MEGATHON 2022
ACCENT DETECTION
Team Name:Le Boys

Problem Statement:
Given an audio clip of a person speaking in English, you are required to identify the native language of the speaker based on their accent when they are speaking in English.

Initial Observations:
After preliminary analysis we realized that the data was skewed, languages like English and Spanish had many data points compared to other languages as shown in fig 1.1.

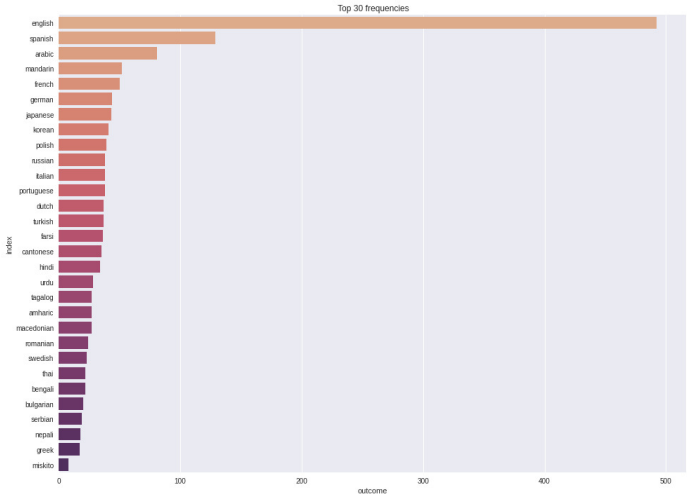

fig 1.1

So to reduce this skewness, we used techniques like data augmentation and oversampling. Data augmentation can be done in 2 ways:

1. Modify the original audio file
2. Augment the Mel-spectrogram

Course of action:
We explored both of these and there were some aspects to be taken care of. For example, we cannot use pitch shift because it might change the accent. We converted the .mp3 files into .wav files for efficient use of libraries.

Approach 1:
We tried fitting the original dataset to a simple dense ANN and achieved an accuracy of around 52%. However the model performed poorly on the test data.

Approach 2:
Tried CNN with oversampling. This gave us a validation accuracy of 71% and training accuracy of 88%. For the minority classes we are attempting to increase the number of audio clips. For this we used librosa.util.frame to slice the data arrays into overlapping frames with the frame length of 10 secs and hop length of 5 secs. Using this method, we observed an increase in the dataset count of 27%. Here we have considered the top 30 classes based on the data count for training the model. Given enough time we can generalize this for all of the classes. 15% of the dataset was used for validation.
Figure 1.2 shows the loss and accuracy curve. Figure 1.3 represents the model architecture.

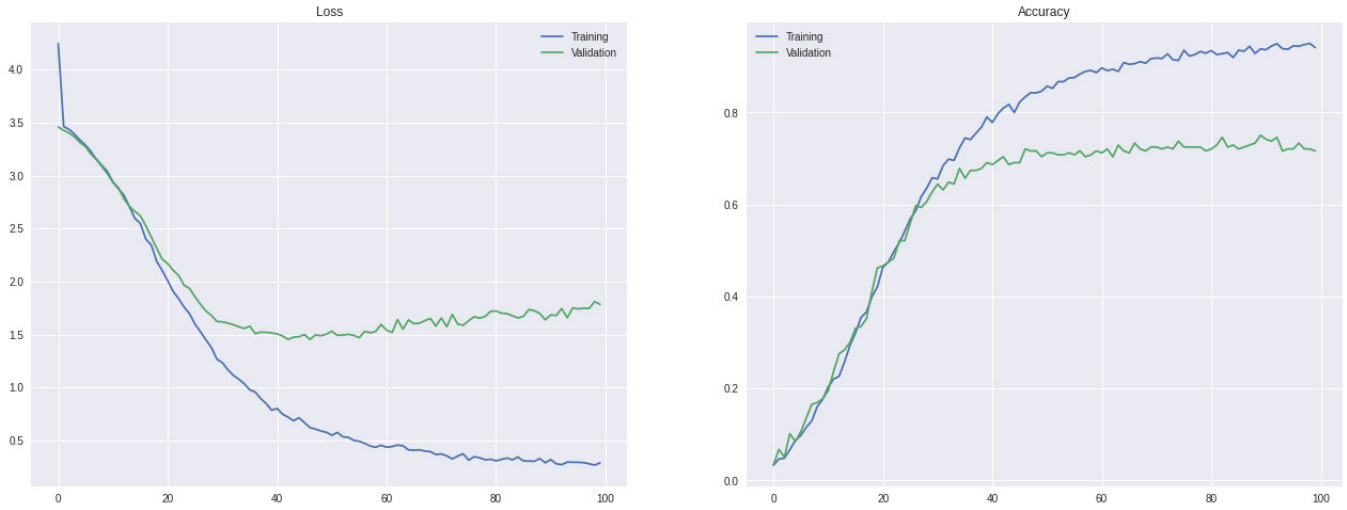

fig 1.2

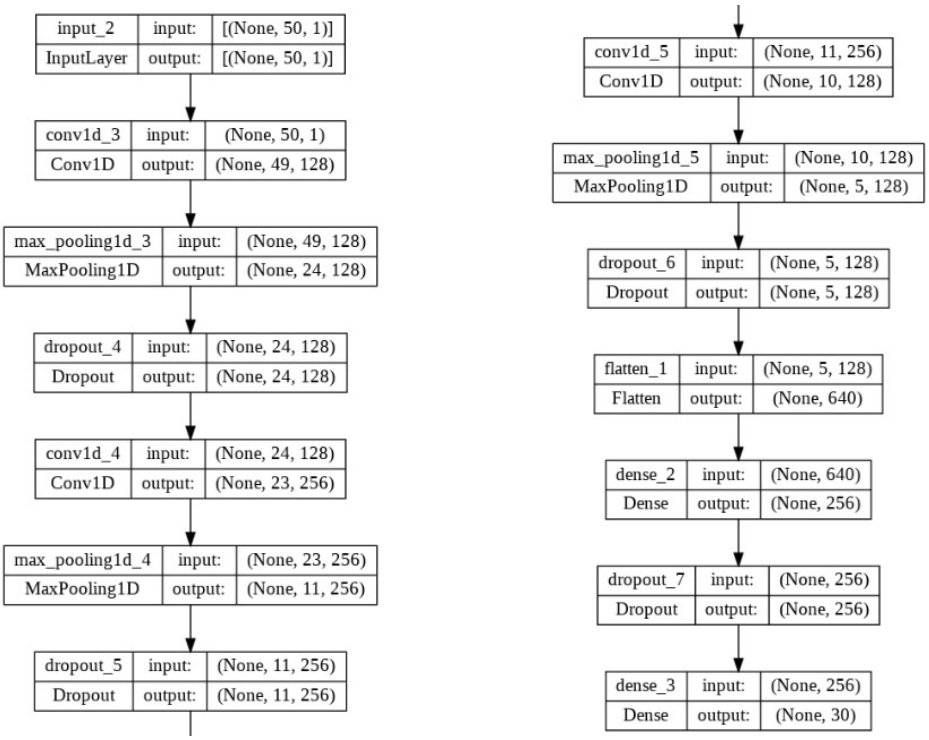

fig 1.3

Approach 3:
Improving the second approach by applying data augmentation. Initially the audio time series data is split into frames during oversampling. These frames are chosen at random and from each frame a new audio file is synthesized. This is done using various techniques like time shifting and by addition of gaussian noise. After this is done, we have obtained a fresh dataset where all of the 30 labels have 100 data points each. Feature extraction is performed on this dataset. Feature extraction is when the time series data of the audio file is converted to a Mel-Spectrogram which acts as an image for our convolutional neural network. Then the CNN is trained and the final model is produced.
Figure 1.4 shows the loss and accuracy curve of this approach.

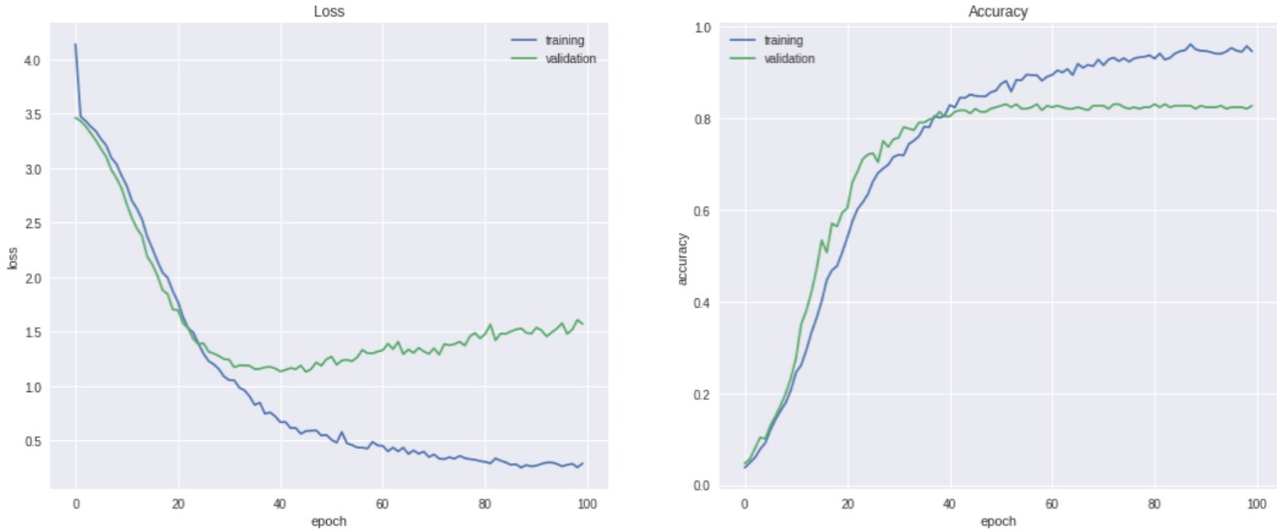

fig 1.4

Future Aspects:

This model can be extended for various classes after collecting more data and adding to the dataset. This will help in increasing the accuracy of the model and make sure of efficient working.