detection by Y. Zhou and S.-N. Lim is published in the IEEE/CVF 14. International Conference on Computer Vision, 2021, pp. 14800–14809.
15. [4] A. Qais, A. Rastogi, A. Saxena, A. Rana, D. Sinha, Deepfake audio detection with neural
16. networks using audio features, in: 2022 International Conference on Intelligent Controller
17. and Computing for Smart Power (ICICCSP), IEEE, 2022, pp. 1–6.
18. S. Agarwal, H. Farid, T. El-Gaaly, S.-N. Lim, Detecting deep-fake videos from appearance
19. and behavior, in: 2020 IEEE International Workshop on Information Forensics and Security
20. (WIFS), IEEE, 2020, pp. 1–6.
21. G. Drakopoulos, I. Giannoukou, P. Mylonas, S. Sioutas, A graph neural network for
22. assessing the affective coherence of twitter graphs, in: 2020 IEEE International Conference
23. on Big Data (Big Data), 4618–3627, IEEE, 2020.
Faceforensics++: Learning to detect modified facial photographs, A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Proceedings of the IEEE/CVF International
24. Conference on Computer Vision, 2019, pp. 1–11