Capstone Presentation - Speech Accent Recognition

Problem statement

People from different parts of the world speak in different accent and hence it is difficult for AI enabled devices like Alexa , Google home or even Self Driving cars to understand user commands.

This project aims to recognize and classify speaker accents by AI enabled devices.

Gathering data

Gathering initial data

Initial dataset was obtained from Kaggle https://www.kaggle.com/rtatman/speech-accent-archive.
Dataset had around 2000 audio files and a CSV file.
CSV File contained biographical information about speaker.

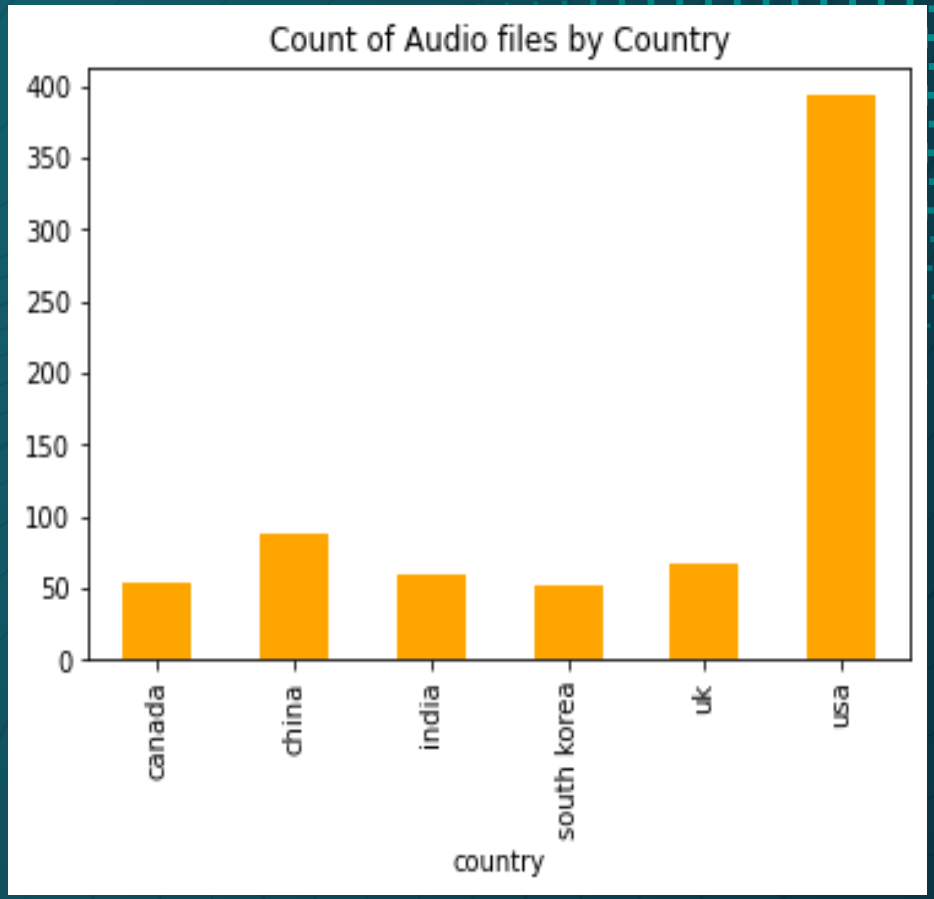
Gathering initial data

Audio files contained below paragraph spoken by all the speakers in duration about 20-30 seconds.

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

Available Data

Most number of audio files available from USA , followed by China , UK , India , Canada and South Korea



Original data story

Original data is hosted @ https://accent.gmu.edu/

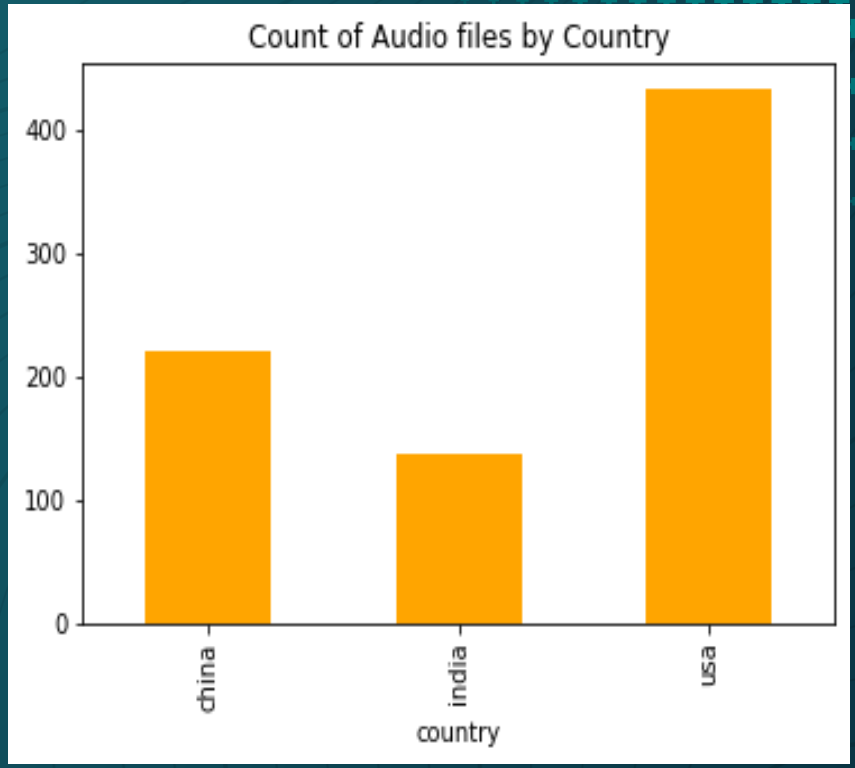
This data is created and maintained by George Mason University's department of English Speech Accent Archive

Gathering more data

Additional data was gathered from Speech Accent Archive

Data was scraped using BeautifulSoup

Data was cleaned and filtered

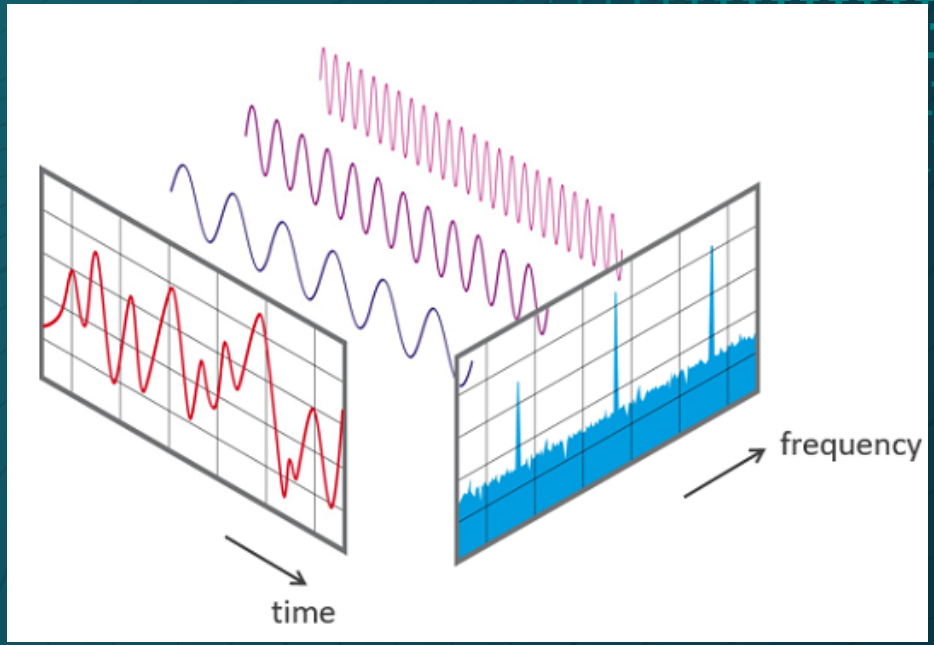


Studying Audio Data

What is Audio Data

Audio is combination of

1. Duration
2. Sampling rate
3. Amplitude
4. Frequency



Source : https://commons.wikimedia.org/wiki/File:FFT-Time-Frequency-View.png

Audio Features

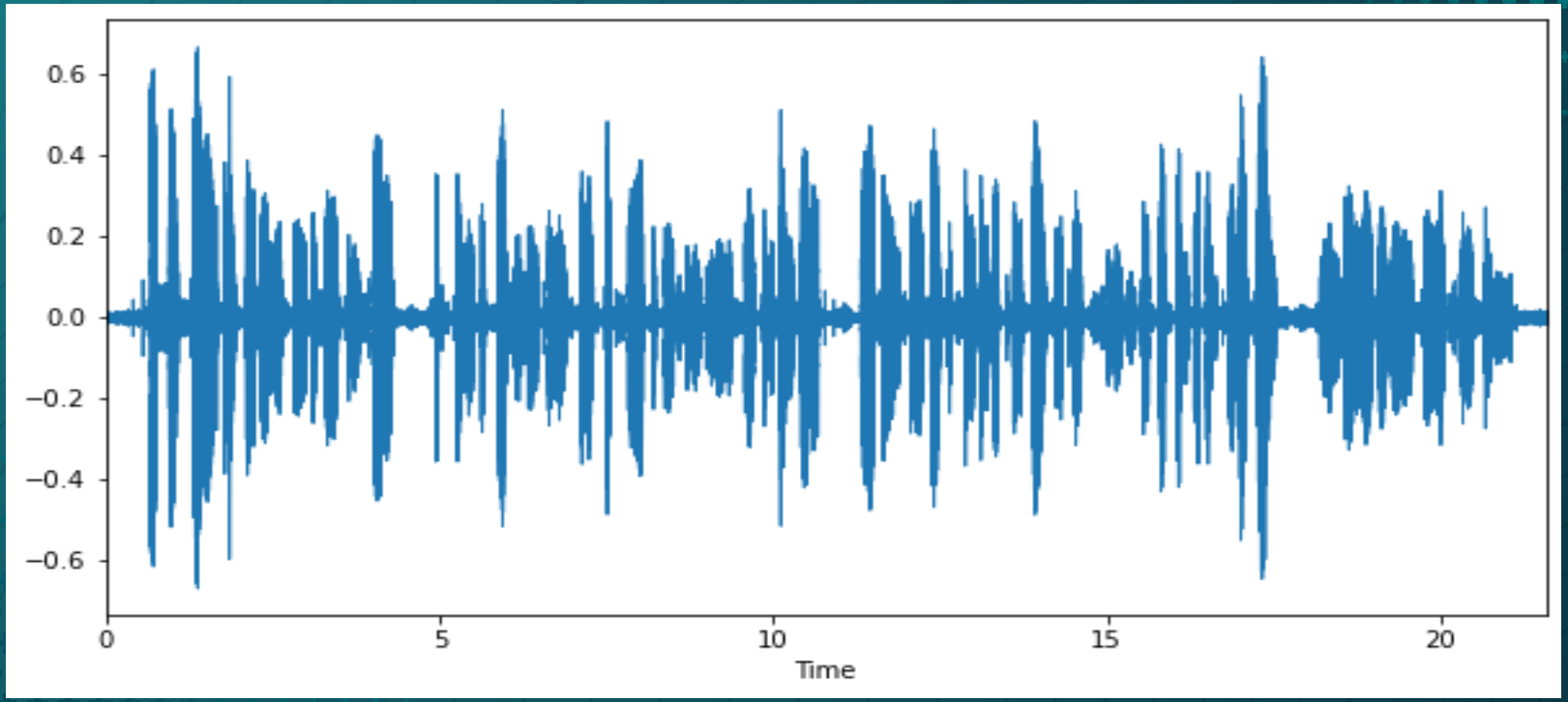
Feature – According to dictionary is: a distinctive attribute or aspect of something.

Some Features you can extract from audio are :

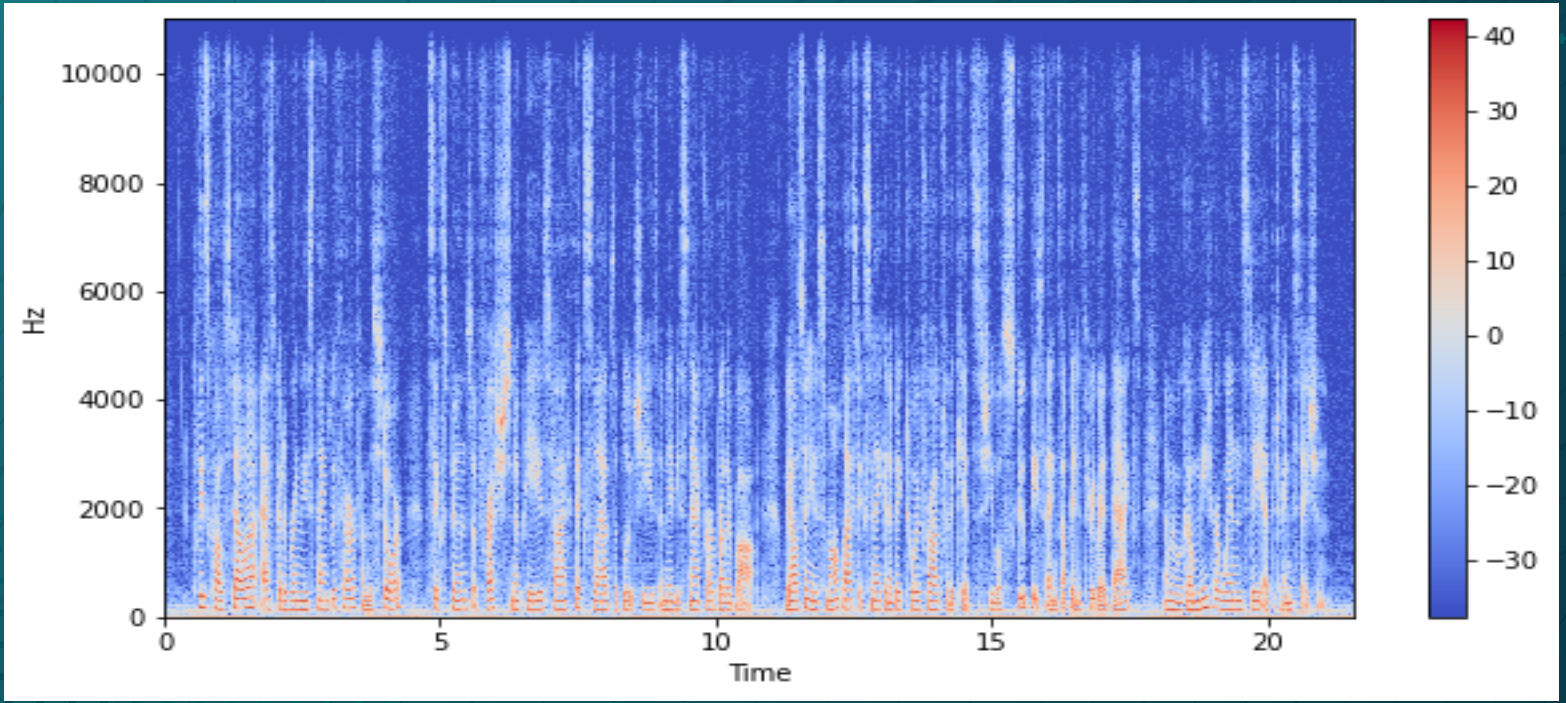
1. Audio Wave
2. MFCC – Mel Frequency Cepstral Coefficients
3. Log Mel-spectogram
4. Chroma

Audio Data EDA

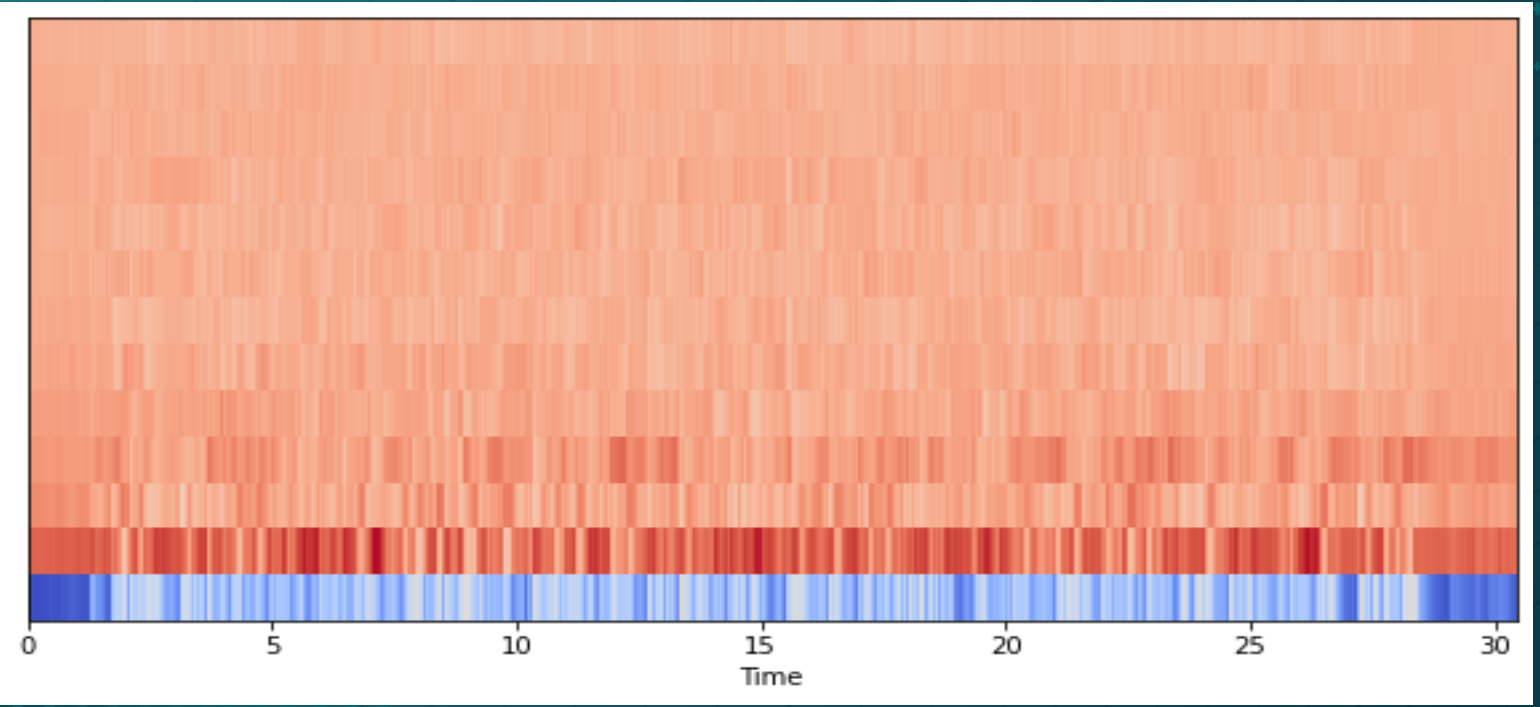
Audio Wave plot from audio file



Spectogram from audio file

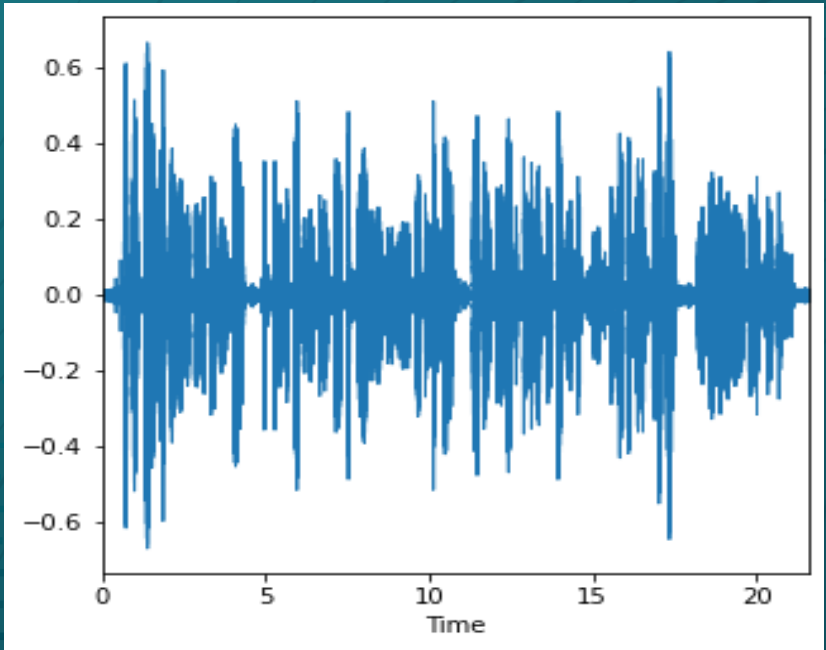


MFCC from audio file

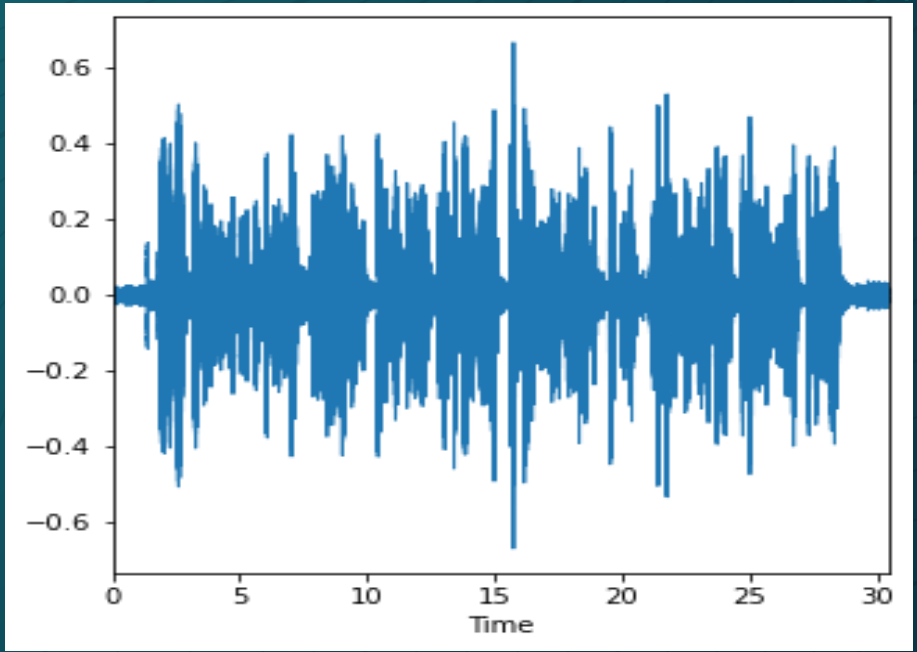


Comparison of audio wave for speakers from 4 different countries

USA

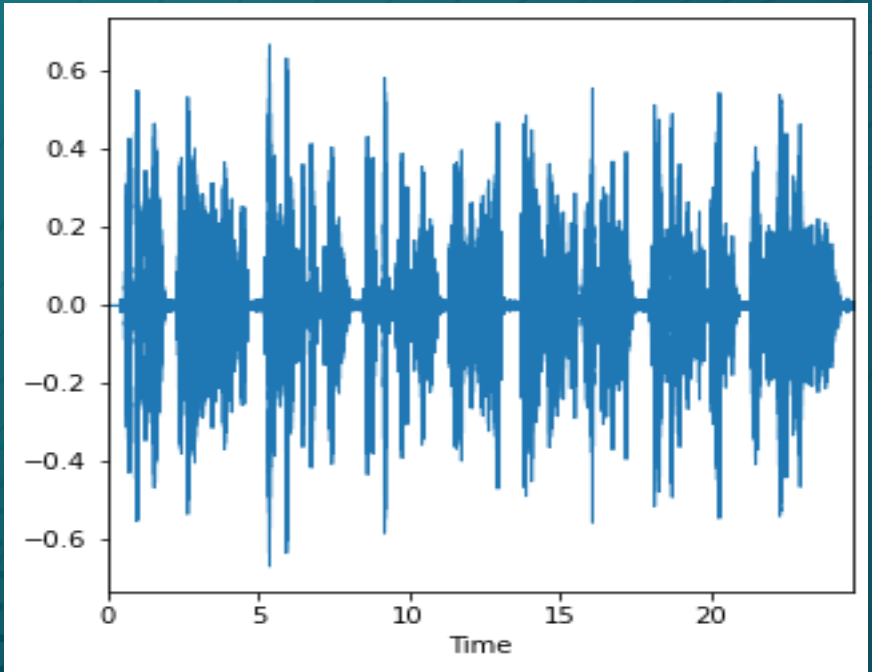


China

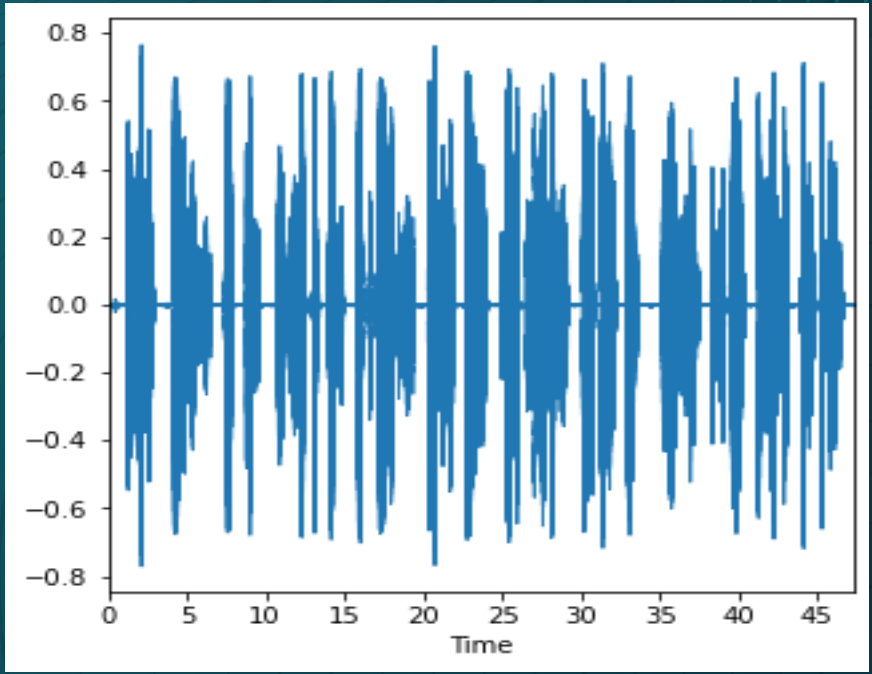


Comparison of audio wave for speakers from 4 different countries

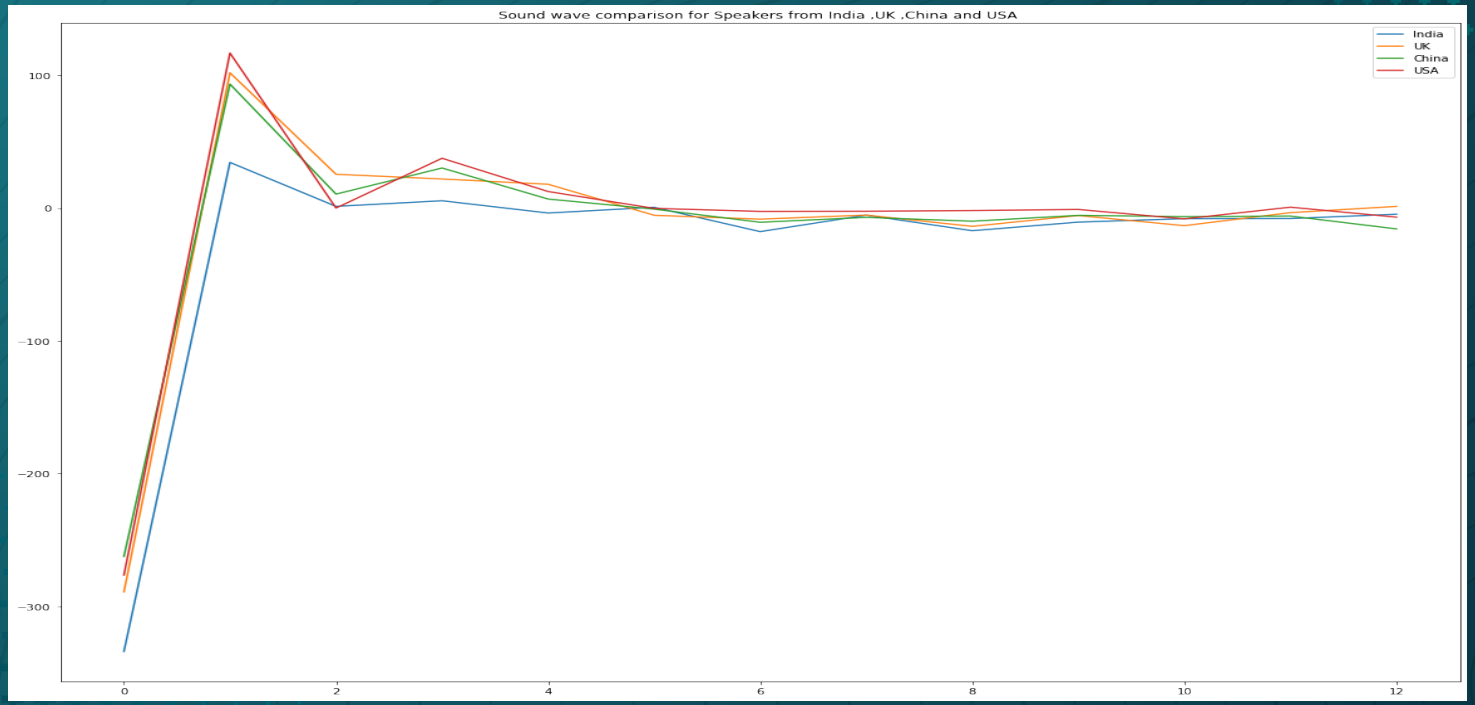
UK



India



Comparison of audio wave for speakers from 4 different countries



Building and testing Models

Prepare for modeling

Below steps were followed to prepare audio data for modeling

- Load audio file using librosa (library for working with audio)
- Extract audio features from audio file using librosa. sampling rate used to extract - 44KH
- Used MFCC feature from each audio file since it is common feature used for speech recognition.
- Convert this to a vectorized array for modeling.

Prepare for modeling

Classes for model	Baseline for each class
1) American	54%
2) Chinies	27%
3) Indian	17%

Different models experimented

Feed Forward Neural Network – to check baseline

Support Vector Machine – LinearSVC

Adaboost with DecisionTreeClassifier

Convolutional Neural Network (CNN) with regularization

CNN with SMOTE

CNN with Undersampling

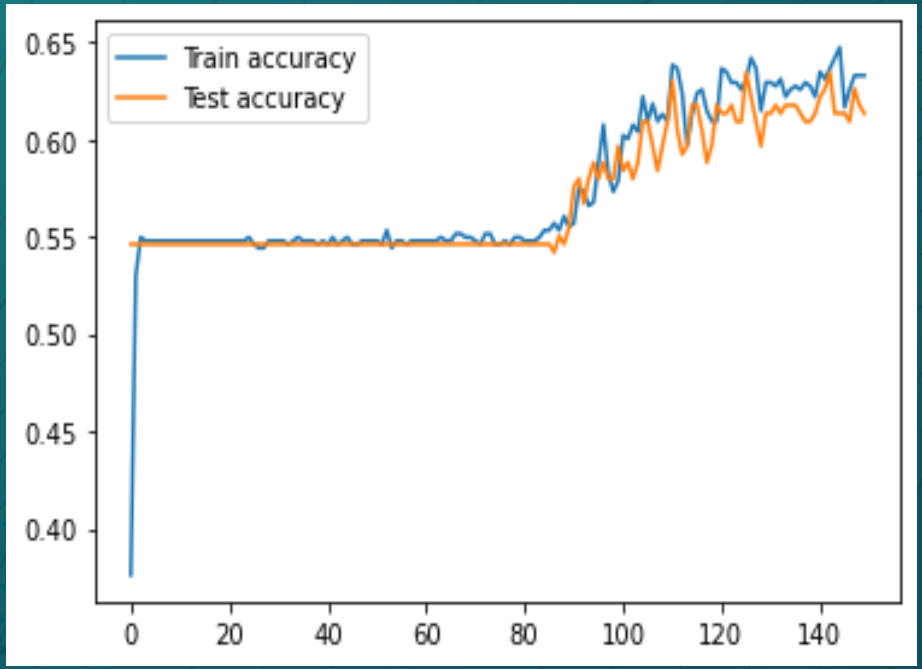
Best performing models

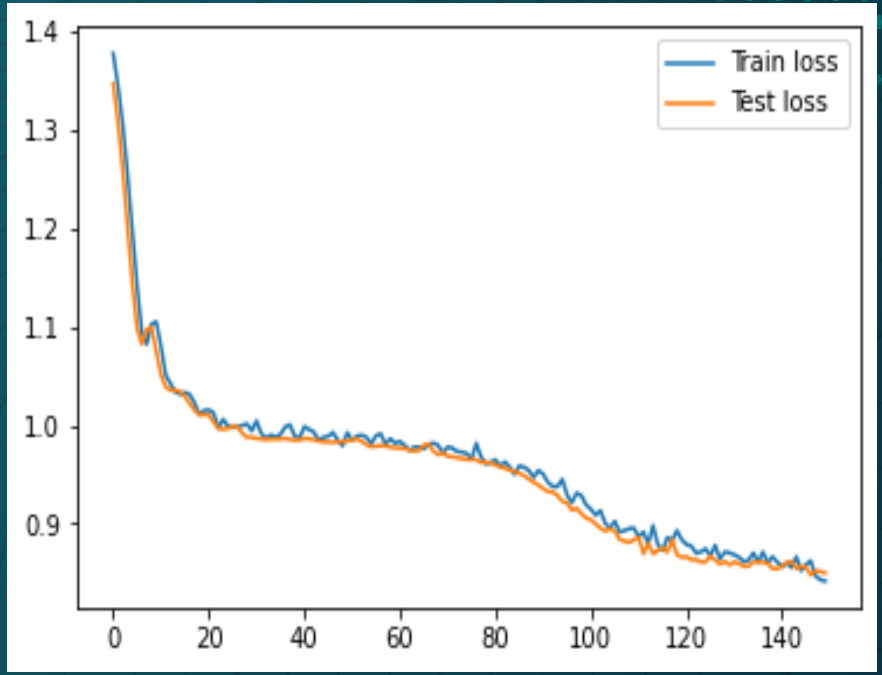
	Train Accuracy	Test Accuracy	Predictability (Y/N)
SVM	68 %	63 %	Y
CNN with SMOTE	62 %	63 %	Y
CNN with Regularization	63 %	61 %	N

Other metrics

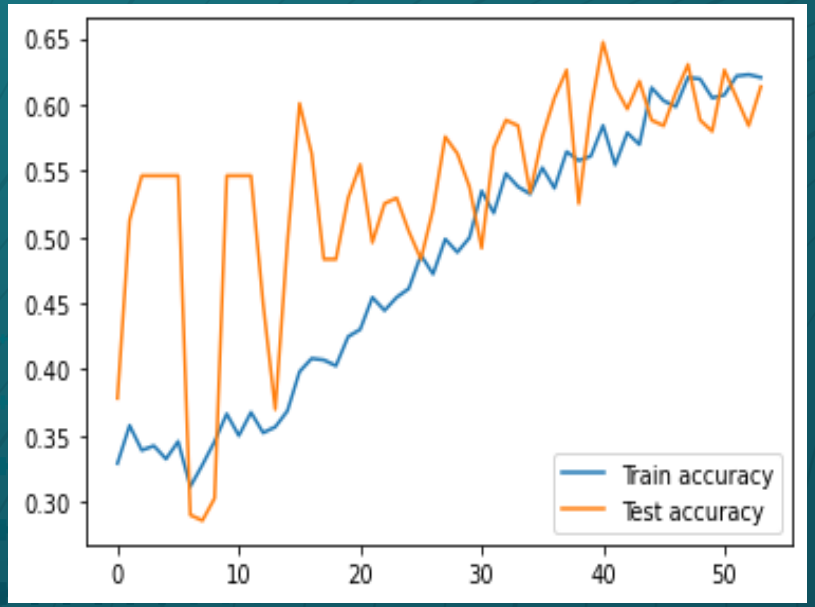
	Precision		Recall		f1 score	
SVM	American	75	American	75	American	75
	Chinese	55	Chinese	67	Chinese	60
	Indian	35	Indian	22	Indian	27
CNN with SMOTE	American	72	American	60	American	65
	Chinese	49	Chinese	54	Chinese	51
	Indian	36	Indian	49	Indian	41
CNN with Regularization	American	65	American	86	American	74
	Chinese	52	Chinese	51	Chinese	52
	Indian	00	Indian	00	Indian	00

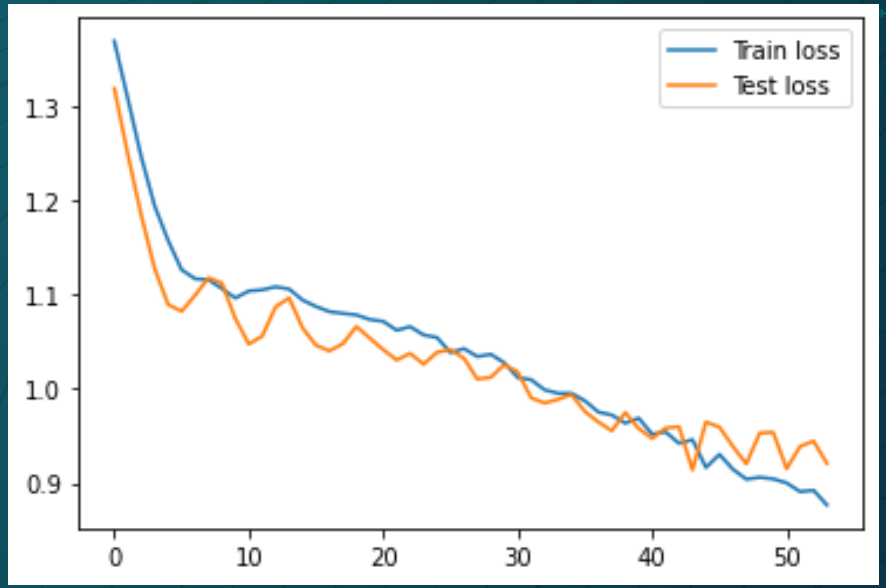
Accuracy and loss - CNN with Regularization





Accuracy and loss CNN with SMOTE

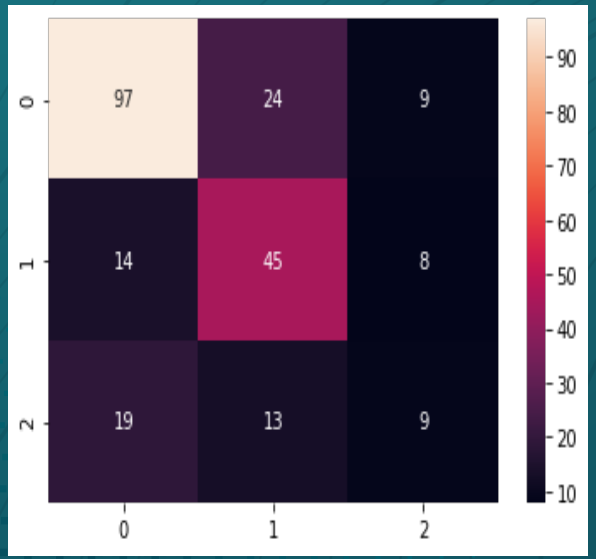




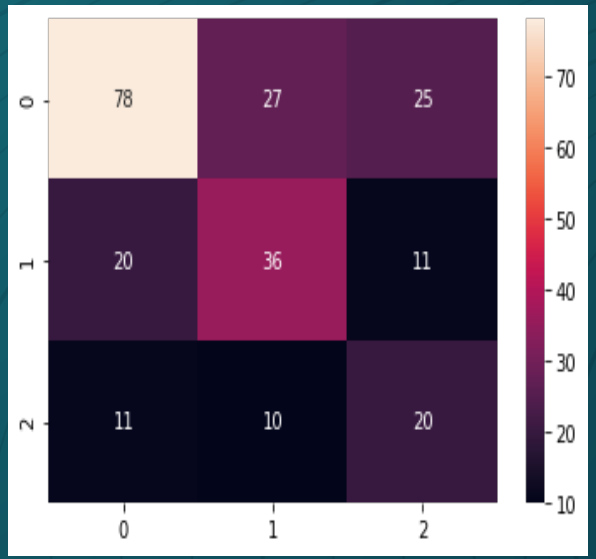
Heatmap for 3 top models

0 - American 1 - Chinese 2 - Indian

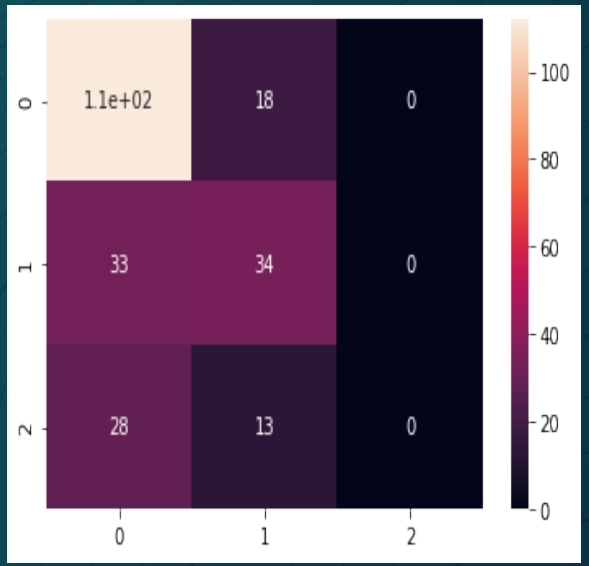
SVM



CNN With SMOTE



CNN with Reg



Conclusion

Summary

SVM is best performing model - good accuracy and optimal precision and recall for all classes.

After using balancing technique SMOTE , precision and recall gets better, and CNN can predict all 3 classes with slight compromise on accuracy.

Data is insufficient and imbalanced.

Due to imbalanced data CNN gives biased results and does not predict 1 class at all.

Next Steps

Gather more data and try to add more classes.

Convert this into a web app for better demonstration.

Try different approach for classification , e.g., try to vectorize with a single word for each user’s pronunciation.

More research with different datasets.

Credits

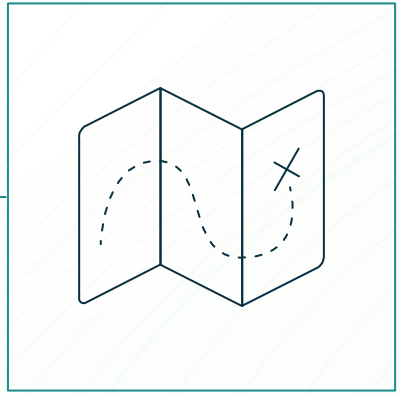
Special thanks to those that made these awesome resources freely available:

https://www.kaggle.com/rtatman/speech-accent-archive

https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-1.html

https://github.com/yatharthgarg/Speech-Accent-Recognition

https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd



Thanks!