HMI Coding Challenge Solution

1. Description of the method

1.1 Split of data into train and test folder.

The actor speech from actors 1 to 20 was put into the train folder. The actor speech from actors 21 to 24 was put into the test folder.

1.2 Loading the audio and feature extraction

The MFCC features are extracted from the audio, by traversing the actor speech dataset. Librosa library load and mfcc functions are used to extract the features. The features are concatenated and returned. The features are extracted for both train and test files. The train folder contains the speech for the first 20 actors and the test folder contains the speech for the last 4 actors.

1.3 Train and test data split

The features extracted from the train folder are further split into X_train and X_val, along with Y_train and Y_val. The train_test_split function from sklearn library is used to achieve X_train, X_val, Y_train, and Y_val. The validation to train ratio was set to be 0.1.

1.4 CNN model

The CNN used in the model has two layers and one fully connected layer. For the training of the CNN, Adam optimizer and learning rate of 0.001 was chosen as the parameters.

1.5 Random Forest Classifer

The features were passed through the CNN, then concatenated and further added to the classifier. The last layer of the CNN network was removed. The mfcc features were passed through the CNN network. The outputs were then passed through the random forest classifier. The random Forest classifier outputs greater accuracy than a softmax layer.

2. Results

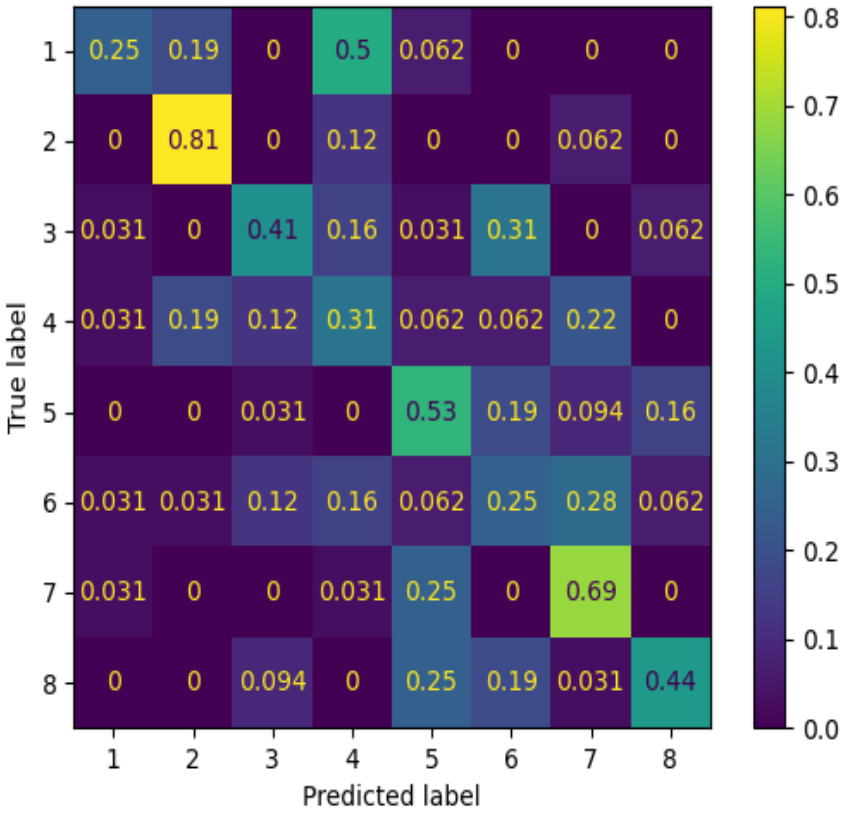
The model classification accuracy is 0.47

recall score:[0.54545455, 0.67567568, 0.39393939, 0.29411765, 0.4, 0.47058824 0.4516129 0.40540541]

f1 score:[0.44444444, 0.72463768, 0.4, 0.3030303, 0.44444444, 0.32653061, 0.44444444, 0.43478261]

precision score:[0.375, 0.78125, 0.40625, 0.3125, 0.5, 0.25, 0.4375, 0.46875]

Confusion Matrix:



emotion_labels :
'01':'neutral',
'02':'calm',
'03':'happy',
'04':'sad',
'05':'angry',
'06':'fearful',
'07':'disgust',
'08':'surprised'

Calm, angry, and disgust classes have higher accuracy than the average.
Neutral, sad, and fearful classes have lower accuracy than the average.
Happy and surprised classes have accuracy similar to average.

3. Future Scope and further improvements to the model

3.1 Adding more features

One of the computationally inexpensive ways to increase the accuracy is to increase the number of features. For instance, Gammatone Cepstral Coefficients is one of the features that could be added to the classifier along with MFCC. The Gammatone cepstral coefficients (GTCCs) are a biologically-motivated type of feature. The features are extracted using Gammatone filters with equal rectangular bandwidth bands. Research shows that GTCC outperforms MFCC in non-speech audio classification.

3.2 Recurrent Neural Network

A Recurrent Neural Network is a class of networks that contains simple neurons which pass information to each other, such that the information persists. RNN are useful in tasks in which the data has an imminent effect on each other. For instance, words of a sentence makes sense when seen in relation to each other, as standalone may not contain as much meaning.The network contains multiple copies of the same network, with one layer of tanh function.

3.3 Using a more robust CNN

Increasing the number of CNN layers can result in improved accuracy. Although this method is computationally expensive, it is the most straightforward way of getting better results. By increasing the number of hidden, we can increase the number of features. However, this method can result in overfitting, which stops the model from being generalized, and can cause a drop in the model's performance.