Analyzing Speech to Detect Emotions using Deep Neural Networks

Satyabrat Bhol[1], Deepak Hirawat[2], Momojit Ghosh[3], Dr. Rupesh Kumar Sinha[4]

[1,2,3]Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India
[4]Faculty of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India
[1]Corresponding author E-mail: satyabrat35@gmail.com
E-mail: dhirawat20@gmail.com, momojit.ghosh@gmail.com, rk.sinha@bitmesra.ac.in

Abstract— **Emotion detection has many applications in understanding the behavior of a person. The speech can be utilized to derive the emotion of a human being. It is used by organizations in analyzing the reviews and feedbacks form their valued customers. We are presenting a classification model of emotions using deep neural network. There is a total of eight different emotions classified using this neural network model – happy, calm, neutral, sad, angry, disgust, fearful and surprise. The model produced an overall F1 score 0.80 where class "Calm" had the best score of 0.84 while class "Angry" had the worst score of 0.74. In or proposed work, we have used the Mel Frequency Cepstral Coefficients (MFCC) features to train our model.**

Keywords— emotion classification; mfcc; ravdess; cnn; rf; knn; f1 score

I. INTRODUCTION

Emotion detection using speech signal, as challenging it is, is beneficial in making the interaction between humans and machines much easier. It will improve the way machines react human emotions and act accordingly. There are various features that contribute to a speech signal like amplitude, speech, and frequency. To recognize emotions many approaches such as Mel frequency cepstral coefficient (MFCC), Linear prediction cepstral coefficient (LPCC) and wavelet features have been used. But Mel frequency cepstral coefficient (MFCC) is the most used feature [1]. In this work, a total of eight different emotions (neutral and calm plus the basic ones as proposed by Ekman [2]) have been used to train the model.

II. RELATED WORK

The complex task of classifying emotions is due to ambiguity of natural and acted emotions as explained by [3]. Their work [3] involved many machine learning techniques to create agents. Those agents were able to predict emotions of five different states (angry, happy, sad, fear and normal) with a varying accuracy ranging from 55-85%. They [3] separated the male and female speech data to improve the accuracy of emotion classification. Their modified MFCC approach produced an overall accuracy of 63% while the standard approach had an overall accuracy of 54%.

Emotion classification using Gaussian Mixture Models (GMMs) was proposed by [4]. A private database with content recorded by male and female actors were used for emotion recognition. The data was categorized into four basic emotions – neutral, sad, happy, and angry. For test data, a male actor's recorded voice was used. The Gaussian Mixture Model (GMM) produced an accuracy of 85% for classification of gender specific emotions. But it was observed that when the model was trained with a mixture of male and female recorded voices, the performance of the proposed GMM system [4] decreased indicating the model was gender dependent.

A study [5] on the emotion recognition based on discrete emotion classification was directed to design an efficient classifier. Their model for emotion recognition was constructed on support vector machine (SVM) and artificial neural network (ANN). Experimental results revealed that support vector machine (SVM) produced an accuracy of 85% and artificial neural network (ANN) produced an accuracy of 75%. It was proposed by [5] that support vector machine (SVM) slightly outperforms the artificial neural network (ANN).

III. DATASET

Dataset used in this proposed work is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6]. The dataset contains 1440 Audio-only (16bit, 48kHz .wav) files made by 24 professional actors (12 females, 12 males). There are total of 8 emotions depicted by the actors (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprise). Song contains happy, calm, sad, angry, and fearful emotions while speech contains calm, happy, sad, angry, fearful and surprise. There are two kinds of emotional intensity normal (01) and strong (02). Files have been randomly split into train and test datasets. Test dataset consists of 33% of the original dataset.

Each of the RAVDESS file has a unique name with a 7-part numerical identifier e.g. 03-01-01-05-02-

10.mp4. All the identifier has their own characteristic like modality (03 for all files since we are using only audio recordings), vocal channel (01 for speech, 02 for song), emotion (eight different emotions ranging from 01-08), emotional intensity (01 for normal, 02 for strong), statements (two different statements are being used for depicting emotion), repetition (two repetitions) and actors (ranging from 01-24. Odd numbered actors are male while even numbered are females).

IV. FEATURE EXTRACTION

In this section, the role and calculation of Pre-Emphasis filter, Framing, Windowing, Filter banks and MFCC are briefly described.

A. Pre-Emphasis Filter

Pre-emphasis boosts the relative amplitude for the high frequencies. In audio signals low frequencies have more power than higher frequencies (because of 1/f characteristics) giving rise to spectral tilt. Boosting the higher frequencies makes information available to the model. We used pre-emphasis to make the system less prone to the noise. Moreover, the pre-emphasis filter helps in improving the Signal-to-Noise Ratio (SNR).

B. Framing and Windowing

The process of decomposing the speech signal into a series of overlapping frames is called framing. The frame block consists dividing the speech signal into short frames of specific samples [7]. We applied window function to each such frame (Hamming window). It prevents the signal from getting contaminated by reducing the side lobe of the frames.

C. Filter Banks

Filter banks are arrangements of low pass, bandpass, and high pass filters used for the spectral decomposition and composition of signals [8]. To compute filter banks, we applied triangular features on a Mel-scale to extract frequency band. Python library ‘LibRosa’ [9] is being used for applying pre-emphasis filters, framing , windowing and finally compute the filter bank values.

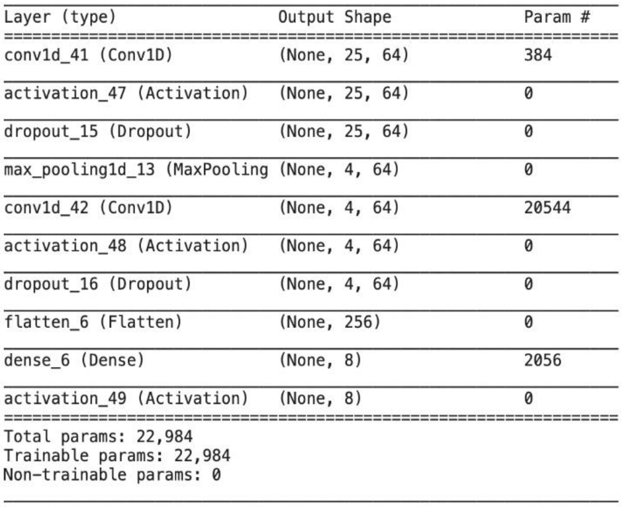
D. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to Mel Scale, and are thus suitable for speech recognition tasks quite well (as they are suitable for understanding humans and the frequency at which humans speak/utter) [10]. 25 Mel Frequency coefficients were considered as features in training our model. We used ‘LibRosa’ library in generating the coefficients as well.

V. PROPOSED MODEL

The proposed model is based on the combination of convolutional neural network and dense layers. Mel-frequency cepstral coefficients (MFCC) were used as features to train the model. From each audio files, a total of 25 features have been extracted.

The deep neural network for the classification task is provided in Fig. 1. Input of size *{number_of_training_files} x 25 x 1* was used. One round of 1D CNN with a ReLu activation function, dropout of 10% and a max pooling of 2x2 was performed. The rectified linear unit (ReLu) returns 0 for any negative input, but for positive input it returns the same value - *f(x)=max(0,x)*. The constant gradient of ReLu helps in faster learning. Pooling layer is used to reduce the resolution of feature map but retaining features of the map for classification [11]. We, then used a Dense layer with a softmax activation function (output size - 8) and calculated the probability distribution of each classes.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_41 (Conv1D)           (None, 25, 64)            384
_________________________________________________________________
activation_47 (Activation)   (None, 25, 64)            0
_________________________________________________________________
dropout_15 (Dropout)         (None, 25, 64)            0
_________________________________________________________________
max_pooling1d_13 (MaxPooling (None, 4, 64)             0
_________________________________________________________________
conv1d_42 (Conv1D)           (None, 4, 64)             20544
_________________________________________________________________
activation_48 (Activation)   (None, 4, 64)             0
_________________________________________________________________
dropout_16 (Dropout)         (None, 4, 64)             0
_________________________________________________________________
flatten_6 (Flatten)          (None, 256)               0
_________________________________________________________________
dense_6 (Dense)              (None, 8)                 2056
_________________________________________________________________
activation_49 (Activation)   (None, 8)                 0
=================================================================
Total params: 22,984
Trainable params: 22,984
Non-trainable params: 0
_________________________________________________________________
```

Fig 1 – Proposed Neural Network Model

VI. MODEL COMPARISION AND RESULT

Different models of classification were used to compare the accuracy with the proposed model (PM). For the first model, we used a K-Nearest Neighbors (KNN) Classifier with 10 neighbors and for the second model, we used a Random Forest (RF) Classifier with 500 trees. The test data size was 474 x 25. We used the F1 score [12] as an evaluation parameter for comparing our models against the proposed model.

The model achieved an overall F1 score of 0.80. The small variation in the F1 score for all eight different classes signifies the robustness of the model. The model was trained on 150 epochs but around 85th-90th epoch the validation loss and validation accuracy reached a stagnant value. Some tweaks made to model in terms of dropout value, regularization value and activation function did not produce much considerable improvement to the accuracy.

The author believes that the proposed model can provide similar accurate results on audio files in real time scenario with environment noise. Future consideration of this work would be using additional data set like Toronto Emotional Speech Set (TESS) [13], Surrey Audio-Video Expressed Emotion (SAVEE) [14] etc.

Emotion Class	**Precision**	**Recall**	**F1 score**
Neutral	0.79	0.77	0.78
Happy	0.82	0.81	0.81
Sad	0.82	0.83	0.82
Calm	0.84	0.85	0.84
Fearful	0.81	0.78	0.79
Angry	0.73	0.76	0.74
Surprise	0.81	0.83	0.82
Disgust	0.82	0.8	0.81
Accuracy			0.80

Fig 2 – Precision, Recall and F1 Score of the Model (PM) for each class

Model	**Avg F1 Score**
KNN	0.67
RF	0.54
PM	0.80

Fig 3 - Average F1 Score for all tested models

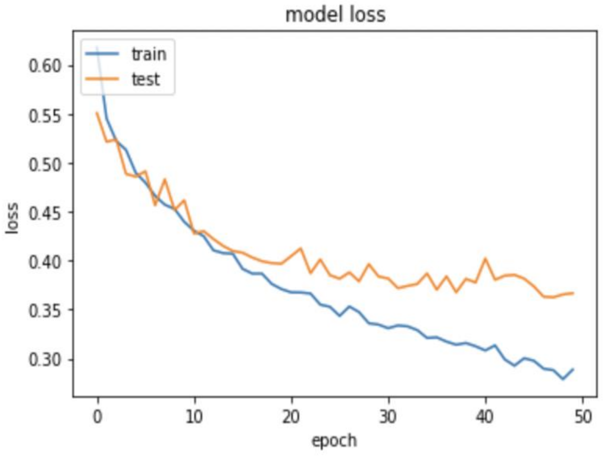


Fig 4 - Model Loss of our proposed model over 50 epochs

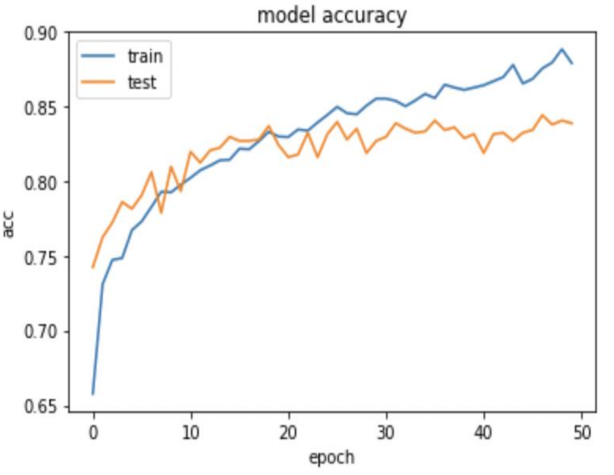


Fig 5 – Model Accuracy of our proposed model over 50 epochs

VII. CONCLUSION

Presented neural network model for classification of emotions using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset having an overall F1 score of 0.80. Model was trained on eight different emotions – calm, happy, angry, sad, neutral, fearful, disgust and surprise. Calm class had the best score of 0.84 while Angry class had the worst score of 0.76. We tried out other models as baseline like K-Nearest Neighbors and Random Forest which had an overall F1 score of 0.67 and 0.54 respectively.

VIII. ACKNOWLEDGEMENT

The authors gratefully thank Electronics and Communication Dept. (ECE) of BIT Mesra for their immense support and encouragement.

REFERENCES

1. S. Demircan and H. Kahramanlı, "Feature Extraction from Speech Data for Emotion Recognition," Journal of Advances in Computer Networks, vol. 2, no. 1, pp. 28-30, 2014.
2. Ekman, P. (1992a). An arugment for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068
3. A. K. Komal Rajvanshi, "An Efficient Approach for Emotion Detection from Speech Using Neural Networks," International Journal for Research in Applied Science & Engineering Technology, vol. 6, no. 5, pp. 1062- 1065, 2018.
4. K. S. Rao, T. P. Kumar, K. Anusha, B. Leela, I. Bhavana and S. V. Gowtham, "Emotion Recognition from Speech," International Journal of Computer Science and Information Technologies, vol. 3, no. 2, pp. 3603- 3607, 2012.
5. X. Ke, Y. Zhu, L. Wen and W. Zhang, "Speech Emotion Recognition Based on SVM and ANN," International Journal of Machine Learning and Computing, vol. 8, no. 3, pp. 198-201, 2018.
6. LIVINGSTONE, S. R., AND RUSSO, F. A. 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 (2018), e0196391.
7. Kamil, Oday. (2018). Frame Blocking and Windowing Speech Signal. 4. 87-94.
8. Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications. Alfred Mertins
9. Vincentius Satria Wicaksana and Amalia Zahra S.Kom, “Spoken Language Identification on Local Language using MFCC, Random Forest, KNN, and GMM” International Journal of Advanced Computer Science and Applications(IJACSA), 12(5), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120548
10. Mel Frequency Cepstral Coefficient – Research Gate Topic
11. Purpose of Pooling - https://www.kaggle.com/questions-and-answers/59502
12. Accuracy vs F1 score - https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2
13. Toronto emotional speech set (TESS) - https://tspace.library.utoronto.ca/handle/1807/24487
14. Surrey Audio-Visual Expressed Emotion (SAVEE) Database - http://kahlan.eps.surrey.ac.uk/savee/