Localizing Deepfake Audio Detection: Enhancing Community Resilience against Synthetic Fraud

CS4991 Capstone Report, 2024

Robert Jason Hudson
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
Rjh8eq@virginia.edu

ABSTRACT

Deepfake audio technology undermines digital communication integrity, particularly in vulnerable communities. The proposed solution within focuses on developing an accessible deepfake audio detection system through advanced computer science methods like spectrogram analysis and the development of convolutional neural networks in order to effectively distinguish real from synthetic audio samples. Preliminary expectations suggest high efficacy in detecting deepfakes, based upon previously existing machine learning systems focused on audio analysis. Future work includes refining this technology through extensive testing and public interface development, aiming to enhance digital communication security and prevent misinformation.

1. INTRODUCTION

The emergence of deepfake audio technology has introduced a sophisticated form of digital deception, exploiting the nuances of human speech to create convincingly fabricated recordings. This technology, while impressive, harbors potential for misuse, particularly in scamming and defrauding unsuspecting individuals. Among the most vulnerable to these digital threats are elderly people, individuals with limited technical literacy, and non-native language speakers. These groups, often less familiar with the fast-evolving landscape of digital technology and potentially more trusting of digital communications, face a heightened risk of being targeted by scammers using deepfake audio as a tool for exploitation.

With these challenges in mind, there is a clear imperative for the development of accessible deepfake audio detection systems. These systems work by leveraging the latest in machine learning to distinguish between real and fake audio by analyzing the unique speech patterns of different communities. This targeted approach not only improves detection accuracy but also provides a safeguard for those most at risk, therein reinforcing local trust in digital communication.

2. RELATED WORKS

Evidence of the danger of synthetic audio has already been widely researched, but a particularly concerning report comes from Mai et al. (2023)[3] in which they detail how the average individual frequently struggles to distinguish between a genuine voice and its deepfaked counterpart in multiple trials of major languages. To combat this troubling development, researchers are proposing various methods for the detection of synthetic audio. A report from Togootogtokh and Klasen (2024)[6] explores the use of Generative Adversarial Networks (GANs) for detection in their work titled

"AntiDeepFake." While acknowledging limitations in generalizability and cost of the described model, their report illustrates a promising initial step towards robust and adaptable detection methods. Finally, Miotti and Wasil (2024) shifts focus to the broader societal implications in their work "Combatting deepfakes." They explore policy solutions to mitigate national security risks and protect individual rights, emphasizing a multi-pronged approach encompassing technology, legal frameworks, and public awareness campaigns. These reviewed works provide a strong foundation for understanding the current landscape of deepfake audio's social implications as well as detection research.

3. PROPOSAL DESIGN

My proposed design presents a thorough approach to developing and implementing a deepfake audio detection system that encourages the testing of potentially fabricated audio in an accessible and efficient manner. Although it is in the conceptual phase, the proposal lays the groundwork for practical development and eventual implementation.

3.1 Development

The bulk of this proposal consists of creating the algorithm that distinguishes between real and synthetic or deepfake audio. First, audio data must be converted into a spectrogram, which is a visual representation of the spectrum of frequencies a sound contains over a period of time. The process begins by dividing an audio file sample into short frames, each subjected to a window function that smoothly tapers the amplitude of the audio signal at the beginning and end of each frame to zero, minimizing abrupt changes in the audio. This, in turn, makes the frames easier to process for the Short-Time Fourier Transform (STFT). The STFT is a mathematical transform that converts these time-domain frames into the frequency domain, which consists of points that can be plotted to show how the frequency content changes over the duration of the audio, producing a spectrogram.

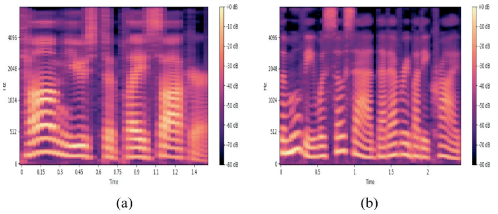

Figure 1: Mel-spectrograms examples comparing a normal audio sample (a) and a synthetic one (b)[1]

There are multiple types of spectrograms depending upon how the data is derived from the audio, but the base methodology of analyzing a spectrogram allows for the detection of a potential synthetic audio sample. In Figure 1, note how the second spectrogram has more irregularities and discontinuities in its structure. These are the obvious signs that the audio file has a synthetic source.

To create an algorithm that can recognize these indicators, the spectrogram must be converted further into a data structure that machine learning models can effectively analyze. The choice of model is important because it will dictate the format and nature of this transformation due to certain models being better at handling certain data types over others. Previously discussed were Generative Adversarial Networks[5]. These operate by containing two competing networks: a generator that creates synthetic data and a discriminator that learns to differentiate between the synthetic and real data. GANs benefit from being able to generate new data the model has not seen before, which helps it learn to catch a wider range of fake audio. However, GANs can be computationally intensive and complex to train, which makes them less desirable for a system that needs to be accessible and efficient.

Instead, a different machine learning model called a Convolutional Neural Network (CNN) is selected. CNNs are a machine learning model known for their proficiency in pattern recognition and ability to learn hierarchical representations, meaning they can automatically detect and learn various features in the input data at multiple levels of abstraction. In order for these layers to properly read the data of the spectrogram it must be converted into a mel-spectrogram which can be seen in figure one. Different from an ordinary spectrogram, a mel-spectrogram applies a mel scale filter bank to the audio signal after the STFT, translating the sound into a 2D image-like format with time on the x-axis, frequency on the y-axis mapped to the mel scale, and the intensity of the frequencies represented through color or brightness. This format plays to the strengths of CNNs, which were originally designed for image data.

The architecture for the proposed deepfake audio detection project involves several key components of CNNs:

3.1.1 *Convolutional Layers*: These layers apply filters to the mel-spectrogram, identifying key patterns and features within small regions of the audio spectrum.

3.1.2 *Pooling Layers*: These layers simplify the data from convolutional layers, reducing its dimensionality to focus on the most critical features, thereby enhancing computational efficiency.

3.1.3 *Fully Connected Layers*: Positioned after convolutional and pooling layers, these layers integrate all learned features to classify the audio as real or synthetic based on the detected patterns.

Together, the layers form a path that transforms raw audio data into readable data for the CNN, allowing it to distinguish between genuine and manipulated audio effectively[2]. To train the CNN model, genuine audio samples will easily be gathered from online databases, while synthetic samples will come from the MLAAD (Multi-Language Audio Anti-Spoofing Dataset) is expected to be used. This dataset is valuable because it contains a wide range of languages and spoofing attack types, offering a realistic training environment for the model[5]. By training the model on a diverse dataset, the system can also improve its accuracy and generalizability across different audio types and spoofing techniques.

3.2 Implementation

To translate this algorithm into a usable web-based interface, the following technology stack is proposed. The frontend will employ the React web framework to create a responsive user interface, selected primarily due to my familiarity with its usage. For the backend, Python is selected due to its extensive libraries and community support for data processing and machine learning tasks. Python's Librosa library will be used for audio analysis as it provides simple and flexible processing of audio signals. The deep learning framework PyTorch is also chosen for its ease of use and ability to accelerate the processing of a convolutional neural network (CNN), which is specifically what is needed if the system is expected to be efficient.

The interface will allow users to upload audio in formats such as mp3, wav, or flac, as well as record live audio, which is then preprocessed by Librosa and converted into a mel-spectrogram. This data feeds into the trained CNN model, which evaluates the audio and provides the user with a real-time assessment of authenticity.

4. ANTICIPATED RESULTS

Due to the preliminary nature of this project, actual data pertaining to it is not yet available. Instead, realistic expectations can be formed based on studies conducted on

similar audio classification tasks using Convolutional Neural Networks (CNNs). For example, research conducted on the SpecRNet model, which processes spectrogram-based information, demonstrates that architectures like this are highly effective in distinguishing genuine from manipulated audios. Their findings show that SpecRNet achieved an Equal Error Rate (EER) of 0.1549, indicating a low rate of misclassification, and an Area Under Curve (AUC) of 99.9941, suggesting excellent overall model accuracy in distinguishing classes across various thresholds[2]. It would take multiple iterations of training my model to develop such high accuracy, but these findings confirm it is indeed possible to create a highly accurate model from these methodologies.

5. CONCLUSION

The proposed project addresses an urgent need in the field of digital communication security by tackling the growing challenge of accessible synthetic audio testing and detection. The system I have conceptualized utilizes spectrogram analysis and machine learning techniques to develop an algorithm that can detect Deepfake audio with a high degree of accuracy. The project is designed to be freely accessible and user-friendly, ensuring that those most vulnerable to digital fraud can easily utilize the tool. This technology is crucial in the fight against escalating incidents of misinformation and digital fraud.

Developing this proposal has greatly enhanced my academic journey at the University of Virginia. I have been able to build upon my knowledge of operating artificially intelligent neural networks as well as learn about the processes utilized by auditory processing systems. Additionally, I learned more about how scams using AI operate, and I have improved my ability to teach others about recognizing these fraudulent tactics and how to protect themselves against them.

6. FUTURE WORK

Moving forward, the immediate next step for this project involves the practical implementation of the proposed detection system. This includes development of a working prototype that can be rigorously tested for its efficacy in identifying deepfake audio. Initial testing could be conducted using a curated dataset of known real and synthetic audio samples to fine-tune the algorithm. The working prototype's user interface will also be evaluated to ensure it is accessible and caters to all groups, including those at risk.

REFERENCES

[1] Gupta, G., Raja, K., Gupta, M., Jan, T., Scott, T. W., & Prasad, M. (2024). A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *Electronics, 13*(1), 95. https://doi.org/10.3390/electronics13010095

[2] Kawa, P., Plata, M., & Syga, P. (2022). SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection. *arXiv* https://arxiv.org/abs/2210.06105

[3] Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLoS One, 18*(8), e0285333. doi: 10.1371/journal.pone.0285333

[4] Miotti, A., & Wasil, A. (2024). Combatting deepfakes: Policies to address national security threats and rights violations. *arXiv*. https://arxiv.org/abs/2402.09581

[5] Müller, N. M., Kawa, P., Choong, W. H., Casanova, E., Gölge, E., Müller, T., Syga, P., Sperl, P., & Böttinger, K. (2024). MLAAD: The Multi-Language Audio Anti-Spoofing Dataset.
arXiv, https://arxiv.org/abs/2301.09512.

[6] Togootogtokh, E., & Klasen, C. (2024). AntiDeepFake: AI for Deep Fake Speech Recognition. *arXiv*, https://arxiv.org/abs/:2402.10218