Speech Emotion Recognition using Time Distributed CNN and LSTM

Beenaa Salian[1,*], *Omkar* Narvade[1,**] *Rujuta* Tambewagh[1,***], and *Smita* Bharne[1,****]

[1]Ramrao Adik Institute of Technology, Navi Mumbai, India

Abstract. Speech has several distinguishing characteristic features which has remained a state-of-the-art tool for extracting valuable information from audio samples. Our aim is to develop a emotion recognition system using these speech features, which would be able to accurately and efficiently recognize emotions through audio analysis. In this article, we have employed a hybrid neural network comprising four blocks of time distributed convolutional layers followed by a layer of Long Short Term Memory to achieve the same.The audio samples for the speech dataset are collectively assembled from RAVDESS, TESS and SAVEE audio datasets and are further augmented by injecting noise. Mel Spectrograms are computed from audio samples and are used to train the neural network. We have been able to achieve a testing accuracy of about 89.26%.

1 Introduction

As humans, our thoughts are best articulated using speech. Therefore, in this increasingly technologically-driven world, the next step forward would be extending this understanding to machines. Although Speech Emotion Recognition (SER) has been around for almost a decade, it has regained attention due to recent developments in this field (for eg. Voice-based virtual assistants like Siri, Alexa, etc and automated help-center assistance, self-driving cars, etc). Although there is a rising demand for voice-controlled technologies, The recognition of emotion from speech is the main challenge in human-machine interaction. Despite significant progress in speech recognition, we are still a long way from determining underlying emotions from the speaker's audio signals since the machine does not understand the speaker's emotional state. Because speech is the simplest and most effective means of communication for humans, a computer must be able to grasp the user's mood in this more technologically-driven world. We'll look at which aspects of speech are the most effective in discriminating various emotions. An improved speech emotion identification model must be able to recognise seven primary emotions: anger, disgust, happiness, sorrow, neutral, surprise, and fear. We must develop a robust deep learning model which can accurately and efficiently classify emotions from speech alone.

There are various hybrid model implementations in the SER domain. The most commonly used hybrid model is the CNN-LSTM model. Where CNN is used for learning local correlations and LSTM is used to learn long-term dependen- cies from the learned local features. In [1] 13 MFCC (Mel Frequency Cepstral Coefficient) with 13 velocities and 13 acceleration components are used as features and a 1D CNN and LSTM model are used for classification. In this paper, the EMODB dataset is used.To compute MFCCs, Discrete Cosine Transform(DCT) is applied. DCT is needed to decorrelate filter bank coefficients. It is a linear operation and therefore discards any non-linear useful information present in the speech signal. In [4] three different features, MFCCs, magnitude spectrogram, and log-mel spectrogram, are compared with several architectures, such as CNN, BLSTM, and CNN-LSTM, to determine which architecture and feature combination is best for speech emotion recognition. All of the models were tested on two different datasets. – EMODB and IEMOCAP where 4 emotions are classified where the length of audio files is kept 3 seconds. The design is shown to perform effectively with Log-Mel Spectrograms when combined with CNN+LSTM architecture in this article. The aim of [3] is to find the relation between the duration of speech length and the recognition rate of emotions. In this paper, analysis is performed using a CNN model having two convolution layers where magnitude spectrograms are taken as features. Performance for the system is analyzed using speech sequences of length from 0.25s to 1.5s. It is observed as the length of the speech signal increases, the accuracy of the system increases.

2 Proposed Methodology

2.1 Speech Corpus

The data sets used here cover people of different ages – ranging from young to old, covers both the genders, and people having different accents.

2.1.1 Toronto emotional speech set (TESS)

200 special words are spoken using the phrase "Say the word - " by two actresses. The two actresses are aged 26

*e-mail: beenaasalian09@gmail.com
**e-mail: nomkar99@gmail.com
***e-mail: rujuta.tambewagh@gmail.com
****e-mail: smita.bharne@rait.ac.in

and 64 years. The audios ae recorded on set potraying one of the seven emotions

2.1.2 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

RAVDESS consists of gender balanced audio samples from males and females (12 females and 12 males). The audio samples contain seven different emotions with two levels of intensity - high and low.

2.1.3 Surrey Audio-Visual Expressed Emotion(SAVEE)

The audio samples in the dataset are recorded with four native english speakers. They are identified as DC, JE, JK, KL. This results in a total of 120 utterances per speaker in which they 7 emotions are covered.

2.1.4 Augmentation

The prediction accuracy of any deep learning model is largely dependent on the amount and the diversity of data available during training. A common method to increase the diversity of your dataset, is to augment your data artificially. To generate syntactic data for audio, we can apply noise injection. We have used the numpy library to add noise to our existing dataset.Table 1 shows the final dataset after augmentation.

Table 1. Final Dataset after Augmentation

Total Dataset After Augmentation								
Code	0	1	2	3	4	5	6	Total
Org	652	652	655	652	652	652	808	4723
Aug	652	652	655	652	652	652	808	4723
Total	1304	1304	1310	1304	1304	1304	1616	9446

$$Emocode\ 0 : Happy, 1 : Sad, 2 : Angry, 3 : Angry,$$

$$4 : Fear, 5 : Surprise, 6 : Neutral$$

2.2 Feature Extraction

There is quite a lot of varied information present in speech signals, most of which doesn't aid us in our objective of recognizing emotions from audio files. Hence, we extract the relevant information from these speech signals and provide these as the feature vector input to our classifier. We have used Mel Spectrograms as our feature to be extracted which would be further used to train our classifier. The steps for feature extraction can be seen in figure 1.

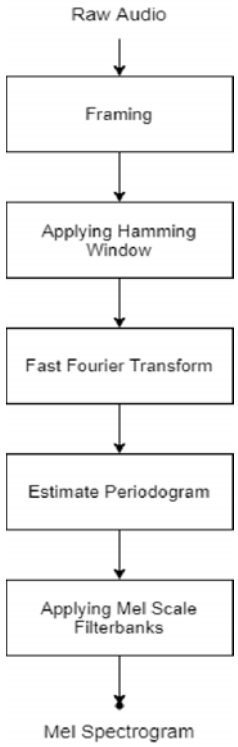

Figure 1. Feature Extraction Steps

2.2.1 Preprocessing

Every audio sample in our dataset is first sampled at 21,500 Hz with an offset of 0.5 seconds and according to the calculations, the average length of audio samples in our dataset is around 3 seconds. Further, Z-distribution, also known as standard normal distribution is performed on the samples and then, we adjust these audio files to be approximately 3 seconds long, i.e 650,000 samples per audio. Adjustment is made by truncating the files, if they have a length greater than average or by padding it with zeroes if it is lesser than the same.

2.2.2 Time Window Framing

We assume that the properties in a non-stationary speech signal signal remain constant over a very short period of time. These short time periods are termed as 'frames', and these frames are long enough to contain vital characteristics and short enough for it to be considered stationary. Here, window size is 23ms with an overlapping of 50frames.Framing begins with the first N = 256 samples, the second frames starts with a hop of M = 128 samples and overlaps the first frame by N M , and this is repeated for the entire signal. This overlap smoothenes the transitions between the frames.

2.2.3 Applying Hamming Window

The signal is further passed through a Hamming window to smoothen out the signal and makes sure that the neighbouring window-ends match. Also, it leads to a better signal clarity and helps in reducing the spectral leakages The

equation to calculate the same is given by (1)

$$w(k) = 0.54 - 0.46cos(\frac{2\pi k}{N-1}) \quad (1)$$

where $w(k)$ is the window function and $0 \leq k \leq N-1$

2.2.4 Fast Fourier Transform

Fast Fourier Transform takes a sequence of discrete signal amplitudes as input, and converts it into it's frequency con- stituents and is given by (2). We perform an N-point FFT on every frame in the signal to calculate the overall frequency spectrum. This is also termed as Short Term Fourier Transform (STFT).

$$X_k = \sum_{n=0}^{N-1} X_k e^{\frac{-i2\pi}{N}kn}$$

$$k = 0, 1, 2, 3, ..., N-1 \; and \; N = 512 \quad (2)$$

2.2.5 Periodogram Estimate

To compute a periodogram estimation, we square the absolute value of the result derived from the complex fourier transform operation. This estimation helps us in identifying which frequencies are present in every frame extracted from the audio sample. The speech frame's periodogram-based power spectral estimate is given by

$$P_i(k) = \frac{1}{N} \mid S_i(k) \mid^2 \quad (3)$$

2.2.6 Applying Mel-Scale Filterbanks

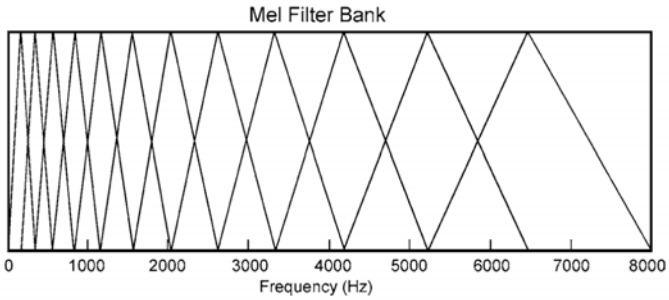

Figure 2. Mel Filterbank

Mel scale is a non-linear transformation and is based on the perception of the audio sample by the human auditory system. To compute the Mel Spectrogram, we just need to apply the mel-spaced filterbanks as seen in figure 2, which is a set of overlapping triangular filters.Starting of one filter overlaps with the centre of the previous one whereas the ending part overlaps with the centre of the succeeding one and so on. It has a response of 1 at the centre and decreases as we move to the ends, where it is 0. The filters are closely spaced and are narrow at the lower frequencies, As the frequency increases, the filters get wider and they turn less discriminative or less sensitive to the variations in the frequency. Mel-Scale tells us how to space the filterbanks on the scale and gives an estimation on how wide to make them as the frequency increases. For converting f Hertz to m mels, we apply

$$Mel(m) = 2595 \times \log_{10}(1 + \frac{\mathrm{f}}{700}) \quad (4)$$

3 Classifier

As seen in Figure 3, a hybrid neural network model consisting of Time Distributed CNN followed by a LSTM layer has been proposed. CNN has proven to be a breakthrough in terms of performance related to image recognition and various computer vision tasks. Long Short Term Memory has proven to be very useful in the case of analyzing sequential data. Therefore, it would help the model to learn both short-term and long-term feature dependencies by using the two in succession. CNN and LSTM, can hence, take advantage of the strengths of both networks.

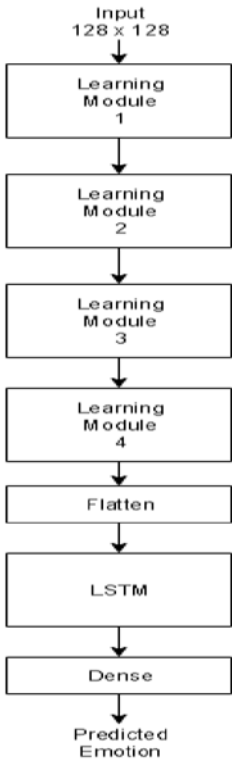

Figure 3. Classifier Model

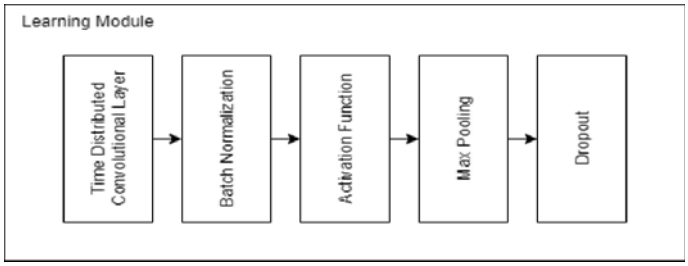

Figure 4. Learning Module

3.1 Time Distributed Convolution Layers

The main idea of the time distributed convolutional layers is applying a rolling window over our input feature, i.e Mel spectrogram. Hence, we get a sequence of images, and the sequence of these images is provided as an input to the first layer in our neural network. As seen in figure 4, this part of the model is subdivided into four Learning Modules (LM). Figure 4 shows that each of these LMs constitutes a time distributed convolutional layer, batch normalization, activation function, max pooling layer, and a dropout layer. Exponential Linear Unit (ELU) is used as the activation function in these four LMs followed by a Max Pooling layer. This is used to lower the size of the image's feature maps, which helps to minimise the number of trainable parameters even more. The dropout regularization helps to avoid overfitting by randomly dropping out a few neurons from the neural network layer.

3.2 Long Short Term Memory

Speech signal is time-varying, and also spectrograms primarily have a time component, so it is well worth trying to explore these temporal properties of speech audio. Long Short Term Memory has proven to be very useful in the case of analyzing sequential data. Therefore, it could help to identify and extract the global temporal features from the mel spectrogram. An LSTM layer having 256 nodes is added as a learning layer in the model followed by a dense fully connected layer having 'softmax' as its activation function.

4 Experiments

The compiled dataset is split into train and test sets, where 80% of the total is used for training the model and the remaining 20% is evaluating the model's performance. Hence, we train with 7557 audio samples and test on 1889 audio samples. After applying a rolling window on the mel spectrogram, a sequence of 6 overlapping images is generated for each audio file, and this is provided as input to our first Learning Module (LM). As seen in figure [3] are four LM's in our model and each one consists of a time distributed convolutional layer, batch normalization layer, an activation function layer, dropout layer, and lastly the max pooling layer. The convolutional layers have a kernel size of 3×3. The layers in the first two blocks have 64 feature maps, and the subsequent two have 128 feature maps. The activation function used in all four blocks is Exponential Linear Unit as this function tends to converge faster and provide better accuracies. After the four LM blocks, the resultant output is flattened and provided as an input an LSTM layer, followed by a fully connected layer with a softmax activation function which provides an output that maps to the predicted emotion for every audio sample. The optimizer for this model is Stochastic Gradient Descent (SGD), with a learning rate of 0.01. We applied early stopping having patience as 15, with respect to maximizing the validation accuracy.

5 Results

The model is trained for a set of 100 epochs and compiled with categorical cross-entropy as the loss function. By using 'Model Checkpoint', we monitored and saved the weights that provide maximum value for validation accuracy. Our best model saved through this provides an accuracy of 90.64% on the training set and 89.26% on the testing set.

5.1 Performance Matrix

Table 2. Performance of the model

	Evaluation Metrics		
Emotions	**Precision**	**Recall**	**F1-score**
Neutral	92	90	91
Happy	82	87	84
Sad	92	81	86
Angry	90	92	91
Fear	91	92	91
Disgust	84	96	90
Surprise	95	87	91

Table 2 summarises our model's performance in terms of various performance metrics. Precision is the ability of our model to check the correctly predicted positives from all the predicted positives and is given by Eq. (5). Our model has the highest precision of 95% for surprise emotion and lowest for happy emotion with precision of 82%.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5}$$

Recall measures the model ability to check the correct positive from all the existing positives in the test dataset and is given by Eq. (6). We can see from table 2 that our model has the best recall score for disgust emotion of 96%and the worst recall score of 81% for sad emotion.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{6}$$

The F1 score is the harmonic mean of precision and recall, which is used to measure an emotion's overall performance. Eq. (7). The F1 score is The best F1 score is for surprise, neutral, angry, fear emotions which is 91% and the worst F1 score is 84% for happy emotion.

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{7}$$

Model accuracy score is the most intuitive performance measure, and it is simply a ratio of correctly predicted observation to the total observations. Our model has an accuracy of 90.64% on training set and 89.26% on validation data set.

Table 3. Comparison with Existing System

Parameters	Proposed Model	System [4] pt. 1
Architecture	Time Distributed CNN + LSTM	CNN
Dataset	RAVDESS + SAVEE + TESS	EmoDB
Features Used	Log Mel-Scale Filterbanks	Mel Spectrogram
Dataset Distribution	7720 and 1450	271 and 68
Accuracy	**89.26 %**	**78.16 %**

Parameters	Proposed Model	System [4] pt. 2
Architecture	Time Distributed CNN + LSTM	Bi-LSTM
Dataset	RAVDESS + SAVEE + TESS	Iemocap
Features Used	Log Mel-Scale Filterbanks	Mel Frequency Cepstral Coefficient
Dataset Distribution	7720 and 1450	4424 and 1107
Accuracy	**89.26 %**	**46.21 %**

In the above tables, we have compared our system to two pre-exiting systems presented in [4], based on various parameters.

5.2 Loss and Accuracy Curves

A learning curve is a diagnostic tool that shows the performance of the model over the period of time. We have in the plotted the loss curve and accuracy curve for the training of our model. We can see that Figure 5 shows a steep decline in training loss for the first 20 epochs and then we notice steady decline of training loss till 100th epoch and Figure 6 shows a sharp increase in the accuracy for the first 20 epochs and then we see that there is a steady increase of the accuracy till 100th epoch.

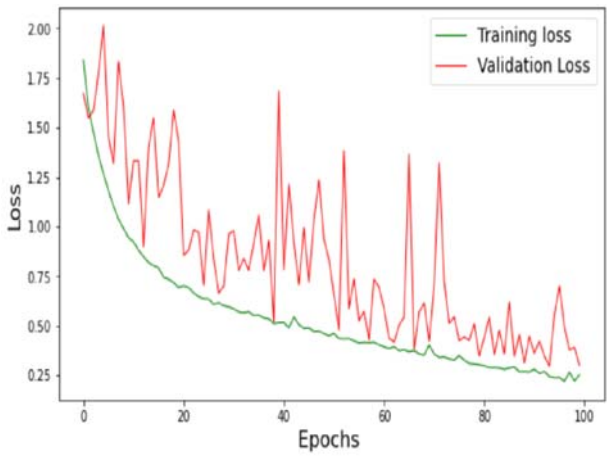

Figure 5. Loss Curves

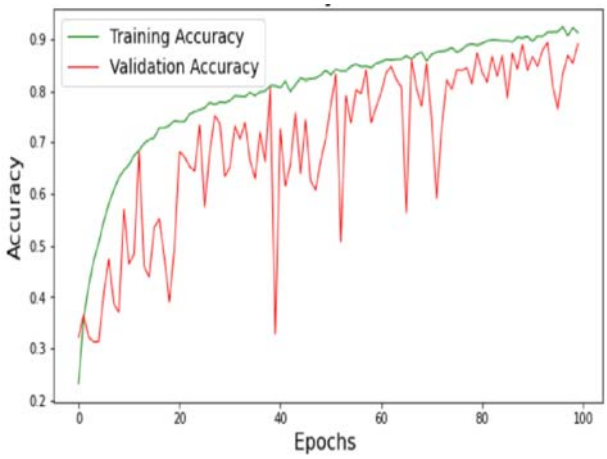

Figure 6. Accuracy Curves

5.3 Confusion Matrix

Confusion matrix helps us evaluate the performance of our neural network, when it makes predictions on the test data, and it helps us to analyse how accurate our recognition model is with respect to specific emotions. From the confusion matrix seen in figure 7, we can conclude that:

1. The model performs exceptionally well, while predicting fear and disgust emotions as these emotions have a result of 251/274 and 249/259 respectively.

2. On the other hand emotions like happy and sad have a result of 220/254 and 208/258,hence there is scope of improvement in these emotions.

3. The model often predicts a sad emotion as neutral emotion. By coming up with a solution for this problem, we can improve the performance of the system to a great extent.

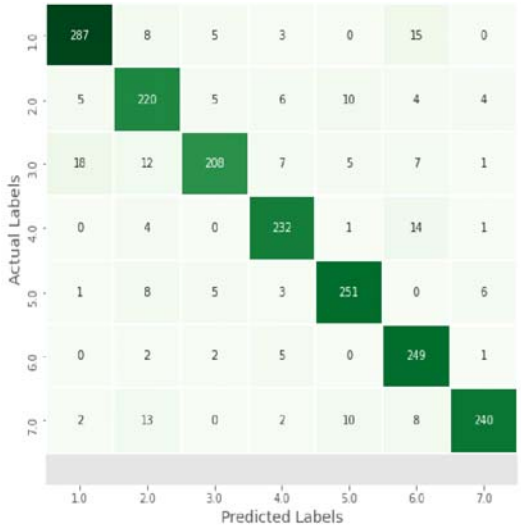

Figure 7. Confusion Matrix

6 Conclusion

This research presents a hybrid neural network strategy for detecting underlying emotions in audio samples. In terms of performance, a Time Distributed CNN + LSTM model trained on a large gender-balanced dataset of speakers with various accents and nationalities outperforms other models. Neutral, Angry, Fear, Disgust, and Surprise all had testing accuracy of 90% or higher. This model's performance could be improved even further through targeted training, by focusing on emotions such as happy and sad, which have accuracy rates of around 84% and 86%, respectively.

References

[1] S. Basu, J. Chakraborty and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 2017, pp. 333-336, doi: 10.1109/CESYS.2017.8321292.

[2] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 177-183, doi: 10.1109/COMITCon.2019.8862221.

[3] B. Puterka and J. Kacur, "Time Window Analysis for Automatic Speech Emotion Recognition," 2018 International Symposium ELMAR, Zadar, 2018, pp. 143-146. doi: 10.23919/ELMAR.2018.85.

[4] S. K. Pandey, H. S. Shekhawat and S. R. M. Prasanna, "Deep Learning Techniques for Speech Emotion Recognition: A Review," 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), 2019, pp. 1-6, doi: 10.1109/RADIOELEK.2019.8733432.

[5] A. B. Abdul Qayyum, A. Arefeen and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," 2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON), 2019, pp. 122-125, doi: 10.1109/SPICSCON48833.2019.9065172.

[6] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 2257-2260, doi: 10.1109/WiSPNET.2017.8300161.

[7] B. Puterka, J. Kacur and J. Pavlovicova, "Windowing for Speech Emotion Recognition," 2019 International Symposium ELMAR, 2019, pp. 147-150, doi: 10.1109/ELMAR.2019.8918885.

[8] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2010

[9] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[10] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-levelinformation for robust recognition of emotions within speech,"Proceedings of Interspeech,pp. 2249–2252, 01 2007.