Speech Assignment PA2 - Report

Rohan Frederick
(M23CSA525)

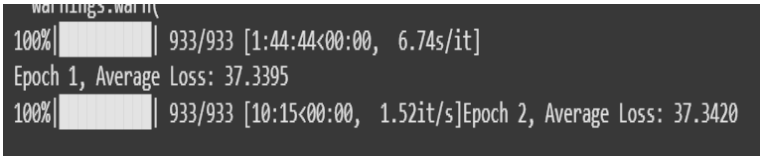
5th April 2025

1 Question 1

1.1 Task I - II

For this question Hugging face wavlm base plus was used. S3PRL library was used to download the model. The evaluation data set shared had approximately 37K files which were preprocessed to resample at 16K sampling rate. Evaluation was done by taking the 'Last Hidden Layer' of the model as the embedding/representation of the audio wav and finding the cosine similarity with the target Audio Wav. All cosine scores were collected and the target label was predicted with a threshold of 0.85

The model was finetuned with LORA on the transformer layers - q proj and v proj. 2 epochs with 933 speech utterances from vox celeb 2 data was used to fine tune the model.



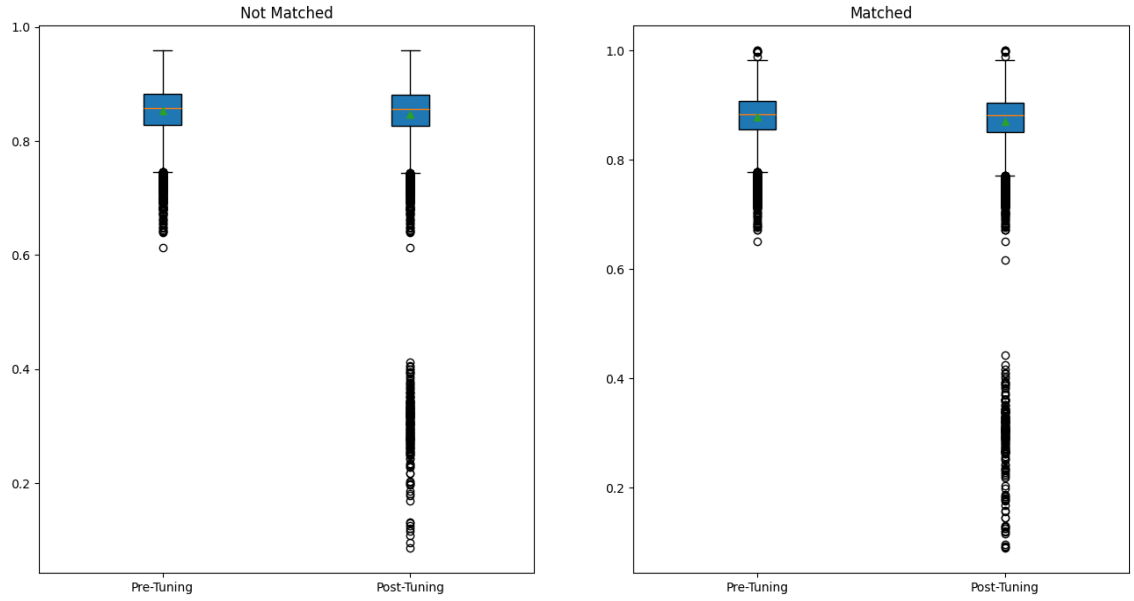
The evaluation matrix was evaluated in the same way, i.e. cosine similarity score between the source and target and labels were predicted with the threshold of 0.8782

Below are the metrics:-

Metric	Pre-trained Model	Fine-tuned Model
EER	36.78%	38.16%
Accuracy	60.6%	61.9%
TAR 1%	7.58	7.15

Table 1: Metrics

The score distribution before and after are as follows:-



1.2 Task III

The speaker samples across first 50 ID were mixed and overlapped using a custom function and Speaker Separation Model from Speech Brain was used to separate the the Separation was evaluated by the following rules:-

- Each pair of constituent speech utterance is matched with each of the separated speech i.e. 2 X 2 = 4 comparisons
- The above comparison is repeated for Cosine similarity and max similarity is considered as the match
- The max SDR value is used to get the match that is used as the ground truth
- compare the match from cosine similarity with the ground truth to get the accuracy

The accuracy of the model with Pre-trained speaker identification model is: 69.5% The accuracy of the model with LORA fine tuned speaker identification model is: 69.5% (Could see changes only in 3rd or 4th decimal places)

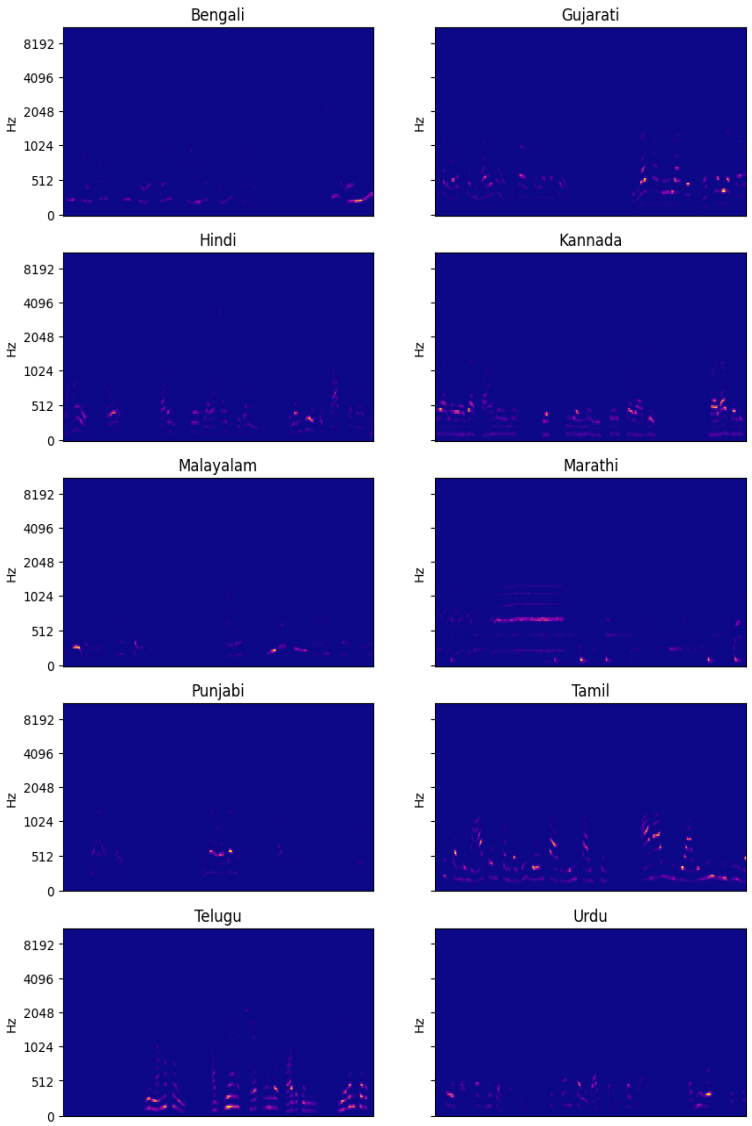
1.3 Task IV

This was achieved using a wrapper class that wraps around both Sepformer model and speaker identification model.

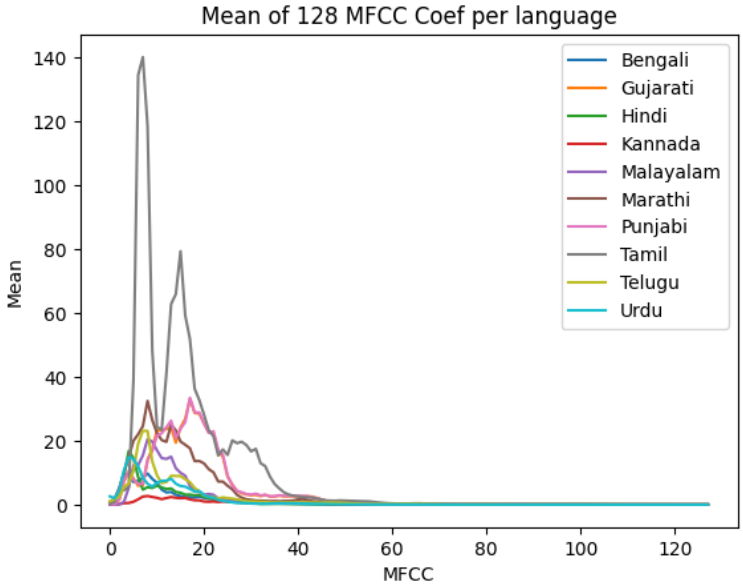
2 Question 2

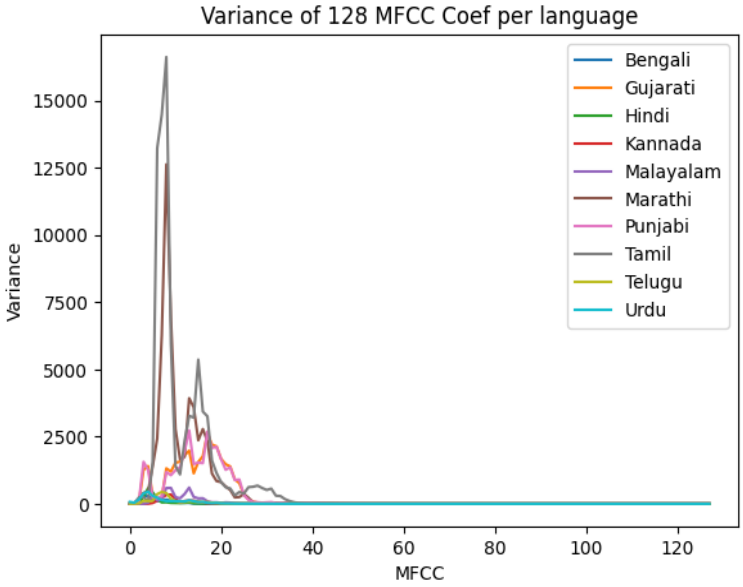
2.1 Task A

The data for 10 languages was downloaded and 128 size MFCC calculated and plotted below:-



The statistical analysis on the MFCC coefficients are as below:-





Comparison on MFCC:-

Language	Observation
Marathi	Has more horizontal lines, indicating sustained harmonics or voicing
Tamil	Shows sharp vertical spikes, possibly reflecting rapid articulation changes
Gujarati	Dense mid-frequency energy, indicating rich consonantal transitions.
Urdu	Balanced energy across low and mid bands, smoother transitions, due to nasal sounds.
Telugu	Bright low-frequency bands with sparse mid-high energy, reflecting tonal transitions.

Table 2: MFCC Comparison

2.2 Task B

The MFCC components across fixed set of windows was flattened and used as a feature to train a Random forest model. The accuracy and other metrics are as follows:-

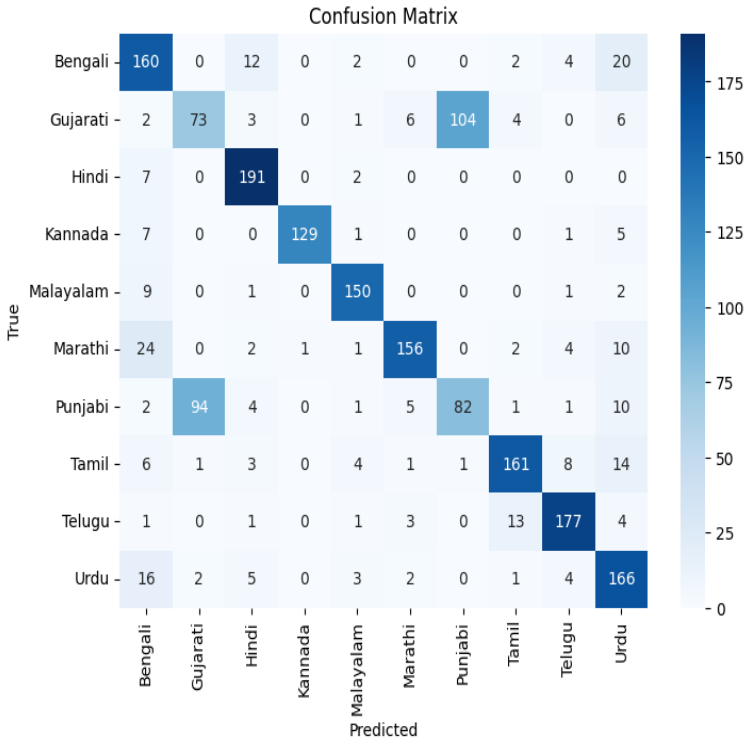
```
Accuracy: 75.93%

Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.80      0.74       200
           1       0.43      0.37      0.40       199
           2       0.86      0.95      0.91       200
           3       0.99      0.90      0.95       143
           4       0.90      0.92      0.91       163
           5       0.90      0.78      0.84       200
           6       0.44      0.41      0.42       200
           7       0.88      0.81      0.84       199
           8       0.89      0.89      0.89       200
           9       0.70      0.83      0.76       199

    accuracy                           0.76      1903
   macro avg       0.77      0.77      0.76      1903
weighted avg       0.76      0.76      0.76      1903
```

The classification output confusion matrix is as follows:-



2.3 Potential Challenges with MFCC

- MFCCs are sensitive to noise, especially low-SNR environments.
- MFCCs capture acoustic properties, which are influenced by the speaker's vocal tract, gender, age, and speaking style. This makes it hard to tell if MFCC differences are due to the language or just the speaker.
- Same language may sound very different across regions (e.g., Tamil spoken in Chennai vs. Sri Lanka).
- MFCCs compress information and mostly emphasize vowel-like energy and lower spectral features.

3 Git Library

- https://github.com/rohanfred/SpeechUnderstanding_PA2.git