CSE616 Project Documentation
Neural Network Gender Classifier Using Audio

Othman Mohamed Othman Ahmed

2002395@eng.asu.edu.eg

Ain Shams University
Faculty of Engineering
Computer and System Department

1. Overview

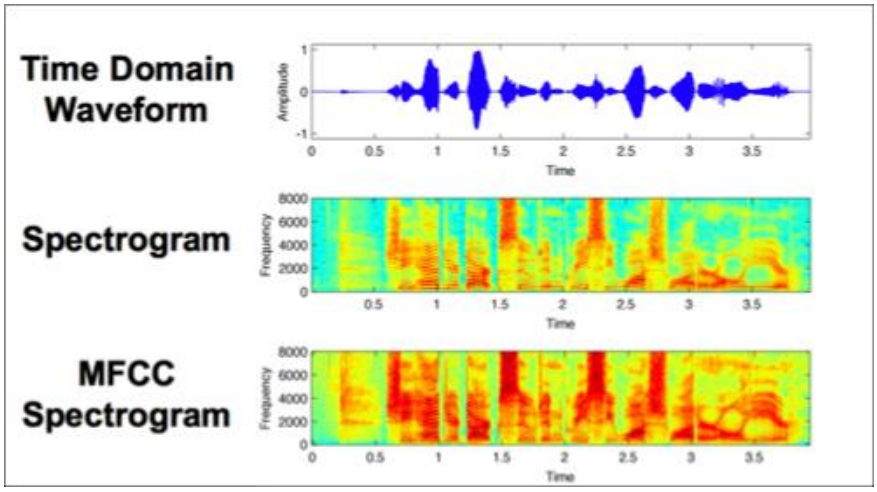
Due to the breakthrough of deep learning, Gender Recognition from audio data has become a critical problem in many applications. Such Applications Include Speakers Separation, Speech Recognition and Speech Translation. Many approaches have been explored to solve this problem using different architectures and techniques. In this project, we explore the Usage of Convolutional Neural Networks as well as LSTM Recurrent Neural Networks to build a gender classifier. Mel-Frequency Cepstral Coefficients (MFCCs) are chosen as our features, as they capture important information from the sequential audio data both in time and frequency domain. Then we build our classifier based on these features to classify any audio file according to speaker gender.

2. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are one of the most common features used in processing audio data. It is observed that extracting features from the audio signal and using it as input to the base model will produce much better performance than directly considering raw audio signal as input.

MFCCs are commonly derived as follow:

1. Calculate Fast Fourier Transform of the signal
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows or alternatively, cosine overlapping windows.
3. Take the logs of the powers of each of the mel frequencies
4. Take the Discrete Cosine Transformation of the list of mel log powers.
5. The MFCCs are the amplitude of the resulting Spectrum



3. Datasets

For our application, we needed audio files with assigned labels for the gender of the speaker. 2 of the open-source popular datasets were chosen: LibriSpeech and TIMIT. Subset of the 2 datasets was chosen, of total 8810 files.

The audio files were divided into 80% Training, 10% Validation and 10% Test.

3.1. LibriSpeech

LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned. For our application, we used the clean 100 hour subset of LibriSpeech. It is divided by speaker into 251 Speakers, 125 of which are females and 126 are males. We can extract the gender of the speaker for each audio file from the meta data. For each of the speakers, we choose 10 random files from different recordings.

	Number of Speakers	Number of files per Speaker	Total number of files
Male Speakers	126	10	1260
Female Speakers	125	10	1250

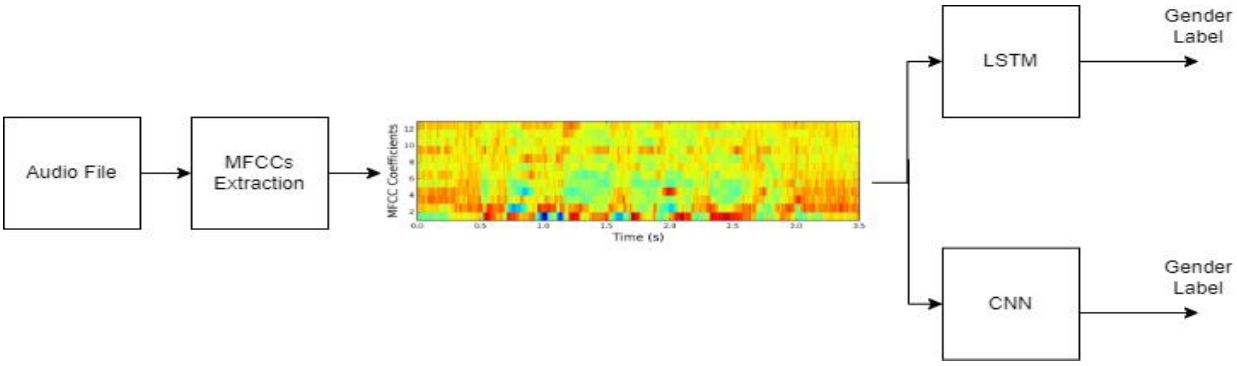
3.2. TIMIT

The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States.

Table 1: Dialect distribution of speakers

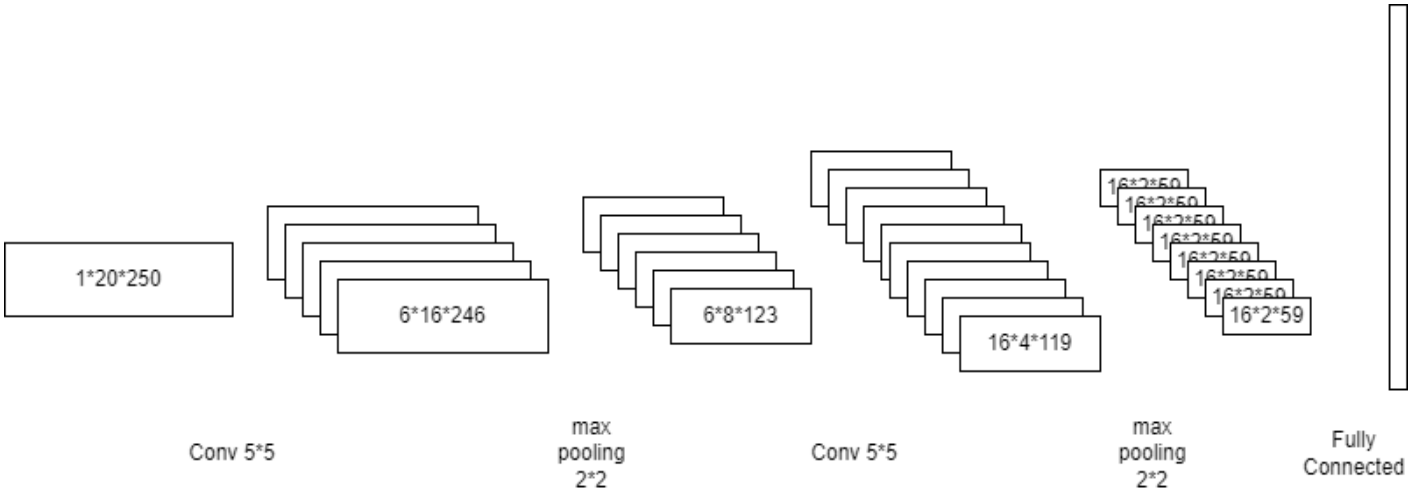
Dialect Region(dr)	#Male	#Female	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
8	438 (70%)	192 (30%)	630 (100%)

4. Model Architecture



For the audio file, the file is read and then the MFCCs are extracted, then the MFCCs are input to either CNN or LSTM network to produce the gender label. For the mel coefficients, 20 mels are chosen which is a common to choice in speech applications.

4.1. CNN



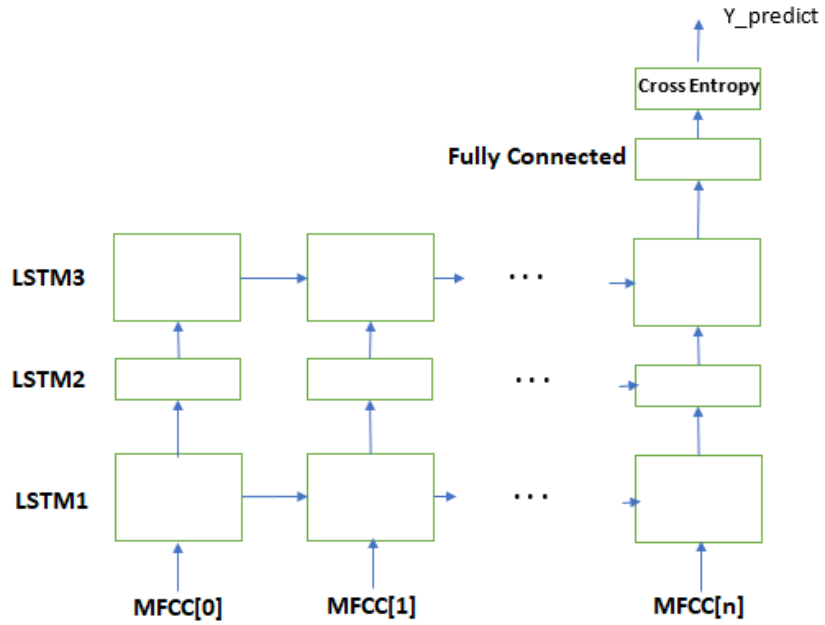
For the CNN model, We stack 2 Conv-Max-Pooling layers followed by fully connected layers.

All inputs are padded to the maximum MFCCs Length (250)

1. Convolution by 6*5*5 Filters
2. Max Pooling 2*2
3. Convolution by 16*5*5 filters
4. Max Pooling 2*2
5. Fully Connected 1888 -> 256 with RELU activation
6. Fully Connected 256 -> 128 with RELU activation
7. Fully Connected 128 -> 1 with Sigmoid Activation

4.2. LSTM

Long Short Term Memory Units are used to explore the ability of Recurrent Networks in this task



The network is made of 3 stacked LSTM layers feeding each other, and then, the last hidden state is fed to a fully connected layer.

For each LSTM layer, the size of the hidden dimension is 2*mfcc_length=40

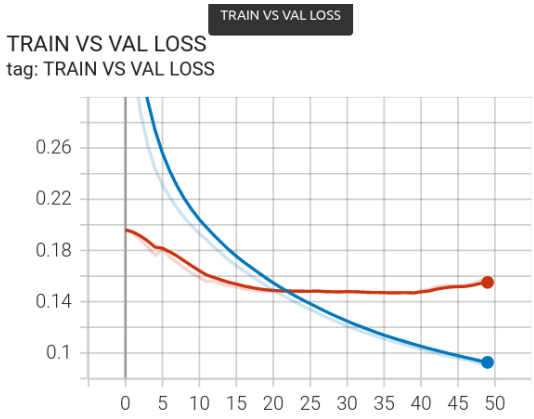
All Sequences are padded to the maximum MFCCs length (250), And dropout is added between different LSTM Layers.

RELU Activations are used in the fully connected layers.

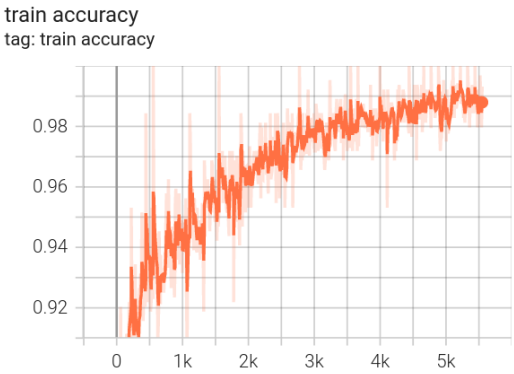
5. Training

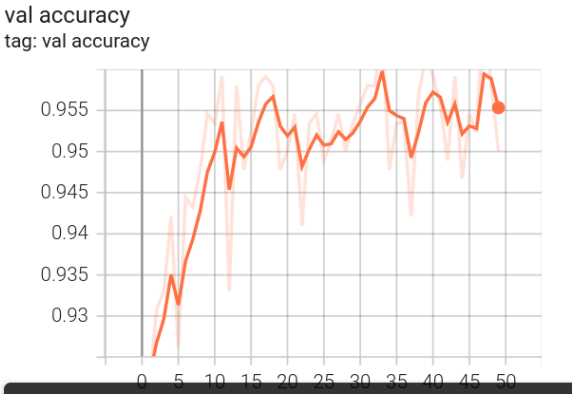
Both the CNN and LSTM networks are trained with similar settings. The loss function is used to be Binary Cross Entropy. The optimizer was chosen to be SGD optimizer. The learning rate was set to constant value of 0.001. The CNN was trained for 50 epochs and LSTM was trained for 100 epochs both with batch size 64. Training was performed on CPU as the number of parameters is not big.

5.1. CNN training



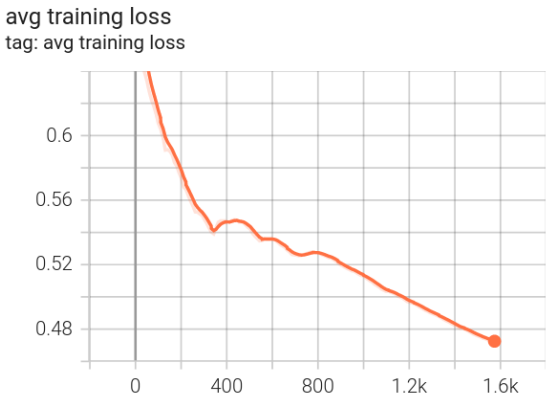
As it can be seen from the plots, The training loss decreases smoothly till the end of the training. However, at some point validation loss starts increasing which means the model may overfit. So the checkpoint at the point of intersection of the train and val curve was chosen to be out best model.

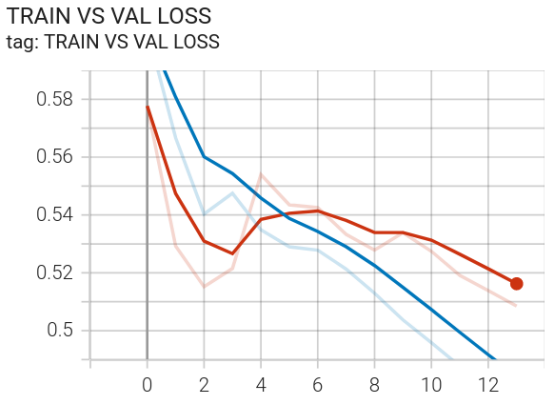




It is shown also that both the train and validation accuracies increase together till the intersection point, so this is a suitable step to take the best checkpoint.

5.2. RNN Training





Although at some point the validation loss starts to increase, it saturates back and decrease again to a new minimum, so the model here doesn't overfit and the last checkpoint is our best model

6. Results

To evaluate our models, the classification accuracy is calculated for the test set. The test contains randomly chosen 10% of the full data set (810 files) not seen by the model during training. We also calculate the precision, recall and f1_score for each of the 2 classes (Male and Female). It is shown that the CNN model outperforms the LSTM model. This may be due to the nature of MFCCs as they can be easily treated as image data.

Results over Test Set Files		
	CNN	RNN
Accuracy	96.25%	84.56%
Recall (male)	0.9763	0.8394
Precision (male)	0.9640	0.9055
F1 (male)	0.9701	0.8712
Recall (female)	0.9399	0.8559
Precision (female)	0.9601	0.7641
F1 (female)	0.9499	0.8074