Audio Identification

Team Members :

Trisanu Bhar - S20190020256
Vikrant Reddy Y - S20190020261
VENKATA NIKHIL VADDIPATI - S20190020259

1. Introduction to the Scientific Problem

Audio is one of the major ways humans communicate through. On a daily basis, an average human identifies and classifies thousands of voices and sounds very precisely. Upon inspection, it was observed that we could mimic this action of the brain in two ways.

1. Through voice familiarity (Classification)
2. Through feel similarity (Clustering)

In this report, we will consider the former way, through voice features and try to teach a model to classify different voices from the same audio sample.

The data we used in the following project is the first US 2020 presidential debate that was conducted between Mr. Donald Trump and Mr. Joe Biden focused on the impact and measures taken during the COVID-19 pandemic.

Our main task is to distinguish three voices of :

- Donald Trump (former president)
- Joe Biden (current president as of 2021 and former vice president)
- Chris Wallace (the debate host)

The data is public and available online. Further, the data was labelled and provided on kaggle.

2. Major Concepts Used

The following concepts helped us achieve our goal efficiently:

Supervised Learning

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output.

Convolutional Neural Nets

A Convolutional Neural Network (CNN) is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data. Convolutional neural networks are distinguished from other neural networks by their superior performance with image, speech, or audio signal inputs. They have three main types of layers, which are:

- Convolutional layer
- Pooling layer
- Fully-connected (FC) layer

Recurrent Neural Network

RNN is a type of Neural Network where the outputs from the previous step are fed as input to the current step. In cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words.

Adam Optimizer for Neural Nets

Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iteratively based on training data.The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients.

Mutual Information Classification

Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

Principal Component Analysis

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. We use PCA to reduce the number of features so that we can make accurate predictions and use them to classify the voice signals into their target classes.

MEL Spectrogram

MEL spectrogram is a frequency representation of an audio signal on the Mel scale. Humans do not perceive audio in a linear scale, hence we use a different scale called the Mel scale that helps us understand the frequency response better.
For instance, humans are better at detecting differences at lower frequencies than at higher frequencies.

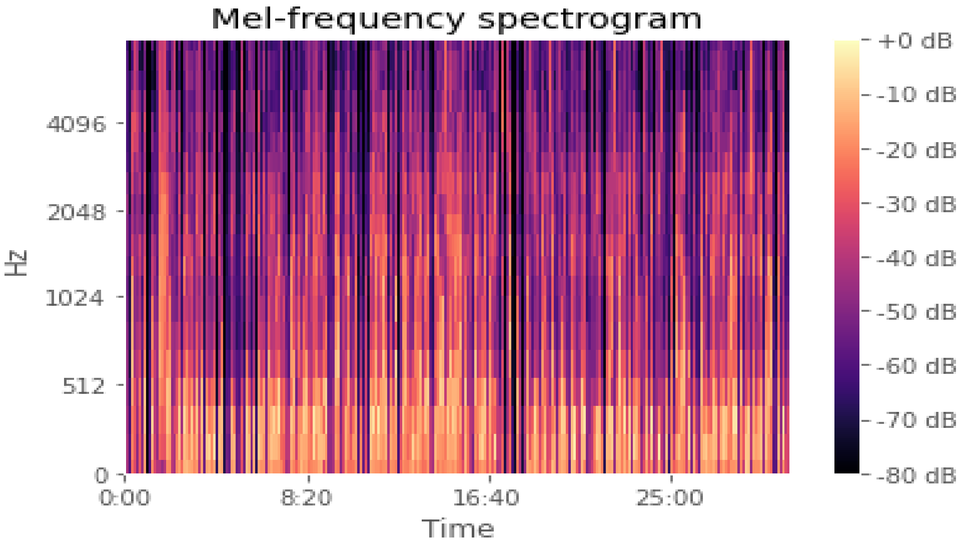
Libraries Used

- ❖ Librosa
    - ➢ Audio analysis - Reading the audio signals complete with the audio features.
- ❖ Numpy
    - ➢ Mathematical operations for working with the audio data.
- ❖ Pandas
    - ➢ Structuring data into tables/dataframes for using or dropping certain features or making edits to the entire column.
- ❖ Matplotlib
    - ➢ Plotting graphs and images

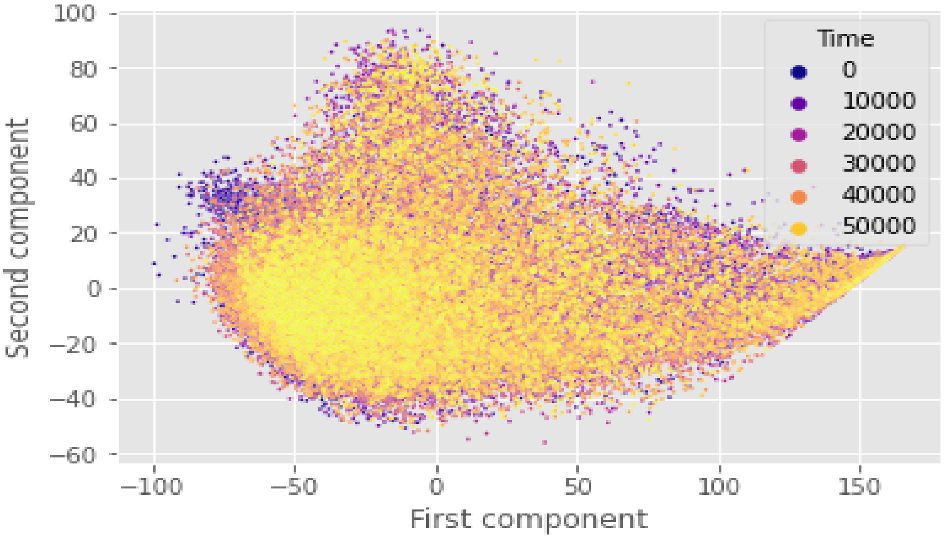
- ❖ Keras
  - ➢ Neural networks library which presents “Layers” that can be stacked on top of one another and a “Model” class that can be used to design the architecture we need.
- ❖ Seaborn
  - ➢ Additional plotting tool for complicated diagrams and images that need to be rendered with good clarity.

3. Data Visualization

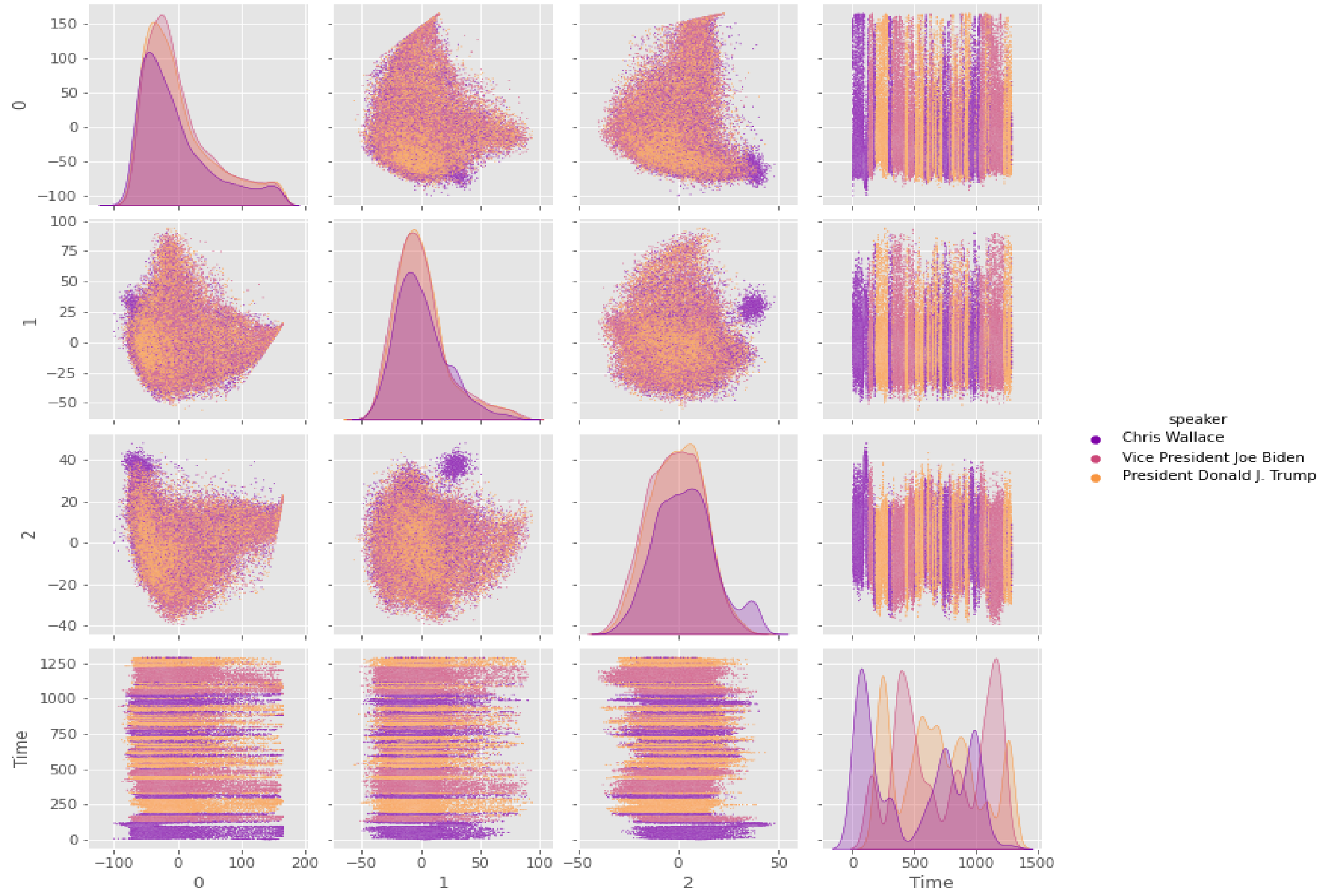
a. *The Mel Spectrogram*



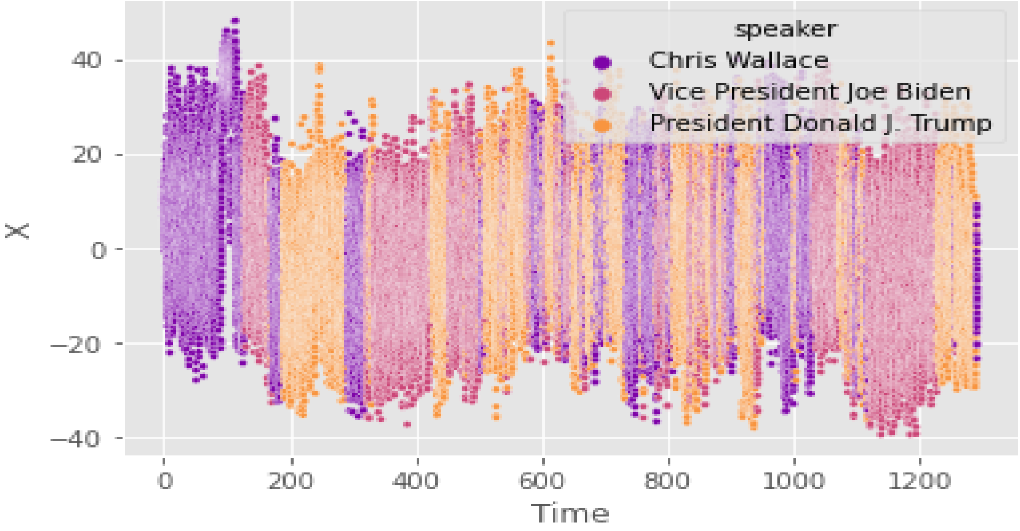
b. *PCA Visualization*



c. PCA Features pair plot



d. Best PCA Feature



4. Methodology

The goal is to use the debate data and predict the speaker label in the time series data with the help of Mel scale coefficients.

To achieve that, we extract plausible features from the data and feed them to the convolutional neural network.

1. We transform time series data into feature space using the mel spectrogram .
2. Since there will still be a lot of parameters, which will make the computation complex, using PCA we reduce the dimensionality and extract elementary features.
3. Finally we set up a Sequential model using Tensorflow and describe the above architecture using Keras layers and train the model. We then use this model for prediction.

Steps involved in this process

1. **Data Set up**
   The audio sample is trimmed and read into the python environment, later converted into frequency domain.
2. **Simplification of the feature space**
   The dimensionality of the features extracted from the audio is reduced by performing PCA and later on, two of the new features are selected by analyzing the mutual information.
3. **Neural Networks**
   The time series data received after data preparation is fed to a neural network that iteratively fits in a ReLU activation function onto the training data using the adam optimizer.

Conclusions based on the analysis have been mentioned below:

Model	Test Accuracy
Support Vector Machine	40.7%
Convolutional Neural Network	61.2%
Recurrent Neural Network	63.4%
Long Short Term Memory	55.5%
CNN + RNN	68.4%
CNN + LSTM	62.6%

5. Conclusion

Since an RNN is more suitable to work with temporal data than a CNN, we could see a better accuracy with the RNN. In the case of an LSTM, we observe a little overfitting, even though the CNN + LSTM model gave a 68.8% accuracy on the train data, the test accuracy dropped to 62.6%.

Improvements that can be made:

- The overfitting problem of the LSTM could be tackled by hyperparameter tuning.

6. Miscellaneous

Project Links

- Model Notebook
  https://colab.research.google.com/drive/1yV4DM1nv5ZqHF19RZiMkHke0pcsloJ9U
- Dataset and Project Idea
  https://www.kaggle.com/headsortails/us-election-2020-presidential-debates

References

- Neural Nets : https://www.youtube.com/watch?v=CqOfi41LfDw&list=PLblh5JKOoLUIxGDQs4LFFD--41Vzf-ME1
- Mel Spectrogram : https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
- How do humans perceive sound: https://www.coursera.org/lecture/3d-models-virtual-reality/how-do-humans-perceive-sound-dTL9Y
- Tensorflow : https://www.tensorflow.org/
- Keras- the Python deep learning API : https://keras.io/
- Matplotlib visualization with Python : https://matplotlib.org/
- Librosa : https://librosa.org/doc/latest/index.html
- LSTM: https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm

Contributions

Trisanu Bhar - Hyperparameter tuning and model architecture

Vikrant Reddy Y - Feature extraction and problem domain research

VENKATA NIKHIL VADDIPATI - Data preparation, analysis and visualization and problem research