SPEECH EMOTION RECOGNITION

Aditi Ravichandra[1], Anoushka Prasad H N[2], Bhavya S[3], Harshitha M[4], Keerthana M G[5]

[1]Assistant Professor, Dept. of Computer Science and Engineering, Atria Institute of Technology, Karnataka, India
[2]Anoushka Prasad H N,UG student, Dept. of Computer Science and Engineering, Atria Institute of Technology , Karnataka, India
[3]Bhavya S, UG student, Dept. of Computer Science and Engineering, Atria Institute of Technology, Karnataka, India
[4]Harshitha M, UG student, Dept. of Computer Science and Engineering, Atria Institute of Technology, Karnataka, India
[5]Keerthana M G, UG student, Dept. of Computer Science and Engineering, Atria Institute of Technology, Karnataka, India

---

Abstract - *Speech Emotion Recognition is a field of artificial intelligence and machine learning which is used to recognize emotion from speech. Speech is language through which humans vocally communicate. Each language uses phonetic combinations of vowel and consonant sounds that form the sound of its words. An emotion expresses a human's mental state and is generated subconsciously. Developing machines that understand, recognize emotion from speech will make human-machine interaction more clear and natural. In this paper, an intelligent model is proposed to recognize the emotion of the user based on their speech using a deep learning algorithm. Convolutional Neural Network (CNN) is the deep learning algorithm used along with feature extraction techniques to recognize emotion from speech.*

Key Words: **CNN, RAVDESS, Mel-Frequency Cepstral Coefficients, Emotion detection**

1. INTRODUCTION

Emotion is a strong feeling which is derived from circumstances, mood or relationships with others and is an important part of a human's life. It is closely related to our decision making approach. With the recent boom in machine learning backed speech processing, numerous tools are being used in emotion recognition. Speech emotion recognition systems are built to identify emotions from speech and classify them into a few categories such as happiness, sadness, anger**,** disgust, fearful and neutral. Determining the emotional state of a human is an idiosyncratic task and may be used as a standard for any emotion recognition model. Recognizing of emotion from speech is a challenging task for many reasons. First is selecting the best features, which are powerful enough to distinguish between different emotions. Second is the presence of various languages, accents, sentences, dialects and speaking styles**.** The types of speakers also add another difficulty because these characteristics directly change most of the extracted features including pitch, energy**.**

2. DATABASE DESCRIPTION

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contain 7356 files with a total size of 24.8 GB. The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All file are available in three modality formats: Audio-only, Audio-Video, and Video-only (no sound). There are no song files for Actor_18 and emotions disgust, neutral and surprised are not included in the song version of the data. Only audio speech and song files have been used for this particular project.

Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics:

TABLE- 1: Filename Identifiers

FILENAME IDENTIFIERS	MEANING
Modality	01 = Full-Audio Video 02 = Video-only 03 = Audio-only
Vocal Channel	01 = Speech 02 = Song
Emotion	01 = Neutral 02 = Calm 03 = Happy 04 = Sad 05 = Angry 06 = Fearful 07 = Disgust 08 = Surprised
Emotional Intensity	01 = Normal 02 = Strong
Statement	01 = "Kids are talking by the door" 02 = "Dogs are sitting by the door"
Repetition	01 = 1st Repetition 02 = 2nd Repetition
Actor	01 - 24 Odd numbered actors are male Even numbered actors are female

Filename example: 03-01-06-01-02-01-12.mp4

1. Audio-only (03)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)
8. Female, as the actor ID number is even.

The total class of samples used in the proposed system:

TABLE - 2: Total Sample class

Emotion	Speech Sample Count	Song Sample Count	Summed Count
Neutral	96	92	188
Calm	192	184	376
Happy	192	184	376
Sad	192	184	376
Angry	192	184	376
Fearful	192	184	376
Disgust	192	0	192
Surprised	192	0	192
Total	1440	1012	2452

3. LITERATURE SURVEY

3.1 A hierarchical support vector machine based on feature-driven method for speech emotion recognition [1]

Through the analysis of one-vs.-one, one-vs.-rest and the decision tree mechanism of binary support vector machine emotion classifiers, a method based on feature-driven hierarchical support vector machine is proposed for speech emotion recognition. For each layer, classifier used different feature parameters to drive its performance, and each emotion is subdivided layer by layer. This method did not rely entirely on the activity-valance dimensional emotion model, but relied on the type of emotion to distinguish. Further more, classifications are constructed by appropriate characteristic parameters ultimately. Experiments on the Chinese-speaker dependent and Berlin-speaker-independent corpus reached conclusions as follows, Chinese-speaker-dependent recognition rate is relatively higher than Berlin-speaker-independent. Feature-driven hierarchical support vector machine in the case driven by effective features improves the speech emotion recognition performance. Meanwhile applying the mean of the log-spectrum to this method can identify high-activity and low activity emotion effectively.

3.2 Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm [2]

Speech is an interactive interface medium as it is possible to express emotions and attitude through speech. In this paper, a hybrid of Hidden Markov Models (HMMs) and Support Vector Machines (SVM) has been proposed to classify four emotions viz. happy, angry, sad and aggressive. Combining advantage on capability to dynamic time warping of HMM and pattern recognition of SVM. HMMs, which export likelihood probabilities and optimal state sequences, have been used to model speech feature sequences i.e. our proposed system is trained using HMM algorithm for emotions considered, while SVM has been employed to make a decision i.e. for classification. The recognition result of the hybrid classification has been compared with the isolated SVM and the maximum recognition rates have reached 98.1% and 94.2% respectively.

3.3 Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine [3]

In human machine interaction automatic speech emotion recognition is yet challenging but important task which paid close attention in current research area. As the role of speech is an increase in human computer interface. Speech is attractive and effective medium due to its several features expressing attitude and emotions through speech is possible. here study is carried out using Gaussian mixture model and support vector machine classifiers used for identification of five basic emotional states

of speaker's as angry, happy, sad, surprise and neutral. In this paper to recognize emotions through speech various features such as prosodic features like pitch , energy and spectral features such as Mel frequency cepstral coefficient were extracted and based on this features emotional classification and performance of classification using Gaussian mixture model and support vector machine is discussed.

4. PROPOSED SYSTEM

4.1 METHODOLOGY

The speech emotion recognition model proposed here is based on CNN. The key idea is considering the MFCC [4] commonly referred to as the "spectrum of a spectrum", as the only feature to train the model. MFCC is a different interpretation of the Mel-frequency cepstrum (MFC), and it has been demonstrated to be the state of the art of sound formalization in automatic speech recognition task [5]. The MFC coefficients have mainly been used to represent the amplitude spectrum of the sound wave in a compact form. As described in [4], the audio file is divided into frames, usually using a fixed window size, in order to obtain statistically stationary waves. The amplitude spectrum is normalized with a reduction of the "Mel" frequency scale. This operation is performed for a significant reconstruction of the wave as the human auditory system can perceive. For each audio file, 40 features have been extracted. The feature has been generated converting each audio file to a floating-point time series. Then, a MFCC sequence has been created from the time series. The MFCC array has been transposed and the arithmetic mean has been calculated on its horizontal axis.

4.2 ALGORITHM

The Convolutional Neural Network (CNN) designed for the classification task is shown in Fig.1. The network is able to work on vectors of 40 features for each audio file provided as input. The 40 values represent the compact numerical form of the audio frame of 2s length. Consequently, we provide as input a number of training files of size (40 x 1) on which we performed one round of a 1D CNN with a ReLU activation function [6], dropout of 20% and a max-pooling function 2 x 2. The rectified linear unit (ReLU) can be depicted as g(z) = max{0, z}, and it allows us to obtain a large value in case of activation by applying this function as a good choice to represent hidden units. Pooling can help the model to focus only on the principal characteristics of every portion of data, making them invariant by their position. We have applied another dropout and then flatten the output to make it compatible with the next layers. Finally, we applied one Dense layer (fully connected layer) with a Softmax activation function, varying the output size from 640 elements to 8 and estimating the probability distribution of each of the classes properly encoded.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_3 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_4 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_3 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 10, 64)            0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 10, 128)           41088
_________________________________________________________________
activation_5 (Activation)    (None, 10, 128)           0
_________________________________________________________________
dropout_4 (Dropout)          (None, 10, 128)           0
_________________________________________________________________
max_pooling1d_3 (MaxPooling1 (None, 2, 128)            0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 2, 256)            164096
_________________________________________________________________
activation_6 (Activation)    (None, 2, 256)            0
_________________________________________________________________
dropout_5 (Dropout)          (None, 2, 256)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 4104
_________________________________________________________________
activation_7 (Activation)    (None, 8)                 0
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
_________________________________________________________________
```

Fig-1: Detailed architecture of the proposed system

4.3 WORKING

The proposed system is trained for accuracy estimations and prediction of emotions. A particular audio file is imported and its characteristics are identified by plotting a waveform.

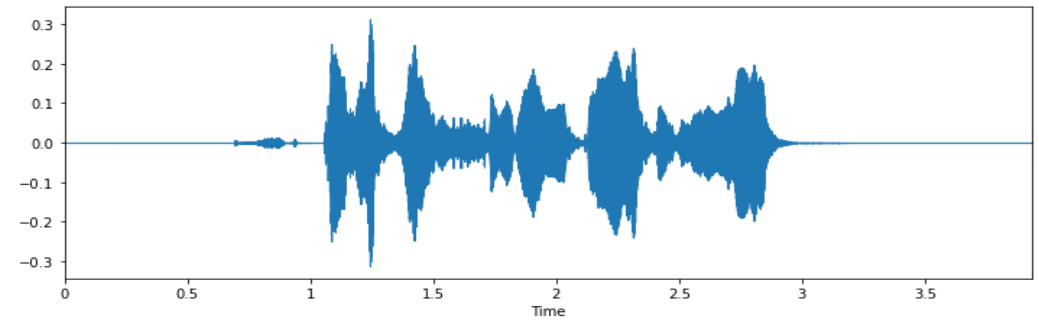

Chart-1: Time domain plot of the speech signal

After the required characteristics have been identified, the entire dataset featuring the audio speech files and audio song files from RAVDESS are loaded into the proposed system. Mel-Frequency Cepstral Coefficients (MFCCs) of each file are extracted and stored as a numpy array. The sampling rate value is obtained using librosa packages and MFCC function. This value holds other variables. Now audio files and MFCC value hold a variable, so consequently it will add a list. Zip the list and apply them to two variables X & y. Then represent (X, y) shape values with the use of numpy package. The proposed system is developed using three 1D CNN layers, four Activation layers (3 ReLU, 1 Softmax), three Dropout layers, two Max-Pooling layers, one Flatten layer and one Dense layer. The loss function is sparse_categorical_crossentropy and the evaluation metric is accuracy. A detailed description of the algorithm is given in the subsection 4.2 and Fig. 1. Trial and error was used to determine the right batch-size and epochs to avoid overfitting of data. Once the ideal values were obtained the proposed model underwent training and a desired accuracy rate was obtained. This model was then used to predict the emotions of specified audio files.

4.4 EXPERIMENTAL RESULTS

A lot of time was spent to determine the test-size, batch-size and epochs. Initially the test-size was set to 0.33 which was not ideal for this model since only 2452 samples were used. This was later changed to 0.2 and it led to an increase in the accuracy rate. The desired batch-size and epochs were obtained through a lot of trial and error. The values for both need to be determined in such a way that it does not lead to overfitting of data which in turn leads to reduced accuracy.

We ended up using batch-size = 4 and epochs = 200 which did not cause overfitting of data and also gave us an accuracy rate of 72.30%. The model was then used to predict the emotions of the specified files. Although only some of the emotions was predicted correctly by the model, this indicated that the model required even more training and testing by using different and more datasets consisting of different languages and accents.

	precision	recall	f1-score	support
0	0.71	0.77	0.74	35
1	0.82	0.72	0.77	80
2	0.71	0.79	0.75	76
3	0.65	0.67	0.66	78
4	0.82	0.80	0.81	79
5	0.83	0.61	0.70	66
6	0.58	0.74	0.65	35
7	0.62	0.69	0.65	42
accuracy			0.72	491
macro avg	0.72	0.72	0.72	491
weighted avg	0.73	0.72	0.72	491

Fig- 2: Details of each emotion's performance

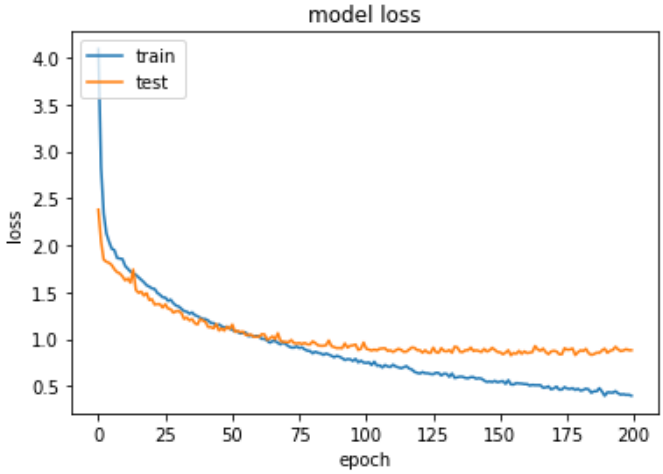

Chart -2: Loss of the model

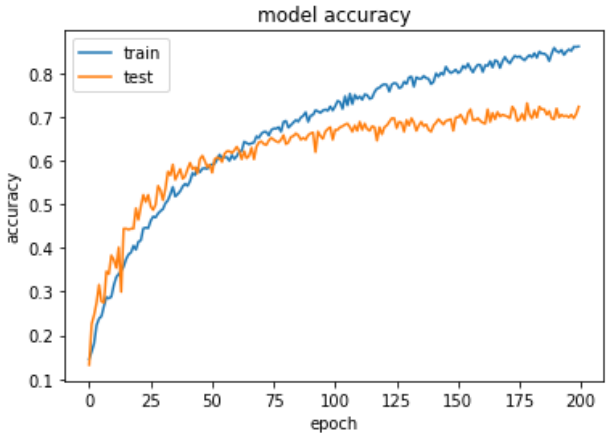

Chart - 3: Accuracy of the model

5. CONCLUSION

In this work, we constructed a speech emotion recognition model based on convolutional neural network (CNN) using the audio file inputs from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The outputs gained show a favorable result with respect to the detection of emotion with accuracy rate of 72.30% and obtained an overall F1 score of 0.72 with the best performance on the angry class (0.81) and the worst performances on the disgust and surprised classes (0.65). For future work, the accuracy and prediction rate of the proposed system can be enhanced by using different datasets so that the system can understand different languages and accents.

REFERENCES

[1] S. Majuran and A. Ramanan, "A feature-driven hierarchical classification approach to emotions in speeches using SVMs," IEEE International Conference on Industrial and Information Systems (ICIIS), 2017, pp. 1-5, doi: 10.1109/ICIINFS.2017.8300369.

[2] Joshi, A.. "Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm." (2013).

[3] Utane, Akshay S. and S. Nalbalwar. "Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine." (2013).

[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[5] Muda, L., Begam, M., and Elamvazuthi, I. : Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. arXiv preprint arXiv:1003.4083 (2010).

[6] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (2010), pp. 807–814.

[7] Rao, K.S., Koolagudi, S.G. & Vempada, R.R. Emotion recognition from speech using global and local prosodic featuresInt. J Speech Technol **16,** 143–160 (2013). https://doi.org/10.1007/s10772-012-9172-2

[8] Srinivas Parthasarathy,Ivan Tashev "Convolutional Neural Network Techniques for Speech Emotion Recognition", 2018 IEEE.

[9] Abhay gupta, Aditya Karmokar, Khadija Mohammed Haneefa, Chennaboina Hemalatha Lakshmi and Shivam Goel "Identification of emotions from speech using Deep Learning". Ins@springer.com

[10] A.Milton, S.Sharmy Roy, S.Tamil Selvi "SVM Scheme for speech emotion recognition using MFCC feature" International Journal of Computer Applications, vol.69-no.9, May 2013.

[11] Xia Mao, Bing Zhang, Yi Luo "Speech emotion Recognition based on a Hybrid of HMM/ANN" Proceedings of the 7th WSEAS International and Communications, Greece, August 24-26, 2007

[12] Jia Liu, Chun Chen, Jiajun Bu, Mingyu You, Jianhua Tao "Speech emotion recognition using an Enhanced Co-training algorithm", pp.1-4244-1017, IEEE 2007.

[13] 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.