CIS 583 – DEEP LEARNING

PROJECT REPORT

Spoof Audio Detection System for Political Discourse Protection

Team Members

Hasan Arif
Sri Haritha Deevi
Fazal Rahaman Pasha Muhammed

1- Introduction

The emergence of deepfake audio technology has raised significant concerns regarding its implications for society. As artificial intelligence (AI) continues to evolve, the ability to create convincing audio imitations of individuals poses a substantial threat to trust and authenticity in communication. The importance of fake audio detectors cannot be overstated, as they serve as a critical line of defence against the potential misuse of this technology.

The impact of synthetic audio on society has already manifested in concerning ways. In January 2024, a significant incident occurred where artificial intelligence was used to create a convincing fake robocall impersonating President Joe Biden's voice, attempting to discourage New Hampshire voters from participating in the primary election. This event highlighted the immediate need for reliable spoof detection systems in political contexts.

According to Gartner's 2024 Emerging Technologies Impact Radar report, by 2026, 30% of organizations will have tools and processes to combat synthetic media and deepfakes, up from less than 10% in 2023. The World Economic Forum's Global Risks Report identifies synthetic media as one of the top technological threats to democracy and social stability.

Without effective countermeasures, the future implications of unchecked synthetic audio are severe:

- Erosion of public trust in authentic political communications
- Manipulation of financial markets through fake executive statements
- Compromise of security systems relying on voice authentication
- Escalation of social engineering attacks using synthetic voices
- Potential triggering of diplomatic incidents through fake political statements

Recent work in the field of audio spoofing detection has been significantly advanced through competitions like the ASVspoof Challenge series, which has become the de facto benchmark for anti-spoofing systems. The ASVspoof 2021 Challenge specifically focused on the detection of deepfake audio and speech synthesis attacks, demonstrating the growing sophistication of both attack and defense mechanisms in this domain.

1.1 Related Work

Several significant contributions have shaped the field of audio spoofing detection:

- The work by researchers at the University of Michigan-Dearborn has been particularly influential in developing lightweight yet effective detection systems. Their approach using modified ResNet architectures with attention mechanisms has shown promising results in detecting synthetic speech.
- The ASVspoof 2019 Challenge introduced the concept of logical access attacks, which closely relates to our current problem of detecting AI-generated voice content. The challenge's baseline systems and evaluation metrics have become standard in the field.
- Recent IEEE papers have highlighted the effectiveness of spectrogram-based approaches, particularly when combined with deep learning architectures. For

instance, the use of Constant-Q Transform (CQT) spectrograms has shown superior performance in detecting frequency artifacts common in synthetic speech.

1.2 Stakeholder Analysis

Primary Stakeholders:

- Government and Regulatory Bodies: Includes election commissions and the FCC, responsible for ensuring election integrity and regulating communications, along with law enforcement agencies that investigate election fraud.
- Political Entities: Includes political candidates targeted by voice spoofing attacks, and campaign teams or party organizations requiring verification of authentic communications.
- Technology Implementers: Includes social media platforms, telecommunications companies, and security solution providers or cybersecurity firms developing and deploying protective measures.

Secondary Stakeholders:

- Voters: Require accurate information to make informed decisions during elections.
- Media Organizations: Play a role in verifying the authenticity of content shared with the public.
- Academic Institutions: Contribute to research and development in combating voice spoofing.
- Civil Rights Organizations: Monitor election integrity and advocate for fair practices.

Key Requirements:

- Implementation of real-time detection capabilities to identify spoofing attempts.
- Maintenance of audit trials and compliance documentation to ensure accountability.
- Development of scalable and integrable solutions for seamless implementation.
- Establishment of clear incident response protocols for timely mitigation.

2- Dataset Construction and Preparation

Our dataset construction methodology builds upon established approaches in the field while addressing the specific challenges of political speech verification. The design choices are informed by successful methodologies documented in ASVspoof challenges and recent IEEE publications.

For our project focused on detecting spoofed audio specifically related to former President Joe Biden, we developed a unique dataset inspired by recent events surrounding the January 2024 presidential election. This dataset comprises original and spoofed audio samples designed to train our detection model effectively.

2.1 Data Collection

Original Audio:- We sourced multiple official addresses from former President Joe Biden, extracting the audio in .wav format. In total, we collected approximately 50 minutes of original audio.

- Collection method: Systematic downloading and audio extraction from verified government sources and official media channels on YouTube. For downloading from YouTube, we wrote a python script which utilizes “yt-dlp”, an all-in-one command line application that can download videos and extract audio from them. The script takes as inputs two parameters, namely the URL of the YouTube video and output directory where the audio files will be saved. An output directory is created systemically when one does not exist.

Spoofed Audio:- Spoofed audios were generated using the Eleven Lab API, which allowed us to create realistic imitations of for President Joe Biden's voice. We compiled around 40 minutes of spoofed audio. Each audio sample was chunked into segments of 3 seconds for analysis.

Content: Varied speech patterns and contexts to ensure robust detection capability.

2.2 Data Processing Pipeline

Our processing pipeline incorporates best practices from ASVspoof challenges and recent IEEE publications:

1- Audio Segmentation

The script split the audio files into chunks according to a particular set time, and a time each chunk ends is exactly the time that the next chunk begins. In this method, there exists a sequence of segments, each completely segmented without any overlap. For example, if the chunk size is 3 seconds, that means the script, that we wrote, would create chunks such as 0 – 3 seconds, 3 – 6 seconds and so on until the audio file is completely done. In this we, the total number of segments are 1,800, where around 1,000 belongs to the original audios while 800 belongs to the spoofed audio category.

2- Spectrogram Generation

In order to generate time-frequency representations of the sound, we utilize the Short Time Fourier Transform. The STFT is computed using a window with 2048 samples, a hop length of 512 samples and a default Hann window, where the individual amplitude values are also scaled to decibels for easier visualization. The spectrograms which are formulated are done using the jet colormap and the images are compiled in a high resolution (100 dpi) of 6x6 inches without axis allowing easy visualization of the audio’s frequency spectrum over time.

2.3 Dataset Organization

The dataset split follows standard machine learning practices while incorporating specific considerations from audio spoofing detection research:

I. Training set: 70% (1,058 spectrogram images)
II. Validation set: 20% (302 spectrogram images)
III. Test set: 10% (154 spectrogram images)

This structured approach allows us to train our deep learning model effectively while addressing the specific challenges posed by spoofed audio detection.

2.4 Comparison with Existing Datasets

Our dataset differs from existing spoofing detection datasets in several key aspects:

I. Specific focus on political speech and single-speaker detection
II. Contemporary synthetic speech generation using state-of-the-art APIs
III. Real-world application context based on actual incidents
IV. Balanced representation of both original and spoofed content

These characteristics make our dataset particularly relevant for practical applications in political discourse protection, while building upon the methodological foundations established by ASVspoof challenges and related research.

3- Model Architecture

The proposed model combines Recurrent Neural Networks (RNNs) in the form of LSTM layers for temporal dependencies and Convolution Neural Networks (CNN) for extracting spatial features This architecture effectively exploits spectrogram data obtained from audio signals and allows detection of spoofing by looking at both the spatial and temporal aspects of the data.

3.1 Convolutional Neural Networks (CNNs)

CNNs form the foundational part of the architecture and consist of four convolutional blocks. Each block contains the following components:

- Convolutional Layer: Extracts spatial features such as edges, textures, and patterns from input spectrograms.
- MaxPooling Layer: It reduces the width and height dimensions, which in turn reduces the amount of computation needed but keeps the important parts.
- Dropout Layer: Compromises the occurrence of overfitting by randomly switching off some neurons in the course of training.

Each convolution block learns a set of features in order of increasing complexity by taking the input spectrogram and producing progressively deeper feature maps.

3.2 Recurrent Neural Networks (LSTMs)

The output of the last CNN block is modified and brought into an LSTM layer. Because LSTM operates on data in a sequence, the model is able to capture the time dependencies present in the spectrograms. This is especially important in the case of audio forgery detection since temporal characteristics of speech can even suggest speaker's spoofing.

3.3 Architecture Summary

The detailed architecture of the model is presented below:

Layer	Feature Map	Size	Kernel Size	Stride	Activation Function
Input Layer	3	(256, 256, 3)	-	-	-
Conv2D	32	(254, 254, 32)	(3, 3)	1	ReLU
MaxPooling2D	32	(127, 127, 32)	(2, 2)	2	-
Dropout	32	(127, 127, 32)	-	-	-
Conv2D	64	(125, 125, 64)	(3, 3)	1	ReLU
MaxPooling2D	64	(62, 62, 64)	(2, 2)	2	-
Dropout	64	(62, 62, 64)	-	-	-
Conv2D	128	(124, 124, 128)	(3, 3)	1	ReLU
MaxPooling2D	128	(62, 62, 128)	(2, 2)	2	-
Dropout	128	(62, 62, 128)	-	-	-
Conv2D	256	(60, 60, 256)	(3, 3)	1	ReLU
MaxPooling2D	256	(30, 30, 256)	(2, 2)	2	-
Dropout	256	(30, 30, 256)	-	-	-
Reshape	-	(196, 256)	-	-	-
LSTM	64	(64)	-	-	-
Dense	1	(1)	-	-	Sigmoid

Table 1. Details of Architecture

The above proposed CNN-LSTM architecture enables:

- Spatial Learning: To identify spectrogram images, Convolutional Neural Networks utilize overlapping 2D images that are created using the convolutional layers, max pooling layers as well as the dropout layers.
- Temporal Learning: The LSTM layer accesses sequential shifts in the data after reformation of the output from the CNN allowing it to track temporal relations about the spectrogram.
- Classification: The sigmoid layer is modelled so as to yield compressed results in the form of ‘feature vectors’ which output a score of how likely the sample is to be a spoof or original audio.

This hybrid CNN-LSTM architecture is well-suited for tasks that require both spatial and temporal understanding. It combines both Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNNs) to exploit spatial and temporal features respectively. The structure has four CNN blocks. Each block consists of Conv2D layers and has been designed to have L2 regularization, and ReLU activation, as well as max pooling followed by Dropout to reduce overfitting. The outputs of the CNN blocks are then fed into the RNN block after being converted to the appropriate shape, whereby the LSTM layer is used to model sequential information. The last fully connected layer uses sigmoid activations for binary output.

The training was performed for 100 epochs with a batch size of 8 and 256×256-pixel input images. The model was built using the RMSprop optimizer with binary cross entropy as the loss function, and accuracy as the metric. In order to enhance the training procedure, certain callbacks including ModelCheckpoint, ReduceLROnPlateau, and EarlyStopping are used, enabling training to be efficient and stopping it early if no further improvement is detected with respect to validation performance. The trained model is saved for future inference, providing a robust mechanism for detecting the spoofed audio for speaker "Former President Joe Biden".

4. Experiments Analysis

The dataset used in the proposed CNN-LSTM model includes a set of spectrogram images created from the audio clips. In order to maintain class balance, a dataset was developed with all original and spoofed audio 529 each. During evaluation out of the 151 spectrograms created, as validation set during training the model evaluation set was created.

The training process started by using the spectrogram images as inputs from the audio clips to help the CNN-LSTM model learn the spatial and temporal features. Convolutional layers would first learn the spatial features out of the hierarchical spectrograms and then the outputs would be converted and fitted into the LSTM Layer. Use of the LSTM layer would allow learning sequential data fitted into the model by learning the time features allowing the model to recognize patterns related to spoofed and original audio. One hundred epochs were used in training the model while using appropriate hyperparameters for proper learning.

The figure below shows the training loss and training accuracy graph.

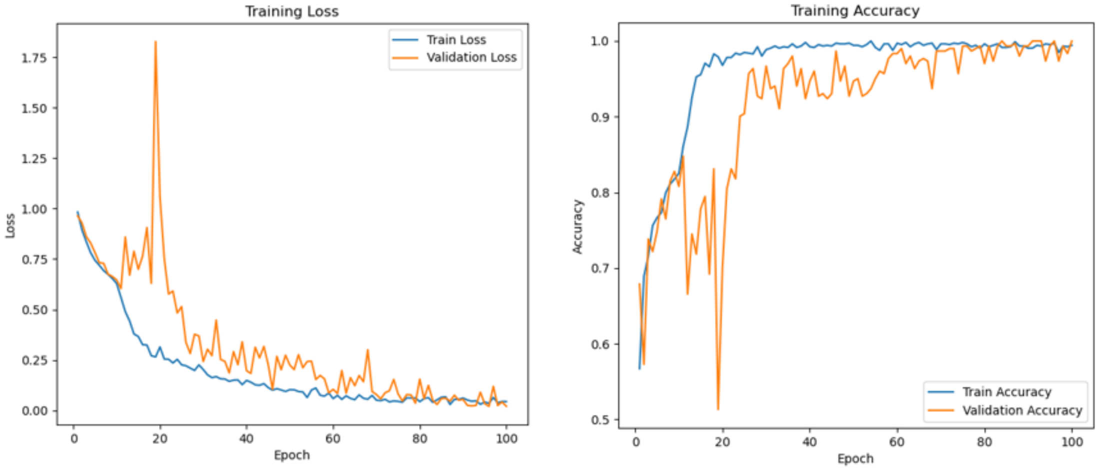

Figure 1

The graph in the above figure provides with the following information:

1- Loss (Left Plot):

- Training Loss: Decreases systematically over time, showing that the model is gaining the knowledge to minimize the errors on the training data in an improving manner.

- Validation Loss: At first tends to fluctuate with a major peak at the 20th epoch and post the peak, seems to go through a phase of reduction and later stages goes down to almost zero value.
- There's consistency with the training loss, in that the validation loss is broadly the same after roughly 50 epochs, which means that the model has good generalization and does not significantly overfit.

2- Accuracy (Right Plot):

- Training Accuracy: Went up almost in a linear manner and went so far as 95 and 100% around the 40th epoch, meaning the model was excellent at performing on the training dataset.
- Validation Accuracy: Presents some early ups and downs in training, then appears to be consistently generalizing around the 30th epoch, and around 50 epoch becomes rounded at an approximate level of the training accuracy.
- Even with variations at the beginning, the validation accuracy does become the same as the training accuracy as the epochs become larger, meaning better generalizability.

Validation accuracy and loss does have some initial ups and downs but does eventually settle down, and once it settles, then the model does show a steady convergence towards the end. Validation loss clearly shows a decrease against training loss

4.1 Evaluation Metrics

The trained model is evaluated on test dataset using standard classification metrics:

- Accuracy: 99.35%
- Precision: 1.0000
- Recall: 0.9870
- F1-Score: 0.9935

These results confirm that the model generalizes well to unseen data, achieving near-perfect performance.

4.2 Confusion Matrix

Analytical performance of the classification model to discriminate between spoofed and original audios can be summarized in a table called confusion matrix. This matrix has two classes: Spoofed and Original.

The confusion matrix also helps evaluate the performance for the trained model. The confusion matrix in figure 2 reveals that:

- True Positives (TP). Model predicted Spoofed to Spoofed in 77 cases.
- True Negatives (TN). The model predicted Original to Original in 76 cases.
- False Positives (FP). Ratios of model predicted Original to Spoofed in 1 of the cases.
- False Negatives (FN). Failure to Spoofed to be recognized as Original (value 0).

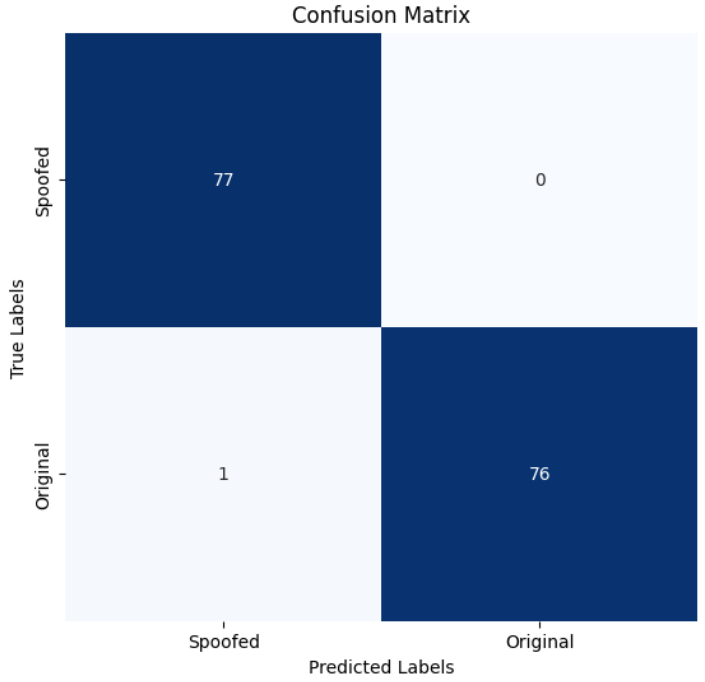

Figure 2

This matrix indicates overall good performance by the model and hints that the percentage of misclassification is low as there was one false positive case and no false negative case which is quite impressive. The accuracy and reliability of the model can also be judged from these ratios, showing a high degree of effectiveness for the discrimination of both classes.

4.3 Generalization to New Data

The trained model was further tested on new spectrogram samples. It consistently identified spoofed audios and original audios with high accuracy, demonstrating its reliability in real-world scenarios.

The model was presented with randomly selected 9 audios spectrogram images from a total sample of 10. The model was able to identify the spoofed audio and original audio accurately.

Links to Implementation and Dataset:

- Code for Audio Spoof Detection
- Dataset and Project Details

5. Conclusion and Learnings

The hybrid CNN- LSTM model combines two neural architectures- convolutional neural networks and Long Short Term Memory- which are quite strong on their own, and with the help of this model becomes really powerful and advantageous in terms of spoofed audio audio detection with the use of spectrogram based representation.

The metrics show that the model has a high level of confidence and accuracy when detecting spoofed audio. The model was able to score an accuracy of 99.35% with a precision performance of 1.0000 and a recall performance of 0.9870 and an F1 score of 0.9935 which shows that the model is very confident in its predictions. It means, the model also does not register many false positives at the expense of discrimination performance against spoof signals. The model's ability to train over 100 epochs until a steady reduction of training losses were observed along a rise of accuracy were in part due to the effective choice of parameters such as the learning rate, batch size and dropout regularization.

Furthermore, the evaluations on the new samples of data that were not part of the samples used for training overt evidence that the model can generalize better and classify the test data correctly demonstrating the relevance of model in practical sense.

This project demonstrates the effectiveness of hybrid structures such as CNN-LSTM networks in audio classification tasks, especially where a detailed spectrogram is required. And using the strengths of the two models, the model lays the groundwork to be able to recognize deepfake or counterfeit audio confidently. These results are helpful in creating better media reliability and security, where deep fakes and other audio attacks can be devastating, and deep learning techniques can be vital in defending such attacks. The project is effective in that it is likely to have further uses not only in authentication devices but also in forensic analyses of sounds as well as real time recognition of deep fake sounds.

References

1. J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "*Audio Deepfake Detection: A Survey*," arXiv:2308.14970, 2023. https://arxiv.org/abs/2308.14970

2. T. Oorloff, S. Koppisetti, N. Bonettini, D. Solanki, B. Colman, Y. Yacoob, A. Shahriyari, and G. Bharaj, "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection," arXiv:2406.02951, 2024. https://arxiv.org/abs/2406.02951

3. J. Yamagishi, X. Wang, M. Todisco, M. S. Md Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "*ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection*," arXiv:2109.00537, 2021. https://arxiv.org/abs/2109.00537

4. X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "*ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild*," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2507–2522, 2023. http://dx.doi.org/10.1109/TASLP.2023.3285283

5. M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," arXiv:1907.00501, 2019. https://arxiv.org/abs/1907.0050

6. L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models," arXiv:2407.01777, 2024. https://arxiv.org/abs/2407.01777

7. M. Alzantot et al., "Deep Residual Neural Networks for Audio Spoofing Detection," in Proc. INTERSPEECH, 2019, pp. 1078-1082. https://www.isca-speech.org/archive/interspeech_2019/alzantot19_interspeech.html

8. H. Ranka, M. Surana, N. Kothari, V. Pariawala, P. Banerjee, A. Surve, and S. Mehta, "Examining the Implications of Deepfakes for Election Integrity," arXiv:2406.14290, 2024. https://arxiv.org/abs/2406.14290