Not My President: A Logistic Regression Audio Classifier for AI-generated Speech

Adarsh Sairam Ambati

June 15, 2023

1 Introduction

One afternoon, Jennifer DeStefano's phone rang with an unknown caller id. To her dismay, when she picked up, the voice was her daughter's, Brianna crying. A man's voice in the background threatened to kidnap and kill her unless a $1 million was sent over as ransom. Only after a number of 911 calls and an ultimately successful attempt to reach her daughter, that this call was exposed as a scam.

The scam artist had scraped Brianna's private TikTok and public Instagram accounts to create a hyper-realistic audio simulation of her voice. It was only due to the quick-thinking of Jennifer that this scam was revealed [Karl].

This whole incident simply reveals how dangerous AI-generated audio can be, especially now that large language models and a growing database of human speech via social media platforms are greatly augmenting the realism of such speech.

It is not unreasonable to imagine that the future of scamming, espionage, or sabotage will heavily rely on AI-generated audio. To that end, I recognize that those committed to solving the problem of AI-content classification will always be playing catch up to those building generative models. However, this research project is a simple example that demonstrates that in the status quo, AI-generated audio still remains distinguishable from human speech. As generated speech develops further. larger, more complex neural nets are likely needed to discern artificially generated audio.

2 Background

2.1 Problem

The overall problem of generative classification is an expansive problem. AI can currently generate audio, text, visuals, and video. For the sake of scope, I limited this research project to simply audio. My justification for this is that textual data, intuitively, do not parlay enough information for the model to accurately discern between artificial and human text. Visual and video generation is not yet at the level where it seems indiscernible from the human eye. Audio data, on the other hand, lies in this middle ground where AI-generated audio will likely not have the same audio statistics (frequency, decibel, etc...) as human speech, but the artificial audio generators are still advanced enough that they could fool the unaware.

Furthermore, as this project serves as a proof-of-concept that AI audio classification is possible, we limit the scope further to be: **classifying the last three President's audio**. This may seem like an arbitrary choice, and to some extent, it is. However, available on Youtube, right now, there are thousands of minutes of data of President Obama, Trump, and Biden's artificially generated audio, mimicking them playing various popular video games or in other contrived scenarios. This coupled with their very public and long State of Union Addresses, makes them the perfect candidate for this proof-of-concept. Finally, I believe that if it is possible to distinguish artificial vs human speech in this specific instance, the problem of extending the classification to all human and artificial audio is a matter of building a bigger model and collecting more data.

2.2 Math/Probability behind Logistic Regression

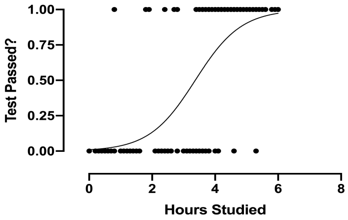
The Logistic Regression Model is one of the most foundational models in all of AI research. It is an extension of the linear regression model to the binary range: $[0, 1]$. As a recall, the linear regression model uses gradient descent and input features to identify an equation that relates the input features to an output prediction:

$$\hat{y} = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + ... + \beta_i * X_i$$

The logistic regression model takes $\hat{y}$ and plugs it into a sigmoid function:

$$\text{pred}= \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\alpha+\beta_1*X_1+\beta_2*X_2+\beta_3*X_3+...+\beta_i*X_i)}}$$

Figure 1: 1-dimensional Logistic Regression Curve



[Pad]

Then, to solve for the optimal α and β parameters, we use the Maximum Likelihood Estimation technique using the gradient descent optimization algorithm to finally, come to a model that would look as Fig. 1 does in a 1-d input scenario.

However, because I will ultimately be using thousands of 10-second audio files each with 13 MFCC features, to speed up processes, I used sci-kit-learns Logistic Regression Model Function

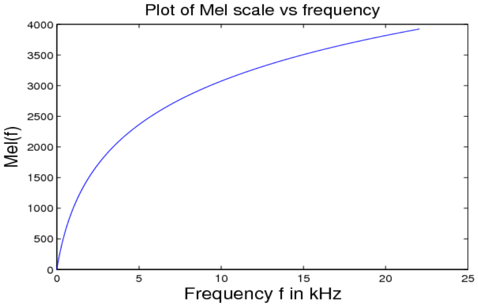
2.3 Audio Classification

This project leverages pre-existing audio classifiers, specifically dhruvesh13's "Automatic music genre classification using Machine Learning algorithms like Logistic Regression and K-Nearest Neighbours" [dhr17]. This classifier like most modern classifiers uses Mel-Frequency Cepstrum Coefficients (MFCCs) as the primary feature inputs for the model.

2.3.1 Mel-Frequency Cepstrum Coefficients

First, let's understand the Mel scale. Unlike frequency in Hz, humans do not comprehend speech in a linear fashion. We have an easier time understanding small pitch changes in lower frequencies than in high frequencies. Thus, the Mel scale is a logarithm transformation of traditional frequency: $M(f) = 1125ln(1 + \frac{f}{700})$. This scale better reflects how humans comprehend and digest sounds at different frequencies. Thus, it stands to reason that when classifying human speech from generative speech, we should use this scale. The actual computation of the MFCCs requires an understanding of

Figure 2: [Ram13]

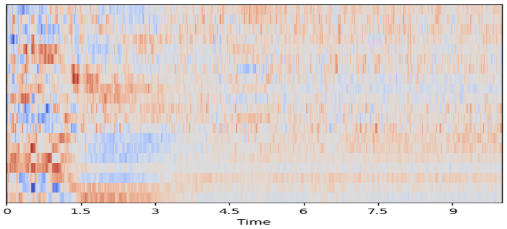


Fourier Transforms and Spectral Densities. On a high level, any signal can be decomposed into the primary frequencies via a Fourier Transform. The spectrum of a signal is average of the frequency content. To find, MFCCs, you take the Fourier Transform of the signal, map the power series of the spectrum of the signal to the MEL scale, take the logs of the powers of each mel frequency, and take the discrete cosine transform of the list of mel log powers [Pra]. Much of this calculation is beyond the scope of this class and my background knowledge.

Thankfully, the library python-speech-features contains a mfcc function that takes in a sound wave and outputs a numpy array containing the MFCCs for that given time interval. We simply have to process this array to ensure that all nans and inf are converted to zeroes to allow the logistic regression model to work correctly (infinities as feature values prevent convergence of the gradient descent during training).

These MFCCs can be used to display an MFCC spectrogram that looks like the following:

Figure 3: MFCC Spectrogram of the First 10 seconds of Biden's 2021 State of Union Address



3 Results

First, I split all the large .wav files (see Appendix) into 10-second chunks.

3.1 Attempted Image Classification: MFCC Spectrogram

Before fully understanding how to manipulate audio data, I attempted to convert all audio files into a spectrogram as Fig 3. Then, I attempted to build a logistic regression model that would classify an audio stream as artificial or real solely based on the spectrogram. However, consistently the accuracy of this model was $\sim$ **0.5**, which is as good as a random classifier. This indicated that the spectrogram alone is not enough to distinguish audio files.

3.2 Direct Approach: MFCCs

Then, I converted all these wave files into numpy arrays that represent the MFCCs. I then used an $80 - 20$ training - testing split to build the logistic regression model using the MFCCs directly. This model proved to be a resounding success.

Accuracy: 0.9748
Area Under PR Curve: 0.99
Area Under ROC Curve: 0.99
For Curve Graphs and Confusion Matrix See Appendix

Ultimately, this project demonstrates that audio classification for artificially generated speech is still viable. A simple, single-neuron logistic regression model using MFCCs with no data augmentation was able to achieve an accuracy of 0.978, with excellent PR and ROC curves, which together indicated that the model made very few false positives and few false negatives.

While in this project, we only tested and train the model on three individuals, early results from general testing (on friends and other AI-generated audio clips) are giving further evidence that this model can be expanded beyond this scope. With further training and larger models, we might not only be able to protect our Commander-in-Chief from AI-generated audio scams but even the average citizen like Jennifer.

References

[dhr17] dhruvesh13. Automatic music genre classification using machine learning algorithms like-logistic regression and k-nearest neighbours. *Github*, 2017.

[Karl] Faith Karimi. 'mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. 2023 url =.

[Pad] Graph Pad. Application of multiway methods for dimensionality reduction to music. *Research Gate*.

[Pra] PracticalCryptography. Mel frequency cepstral coefficient (mfcc) tutorial.

[Ram13] Ajay Ramaseshan. Application of multiway methods for dimensionality reduction to music. *Research Gate*, 2013.

Appendix

Human Videos

- President Biden's State of the Union Address
- Watch Joe Biden's Full 2022 State Of The Union Address
- 2014 State of the Union address (Full speech)
- President Obama Delivers his Final State of the Union Address
- President Trump 2018 State of the Union Address (C-SPAN)
- WATCH: Trump's full 2019 State of the Union address

AI Generated Videos

- US Presidents Play Five Nights at Freddy's (FNAF) FULL SERIES
- US Presidents Play Five Nights at Freddy's 2 (FNAF 2) FULL SERIES
- US Presidents Play Minecraft 1-20
- The Presidents Have a Sleepover COMPLETE SERIES!
- U.S. PRESIDENTS PLAY Mr. President! (Ai voices)

Figure 4: Confusion Matrix of MFCC Logistic Regression Model

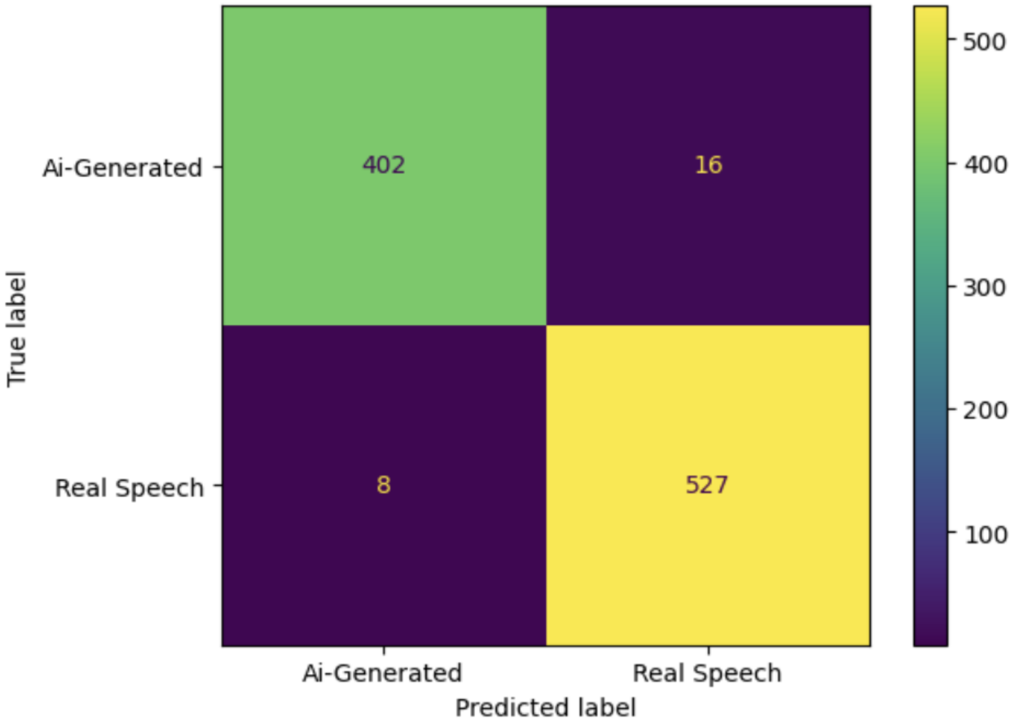


Figure 5: Reciever Operating Characteristic (ROC) Curve of MFCC Logistic Regression Model

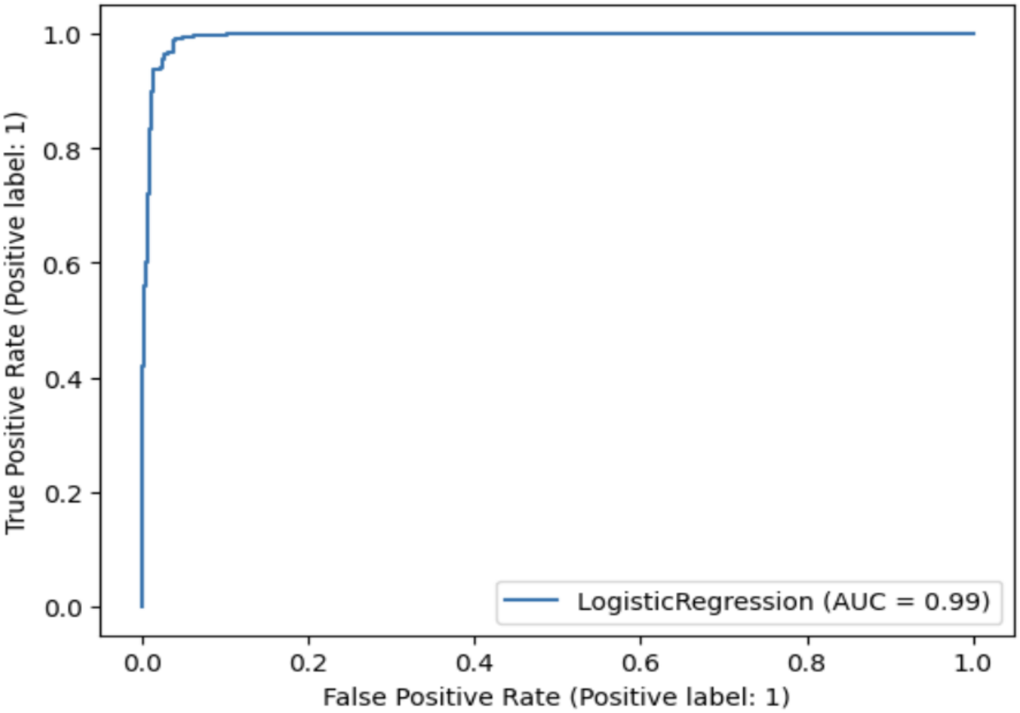
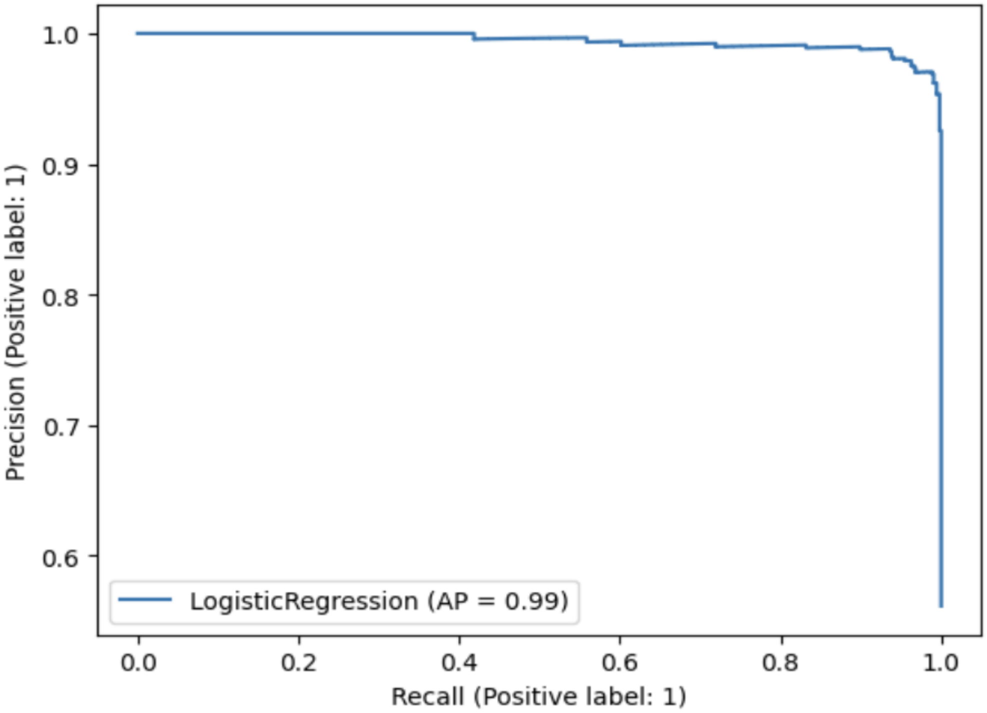


Figure 6: Precision-Recall Curve Spectrogram of MFCC Logistic Regression Model


Precision (Positive label: 1)
LogisticRegression (AP = 0.99)
Recall (Positive label: 1)