**Jio Internship experiment 2**

**Problem Statement:**

Create, train and test a classifier that will accept speaker audio files and then classify them into specific age groups and also classify a gender for them.

**Dataset:**

For the dataset, I use the audio files given in this github repo (<https://github.com/iiscleap/NISP-Dataset/tree/master>). This dataset had a lot of audio files to choose from, also the languages that the speakers spoke were probably most aligned with the speakers that the customer base of Jio spoke as well. This dataset consists of information of 345 speakers of which 219 are male and 126 are female. The audio files contain sentences taken out from newspaper. Each speaker has contributed about 4-5 minutes of data that includes recordings in both the English and mother tongue.

**Solution:**

The solution consists of 2 parts: Data Preprocessing and the Convoluted Neural Network model.

**Data Preprocessing:**

For the data pre-processing, I used the librosa library to remove every silent aspect of the audio files and then split the audio files into multiple 30ms frames form which various features were extracted. I experimented with various features and there are a lot of features to work with but there are some usual suspects: mfcc, delta\_mfcc, delta2\_mfcc, pitch, zcr , rms, spectral\_centroid, spectral\_bandwidth, mfcc\_mean, mfcc\_std, mfcc\_skew, mfcc\_kurt. A substantial number of these features are arrays, so even though the number of features may seem to be only 12, in the dataset we end up with 67 feature columns. Again, it is important to reiterate that there are more features that we can obtain.This resulted in the creation of over 16 million training frames and thus over 16 million training records that each had 67 features and 5 labels ( 1 for gender and 4 for age group). From the test audio files we generated over 3 million test records with the same number of labels and features. We did not overload all the data into single csv files as the machine would not be able to handle it. Thus each csv file contained only 2 million records each and resulted in the creation of multiple csv files.

**Convoluted Neural Network**

I read all the test and training data and trained and tested multiple CNN models , one for each label and then hyperparameter tuned them to get the best possible accuracy. We first normalized the training data and then began training it with the model.

The model that we are using contains 3 convolution layers each of kernel size 3 and relu activation and the number of filters decreasing from 256 to 128 to 64. Each layer also consists of batch normalization and the first 2 layers having MaxPooling as well and the final convoluted layer having GlobalAveragePooling1D. This is then followed by 2 dense layers also having relu activation and filters decreasing from 128 to 64. These layers also have a dropout layer of weight 0.5. Finally, we have the output layer which is just 1 layer with a sigmoid activation function. The model was compiling with the Adam optimizer of learning rate = 0.001 and the loss was binary\_Crossentropy and the metric used was accuracy. We also used early\_stopping with a val\_loss monitor, patience=5, restore\_best\_weights,

Finally the model was trained on 30 epochs with each batch size of 1024 and validation split of 0.2 and evaluated on the test data. This model is used for all the labels.

Though not all of the models have been completely tested. I have great expectations that we should be able to get atleast a basic accuracy of 70% minimum.

**Results:**

Currently the models are giving a basic test accuracy of 67%. The loss gradient is exploding after each epoch and the testing accuracy is around 60%

**Problems:**

None of the audio file datasets found online have an equal distribution of male and female speakers as well as equal distribution of age groups. Also there are multiple features that can be extracted from the audio files but were not possible due to hardware difficulties and time constraints. Model training and modifications were also constrained by the previous factors. Further research and development will be required to get better results. Also as there is to much amount of data running and training this model on services such as Google colab or even on local machines becomes impossible after you include a certain number of features.

**Improvements:**

Try to find better models. Extract more features. Try to circumvent the hardware and software issues.