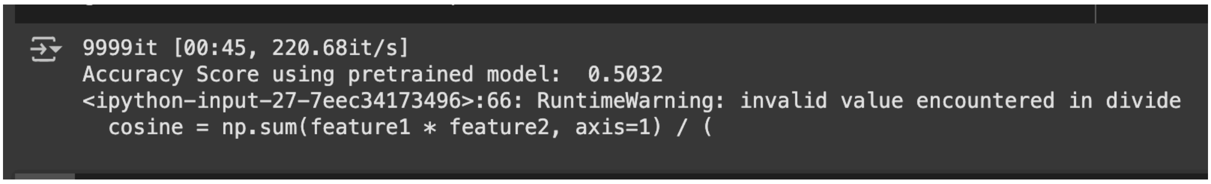
Speech Understanding Programming Assignment - 2

Vikas Kumar Singh (M23AIR545)

Question 1:

For question 1 I have used the pretrained model "facebook/wav2vec2-base-960h" from hugging face to extract features for pairs of audio files. Utilized the `Wav2Vec2FeatureExtractor` to get the features vectors and then calculated the cosine similarity for pairs. Labeled the pair with 1 if cosine similarity was greater than 0.5 otherwise 0. Compared this prediction using a pretrained model with provided test dataset and got the accuracy of 0.503.



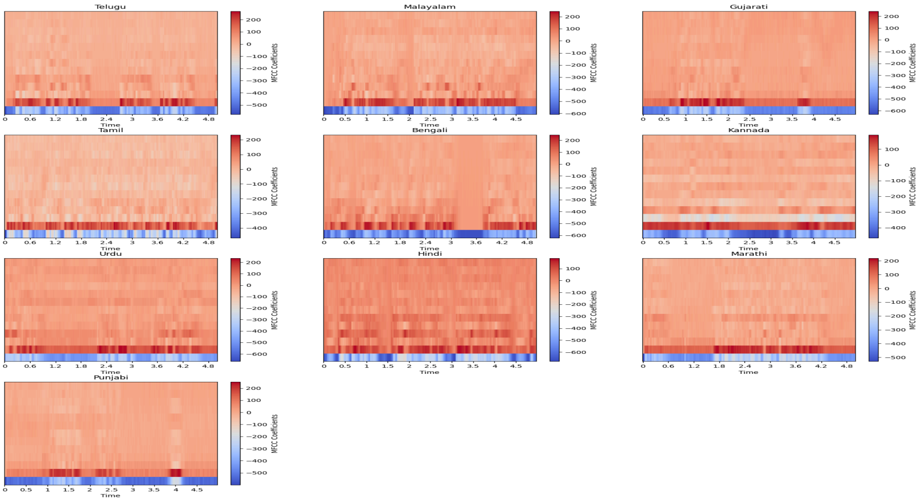
```
9999it [00:45, 220.68it/s]
Accuracy Score using pretrained model:  0.5032
<ipython-input-27-7eec34173496>:66: RuntimeWarning: invalid value encountered in divide
  cosine = np.sum(feature1 * feature2, axis=1) / (
```

Then finetuned the model "facebook/wav2vec2-xls-r-300m" on the provided dataset. Fine Tuning process included dataset preparation also for which I used the `datasets` library, kept 100 id's into a training set and 18 to test. Used LORA implementation using PEFT to finetune the model.

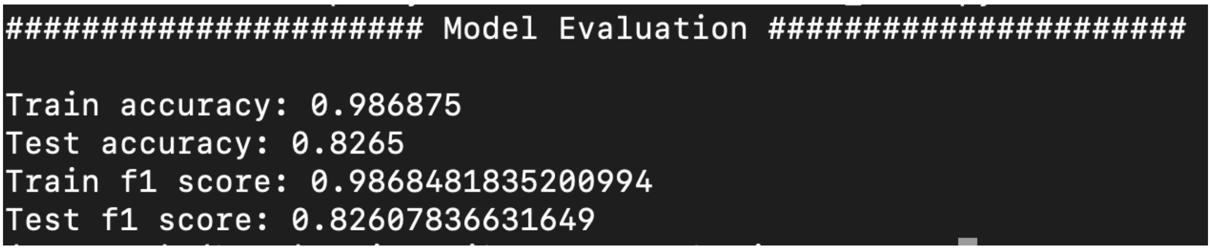
Question 2:

For question 2, first I have calculated the 13 MFCC features for the provided dataset which include audio dataset for 10 indian languages.

Generated the features using these 13 MFCC features which include: mean, standard deviation, min and max. Plotted the Spectrogram of one randomly selected audio file from each language and plotted the spectrogram.



Trained a Random Forest Classifier to classify the language of audio and got training accuracy of **0.98** and test accuracy of **.82**



```
###################### Model Evaluation ######################

Train accuracy: 0.986875
Test accuracy: 0.8265
Train f1 score: 0.9868481835200994
Test f1 score: 0.82607836631649
```