International Journal of Future Generation Communication and Networking
Vol.14, No. 1, (2021), pp. 4322 - 4330

Comparative Study of Convolutional Neural Network and Support Vector Machine for Emotion Detection from Speech Signal

Rajdeep Jadhav[1], Shounak Dey[2], Chinmay Mahajan[3], Nilesh Kulkarni[4], Kokila Kasture[5]
Department of E&TC, SKNCOE, SPPU, Pune
[1]rajdeepj1363@gmail.com,
[2]shounakdey@ymail.com
[3]cpmahajan1999@gmail.com,
[4]nileshkulkarni992@gmail.com,
[5]koki.thakur@gmail.com

Abstract

Human emotion recognition plays a key role in developing interpersonal relationship. Depiction of emotions is done by speech, hand and gestures of the body and through facial expressions. Speech Emotion Recognition (SER) has been a topic of research since many years in human machine interface application. Developments of many systems have taken place for solely identifying emotions. This paper explains how a real time system works for detection of a person's emotion from speech. The classifiers used can predict emotions such as Happy, Angry, Fear, Calm and likewise. The databases used for the speech emotion recognition system are Ryerson Audio-Visual Database (RAVDESS) and Toronto emotional speech set (TESS). The features extracted from these datasets are Energy, Zero Crossing Rate, Mel frequency cepstrum coefficient (MFCC). The study is based on comparison of two classification models: Convolutional Neural Network (CNN) and Support Vector Machine (SVM). To enrich the interface with the system, Tkinter Python package is used for building a Graphical User Interface (GUI).

Keywords *— Support Vector Machine (SVM), Convolution Neural Network (CNN), Mel Frequency Cepstrum Coefficient (MFCC), RAVDESS, TESS*

I. INTRODUCTION

Speech can be considered one of the fastest and legitimate methods for holding a communication between machines and humans, compared to other ways. Humans have a natural tendency to use all their different senses for maximum awareness of the received signal. Through this senses itself, one can perceive the emotions from the speech of their communication partner. While it is quite natural for humans, in case of machines it becomes a challenge. Therefore, the objective of emotion recognition system is to apply emotion related knowledge for the betterment of human machine communication [2]. Regardless of the semantic contents, a SER system should be able to distinguish between emotions. Humans can easily performthis task as a natural part of speech communication; to train the machines so that they can conduct it automatically using programmable devices is still an ongoing subject of research. Better SER results were produced by using more complex parameters such as the Mel-frequency cepstral coefficients (MFCCs), Teager Energy Operator (TEO) features, spectral roll-off, spectrograms and glottal waveform features [5]. Emotion recognition from the speech information is of two types: Speaker dependent or Speaker independent.

ISSN: 2233-7853 IJFGCN

A. *Applications of Speech Emotion Recognition*

Speech emotion recognition have a lot of applications in many fields, which includes medical, e-learning, entertainment, law and many more. In medical field, SER can be used to help monitor a patient condition for rehabilitation. It can also be used for finding out client's emotional state just from their speech and provide counselling. In e-learning, presenter can adjust to the learner by observing emotional status in speech. It can be used for monitoring call centre systems for detecting voice signal with and using them as a feedback for the employees. In the field of entertainment mood and emotion of the user and satisfy the needs using the SER. It can also be used in the field of law for lie detection.

B. *Significance of MFC-*

The Mel-Frequency Cepstrum (MFC) is basically conversion of Linear Cosine Transform of the periodogram of the signal into non-linear Mel scale of frequency [4]. That is, MFC is representation of power spectrum of sound. Whereas Mel Frequency Cepstral Coefficients (MFCCs) are the coefficients that constitutes an MFC. MFCCs are derived from non-linear representation of an audio clip. The distinguishing factor between Mel-Frequency Cepstrum and Cepstrum is that in MFC, the frequency band are uniformly spaced on Mel-Scale, this approximately depicts the human auditory system's response more accurately than the linearly spaced frequency bands used in a normal Spectrum.

Mel scale can be obtained by:

$$M(f) = 1125 \ln (1 + f/700) \quad (1)$$

- Step In Calculating MFCCs

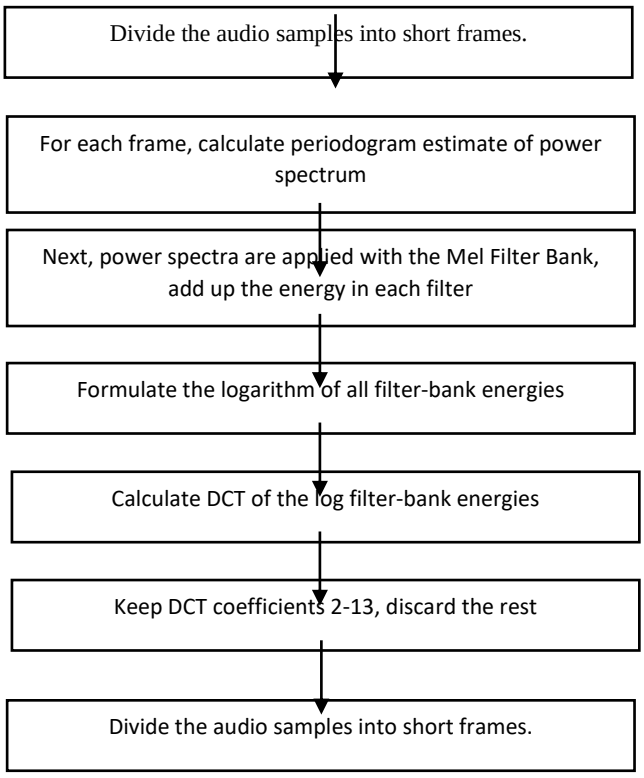


Fig. 1 Flow Chart for steps to calculte MFCC

ISSN: 2233-7853 IJFGCN

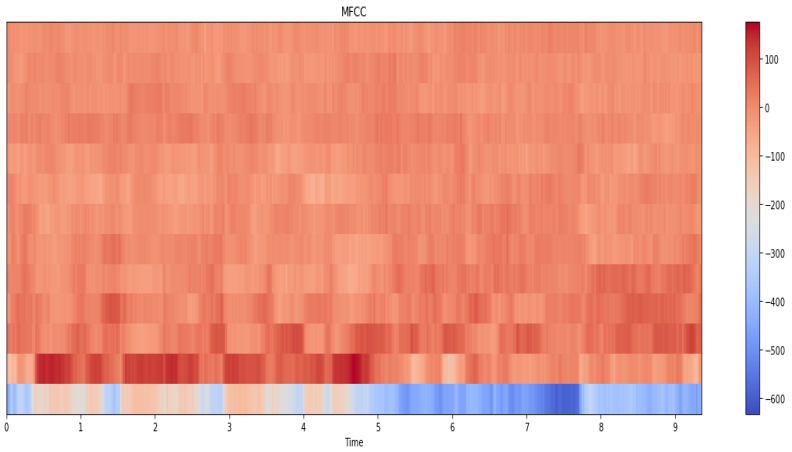


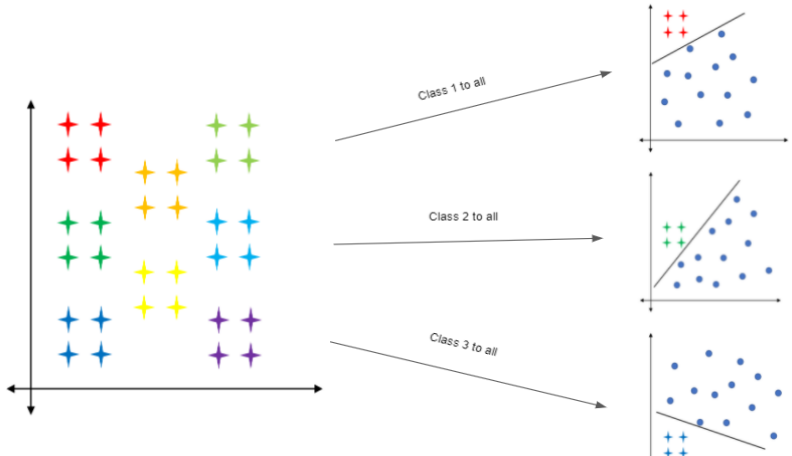
Fig. 2 MFCC feature graph

II. CLASSIFICATION SCHEMES

Initially, the ideology behind the proposed paper was to build a model based on Support Vector Machine (SVM) but later it was decided to take Convolutional Neural Networks (CNN) [3] into the study and compare both the results. Both the techniques are unique on their own. SVM incorporates a Machine Learning approach and a supervised where the user needs to define the model based on the requirements. Whereas CNN leans more towards Deep Learning Approach where a neural network is laid out with a defined number of layers and rest is carried out by the model itself. Generally, there are two kinds of trained models: Speaker Dependent & Speaker Independent. Speaker Dependent model is trained by an individual who shall be using it. Hence, the accuracy for that specific person tends to be on a greater side whereas for other people it lacks considerably. The Speaker Independent model is trained on variation. Regardless of the speaker, the system is expected to generate optimum results.

SVM

Support Vector Machine is a supervised algorithm. It is generally used for Classification as well as Regression based problems. Support Vector Machine technique aims to find out an optimal hyperplane solution in a N-dimension space (where N = Number of Features) that distinctly separates the datapoints [8]. An ideal hyperplane is such that the margin between the line and data points remains maximum. SVM tends to output an accuracy of 83% for a Speaker Independent model, with MFCC, ZCR and RMSE as a feature taken into consideration.



ISSN: 2233-7853 IJ

Fig. 3 SVM One vs All Multiclass Classification

CNN

Convolutional Neural Network is a Deep Learning approach mostly applied in the field of imagery. Deep Learning is a subset of Machine Learning. CNN depicts a feedforward network, which means that it resembles a regularized multilayer perceptron. CNN tends to output an accuracy of 80% for a Speaker Independent model, MFCC, ZCR and RMSE as a feature taken into consideration.

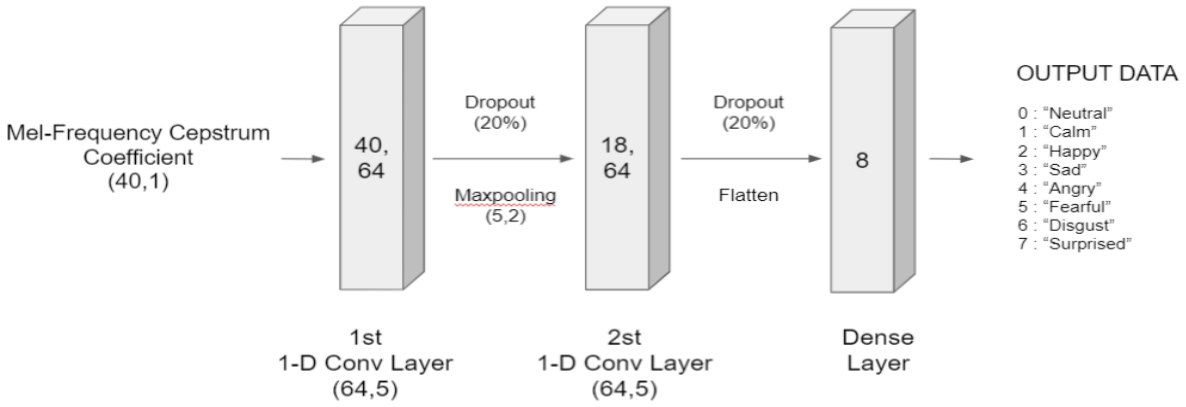


Fig. 4 CNN Model Structure

III. METHODOLOGY

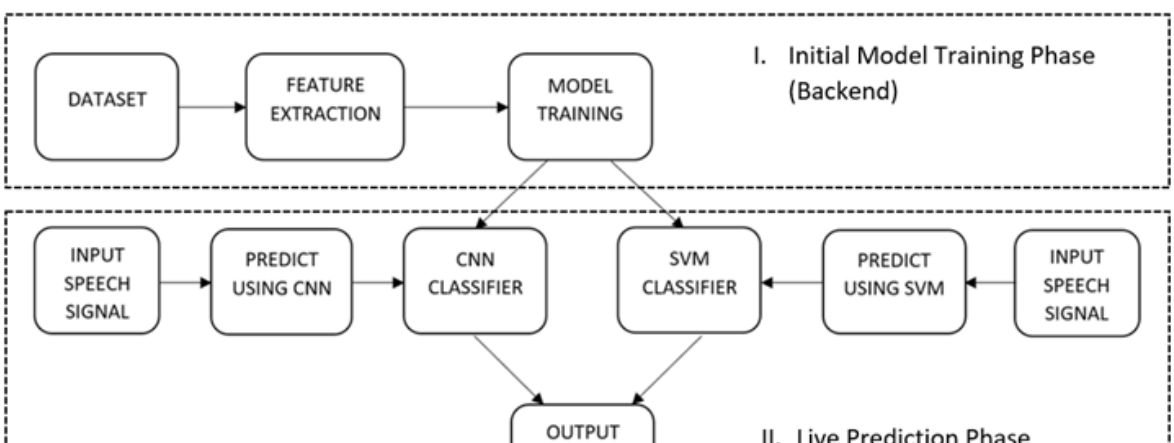


Fig. 4 Block Diagram of Proposed System

Fig.4 depicts the actual working of the proposed system with the help of block diagram. The proposed block diagram is bifurcated into initial Model Training Phase and Live Prediction Phase. Initially, two major datasets which are RAVDESS & TESS were merged together and then used for feature extraction. In feature extraction, the focus is on three essential features: MFCC, Root Mean Square Energy & Zero Crossing Rate. The audio files from which these features were extracted are of 16 bits, 48 kHz & in '.wav' format. Once the features are extracted, they are fed for model training. In this system, two classification models are proposed which are SVM (supervised model) & CNN (unsupervised model). SVM is incorporated using Sklearn API whereas CNN is incorporated using Keras API. The dataset is split in 70:30 ratios, where 70% is for training and 30% belongs for testing purpose. Once the models have been trained, testing can be carried out upon them. For testing one can record a speech sample of 5-10 seconds and use it in '.wav' format for getting prediction results from the respective trained models. The last block that has been shown in the block diagram, 'Out Results to GUI', is a GUI built for the proposed system. The GUI comes with features such as:

a) Browsing a file
b) Plotting the audio signal of the selected file
c) Predicting result using either CNN, SVM or both

IV. RESULTS

After training both the models, CNN produced the most optimum accuracy of 85% whereas SVM produced an accuracy of 83%. The accuracies of the eight emotions that the system has generated for both the models each of are as follow:

Emotions	**Classification Accuracy (%)**
Neutral	**85%**
Calm	**62%**
Happy	**78%**
Sad	**83%**
Angry	**90%**
Fearful	**86%**
Disgust	**82%**
Surprised	**83%**

ISSN: 2233-7853 IJFGCN

Emotions	Classification Accuracy (%)
Neutral	**88%**
Calm	**69%**
Happy	**80%**
Sad	**82%**
Angry	**88%**
Fearful	**86%**
Disgust	**84%**
Surprised	**89%**

Table. 1 Emotion Classification Accuracy rates of SVM

Table. 2 Emotion Classification Accuracy rates of CNN

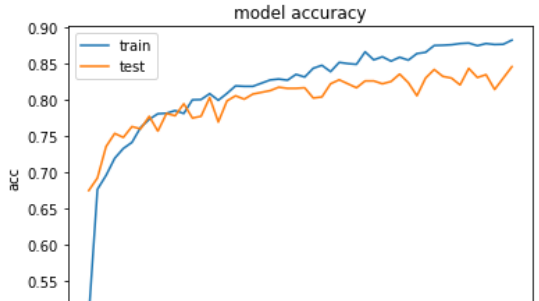
CNN

The model produces an accuracy of 85% with 50 epochs into consideration. For the CNN model of the proposed system, two 1-D Convolutional Layers and a Dense layer have been used. The descriptive model summary can be referred from below table

Layer (type)	Output Shape	Param #
Conv1D	(None, 40, 64)	384
Activation	(None, 40, 64)	0
Dropout	(None, 40, 64)	0
Maxpooling	(None, 18, 64)	0
Conv1D	(None, 18, 64)	20544
Activation	(None, 18, 64)	0
Dropout	(None, 1152)	0
Flatten	(None, 8)	0
Dense	(None, 8)	9224
Activation	(None, 8)	0
Total params: 30, 152 Trainable params: 30, 152 Non-Trainable params: 0		

Table. 3 Model Summary (CNN)

The below figures depict the accuracy and validation loss plot. As it can be inferred from the accuracy plot post 40 epochs the test accuracy changes negligibly. The system has also been tested on 1000 epoch run, and ideal number of epochs was selected as 50. From figure 5, it can be observed that as the number of epochs increases the training data's accuracy increases. Similarly, the test accuracy increases with no overfitting pattern observed. In case of loss plot, both train and test losses decrease as the number of epochs increase. From the pattern of the line plotted in loss plot, one can say that the learning rate is high.



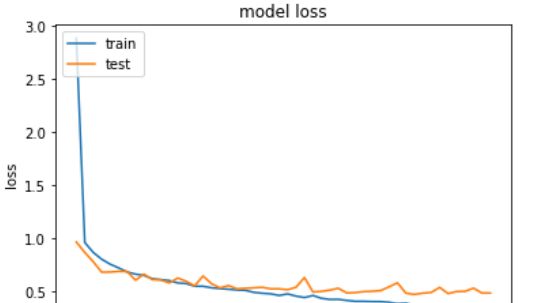


Fig. 5 Model Accuracy (CNN) ***Fig. 6 Model Loss (CNN)***

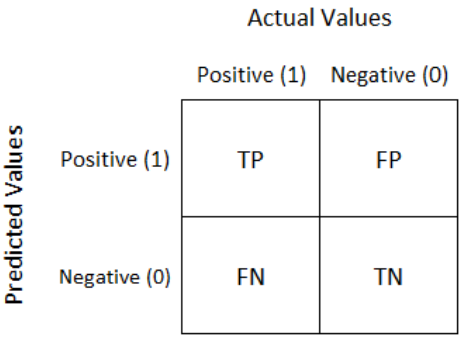
The performance of the CNN model on the set of test data can be described with the help of the Confusion Matrix. Performance parameters such as Accuracy, Precision, Recall & F1-Score can be formulated from the values mentioned in the matrix.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 - Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (5)$$

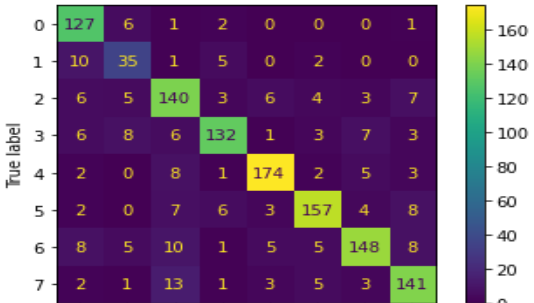



```
[[123   1   2   7   2   0   2   0]
 [  8  36   1   4   0   2   2   0]
 [  2   6 141   5   5   7   3   5]
 [  7   2   5 135   3   4   8   2]
 [  2   0  10   0 177   2   2   2]
 [  0   2   9  11   5 154   5   1]
 [  1   4   4   2  11   2 157   9]
 [  0   0   8   1   4   1   4 151]]
```

Fig. 7 Representation of a Confusion Matrix ***Fig. 8 Confusion Matrix (CNN)***

SVM

This model has been trained using features such as MFCC, RMSE & ZCR. The kernel defined for the model is "polynomial". As mentioned earlier SVM model tends to produce an accuracy of 83%. The confusion matrix is shown below:



ISSN: 2233-7853 IJFGCN

Fig. 9 Confusion Matrix (SVM)

V. CONCLUSION

Speech Emotion Detection system based on comparative analysis of two classification models is illustrated. Both the models can detect amongst the most common emotions such as happy, sad, fear, calm and likewise. In case of SVM classifier, the highest accuracy is obtained for 'angry' emotion which is equal to 90% whereas the least accuracy being for 'calm' emotion which is equal to 62%. In case of CNN classifier, the highest accuracy is obtained for 'surprised' emotion which is equal to 89% whereas the least accuracy being for 'calm' emotion which is equal to 69%. The reason behind the lower accuracy of the 'calm' emotion compared to other emotions is that, one of the databases doesn't contain 'calm' emotion sample. Overall, by looking at the performance parameters, CNN leverages the average accuracy of the model by performing better in all classes and thereby produces better results. With the help of this comparative study one can clearly notice how a deep learning classifier and a machine learning classifier perform, and how important role feature extraction plays in case of SVM.

REFERENCES

[1] B. Zhang, G. Essl, E. M. Provost "Recognizing Emotion from Singing and Speaking Using Shared Models", Computer Science and Engineering, University of Michigan, Ann Arbor, 2015.

[2] A. B. Ingale, D. S. Chaudhari "Speech Emotion Recognition", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, 2012.

[3] R. Khan, O. Sharif "A Literature Review on Emotion Recognition using Various Methods" Global Journal of Computer Science and Technology: F Graphics & vision Volume 17 Issue 1 Version 1.0 Year, 2017

[4] A. P. Reddy, V. Vijayarajan "Extraction of Emotions from Speech-A Survey" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 16, 2017

[5] B. Basharirad, M. Moradhaseli "Speech emotion recognition methods: A literature review", AIP Conference Proceedings 1891, 020105, 2017

[6] P. Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference on Electronic And Mechanical Engineering And Information Technology, 2011.

[7] C. M. Lee, S. S. Narayanan, "Towards detecting emotions in spoken dialogs", IEEE transactions on speech and audio processing, Vol. 13, No. 2, March 2005.

ISSN: 2233-7853 IJFGCN

[8] L. Fu, X. Mao, L. Chen, "Speaker independent emotion recognition based on SVM/HMMs fusion system," ICALIP 2008 - 2008 Int. Conf. Audio, Lang. Image Process. Proc., pp. 61–65, 2008.

[9] M. W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," Proc. Int. Symp. Circuits Syst. 2004, vol. 2, pp. 181–184, 2004.

[10] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 10–21, Jan. 2011.

ISSN: 2233-7853 IJFGCN