PSI Lab: HW

Tharun Dubba

February 17, 2024

This report supplements the Jupyter notebook files and show the results, plots, reasoning. Separate Utils files for individual functions are not used because of Google colab environment.

1. Gender classification:

   (a) Dataset used - LibriSpeech, dev-clean. It has audio data of 20 males and 20 females. For this assignment, all data has been used.

   (b) Since this task is about gender classification, we only have 2 classes. It is derived based on the data given in SPEAKERS.TXT file. Train and test split is 80% and 20% respectively, giving us 16 male speakers and 16 female speakers in the train dataset. Test dataset contains 4 male speakers and 4 female speakers.

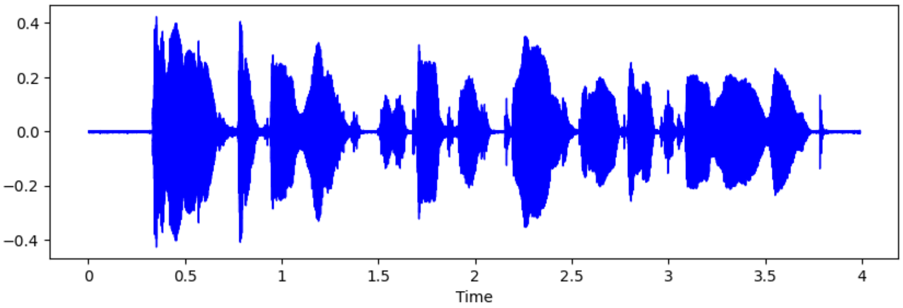


Figure 1: Sample waveform

   (c) I tested on both MFCC = 13 and MFCC = 40 features as well as hop length of 512 and 1024. When 40 coefficients are used, more data is captured giving better results. Hop length of 512 shows 75% overlap between frames.

   (d) Shape of MFCC features for the training dataset is (2212, 40) and the corresponding gender label is of shape (2212, 1). Please note that mean has been taken across each frames in a single audio file.

(e) Usually 1st MFCC coefficient shows us the energy fo the signal. Below image shows the pdf of both male and female training dataset.

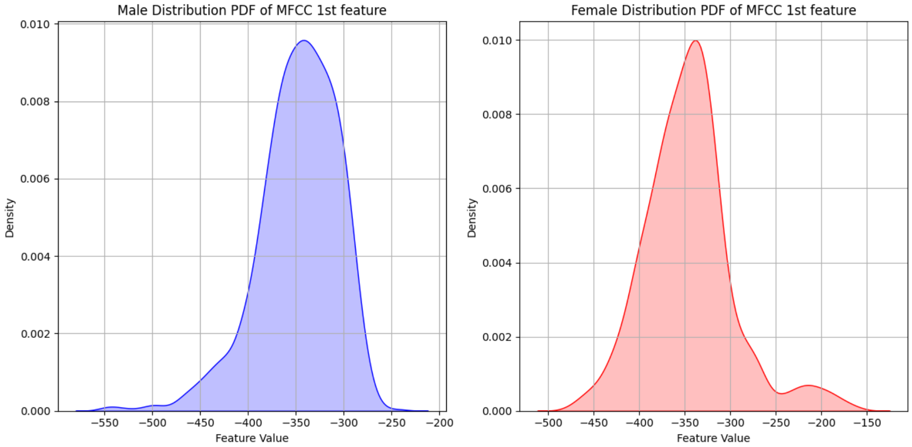


Figure 2: 1st MFCC coefficient visualization of males and females

(f) Further distribution plots across different coefficients are shown in the notebook file.

(g) For outlier detection, boxplots are generally used. Assuming most of the data falls in between 2nd and 3rd quartile, 1st and 4th quartile samples are considered outliers and are discarded to not skew the model. As we notice in the graph, most of data has values in between -300 and -400. However, it is important to find out the most important component and remove the outliers.

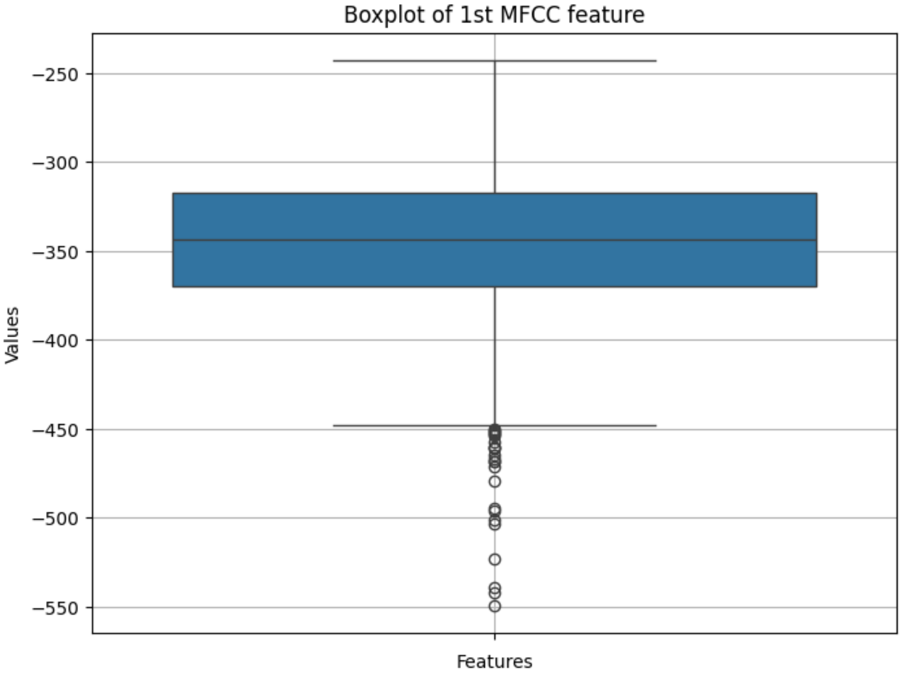


Figure 3: Boxplot of 1st MFCC coefficient in males

(h) Scaling plays a huge role in the result of the model. Some of the features might be very large and model will try to adjust it's weights to adapt these features, causing a highly skewed model and lower accuracy. I used standard scaler method, where we subtract the mean and divide by it's variance.

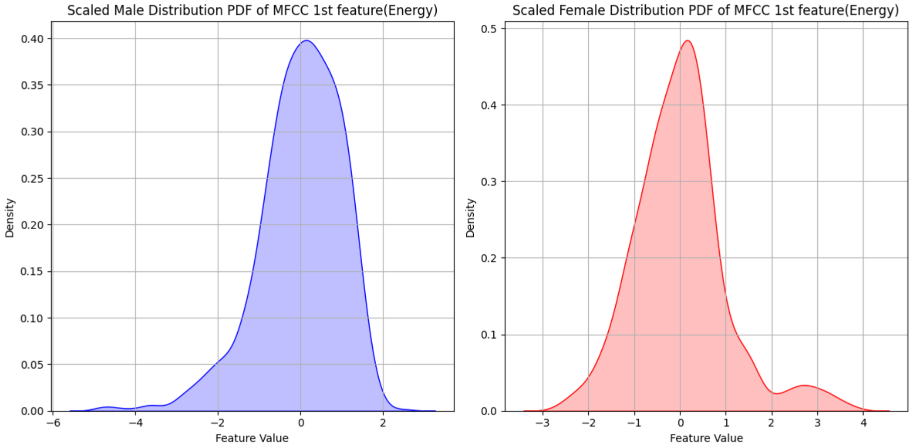


Figure 4: 1st MFCC coefficient visualization of males and females[scaled]

(i) **Naive Bayes classifier** For the 1st classifier, I used Naive bayes classifier. It is a probabilistic ML model which predicts class based on given features. In this case, males are classified perfectly whereas females have 32 mis-predictions. It might be because the data is skewed towards males as seen in the training accuracy. I have used random shuffle to fetch 20% test data and in this particular shuffle, majority of them are females, meaning more training samples for males data.

Table 1: Classification report on test data

Class	Precision	Recall	F1-score
Class-0	1.00	0.85	0.92
Class-1	0.90	1.00	0.95
Accuracy	-	-	0.93

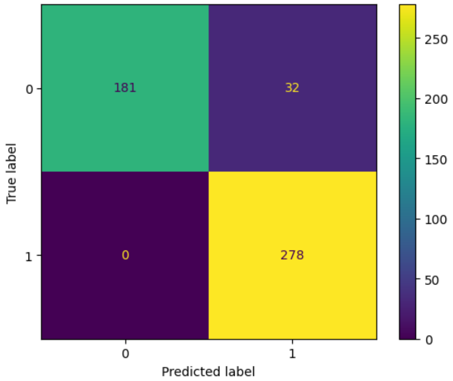
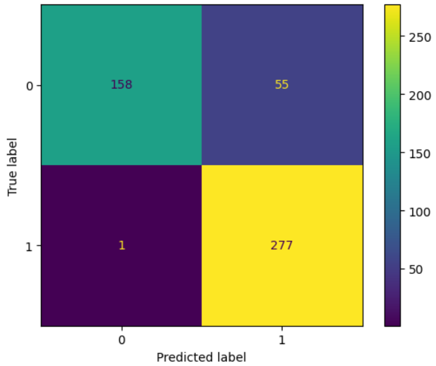


Figure 5: Confusion Matrix

2. **SVM classifier** For the 2nd classifier, I used SVM. It is a non linear function as I used polynomial kernel. Accuracy follows similar pattern to that of Naive Bayes classifier.

Table 2: Classification report on test data

Class	Precision	Recall	F1-score
Class-0	0.99	0.74	0.85
Class-1	0.83	1.00	0.91
Accuracy	-	-	0.89



3. **MLP classifier** For the 3rd classifier, I used MLP. It is a non linear Neural network with 2 small hidden layers. Accuracy follows similar pattern to that of other classifiers

Table 3: Classification report on test data

Class	Precision	Recall	F1-score
Class-0	0.98	0.80	0.88
Class-1	0.86	1.99	0.92
Accuracy	-	-	0.91

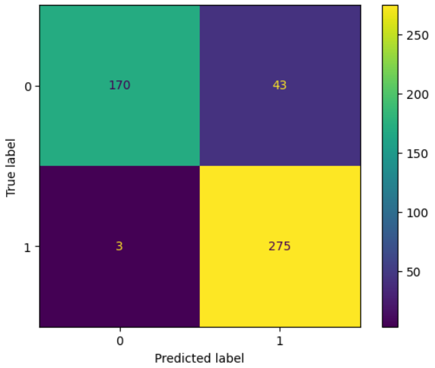
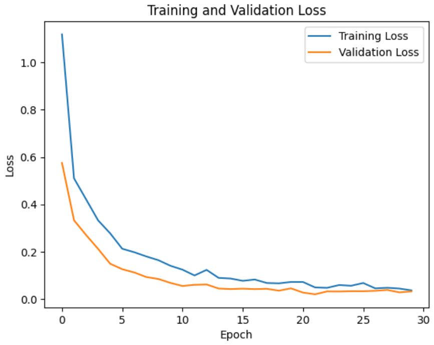


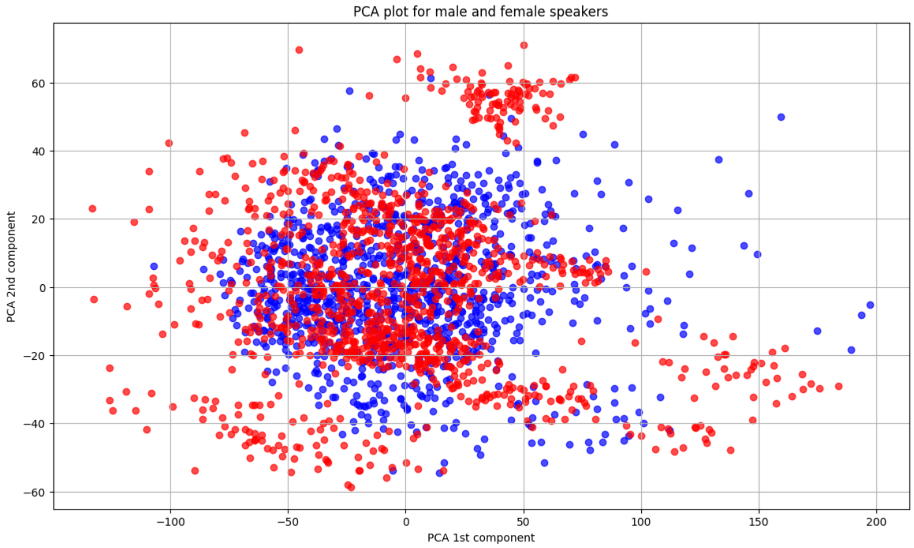
Figure 6: Confusion Matrix

4. **CNN based classifier** Accuracy is slightly higher as compared to other classifiers, it can be attributed to the complex representations of the MFCC features and it's ***temporal relations***. As shown in the image, training loss and validation loss converged around 30 epochs. Since there are 2 classes, I used binary crossentropy loss function.
   Accuracy = 0.93

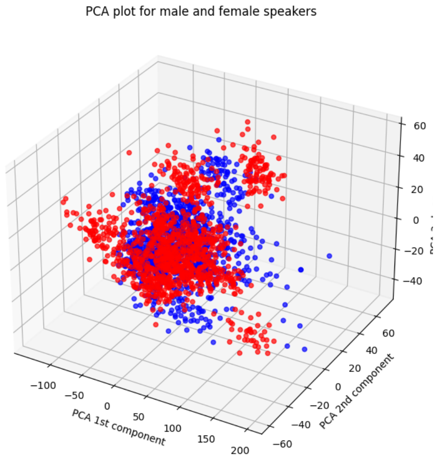


5. **Dimensionality reduction** PCA for 2 components is derived and scattered on a 2D plot. No possible separation with just 2 components as points are spread equally. Red shows female speakers and Blue shows

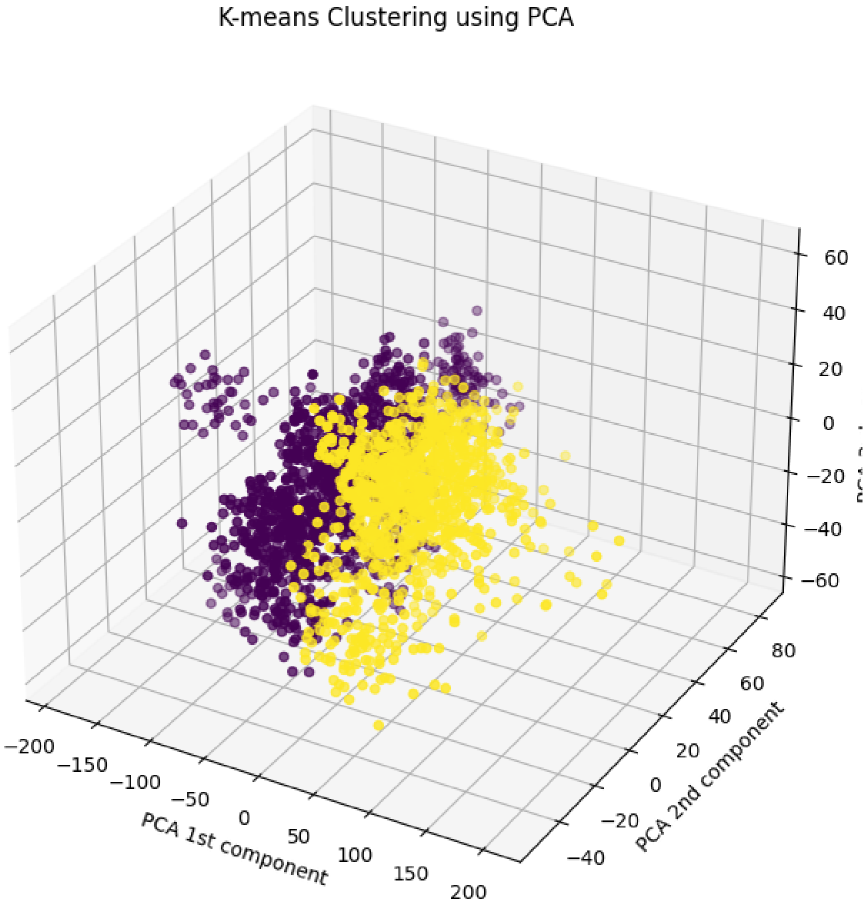
male speakers.



6. **Dimensionality reduction** PCA for 3 components is derived and scattered on a 3D plot. Separation is not clearly seen but can be classified when k means clustering is performed



7. **K means clustering with PCA=3** As seen in the figure, both genders are separated using 3 principal components. It shows that just 3 main components is needed for gender separation task. Although accuracy needs to be measured.



8. Part 2: Speaker classification:

    (a) Speaker embeddings are extracted from wav2vec2 base model with no finetuning. Method described in the ReadMe is somewhat different from the model checkpoint. I followed below link to get the extractions.
    https://github.com/facebookresearch/fairseq/issues/3134

    (b) **Data processing:** Features are extracted based on the 2 second chunks in each audio file and appended together. One key difference as compared to gender classification is making sure of uniform distribution of speakers in the training and testing dataset.

    (c) **Naive Bayes classifier** Accuracy is very low as compared to gender classification. It is because of high complexity in the data. Each vector is of 768 dimensions, whereas with the short amount of speaker data, it is difficult to estimate the probability distribution.
    Accuracy = 0.24
    Confusion matrix is not shown as there are many speakers.

    (d) **SVM classifier** Accuracy is significantly higher compared to Naive bayes classifier. SVM kernel is polynomial making it non-linear and able to represent complex data.
    Accuracy = 0.88

    (e) **MLP classifier** Accuracy is similar to that of SVM classifier as it is a non-linear model.
    Accuracy = 0.88

    (f) **CNN classifier** Since the CNN model only has few layers and low data, model is unable to learn. Hence, the accuracy is low.
    Accuracy $\leq$ 0.5
    However, model is trained only on 40 epochs because of computational issues. As seen in the graph next page, accuracy can be improved with more epochs and deeper layers.

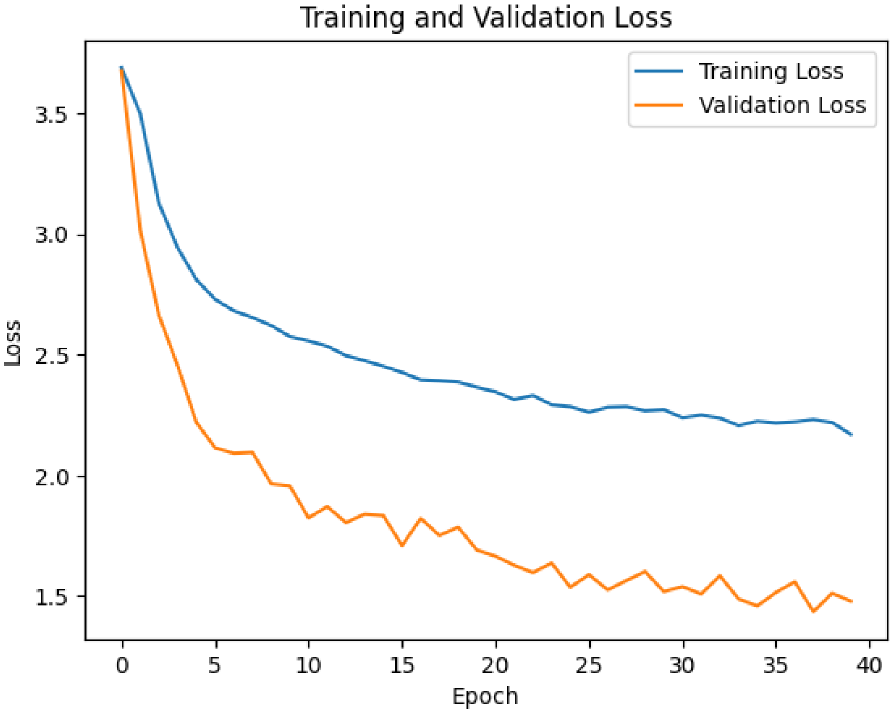


Figure 7: Training loss and Validation loss vs Epoch