|| Volume 5 || Issue 7 || July 2020 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

SPEECH EMOTION RECOGNITION BASED ON MFCC AND CONVOLUTIONAL NEURAL NETWORK

Suchita Sonawane[1], Prof.N.S.Kulkarni[2]

Department of E&TC, Siddhant College of Engineering, Pune, India.[1,2]
sps.s.sonawane@gmail.com,nandakulkarni9@gmail.com

Abstract: Speech based emotion recognition systems please crucial part in audio conferencing. The machine learning speech based emotion recognition techniques has less robustness due to the sensitivity to noise reverberation accent change and language change. Traditional popular Mel Frequency Cepstral Coefficient (MFCC) algorithm for feature extraction performs poor in background non stationary noise. To deal with this problem this paper presents emotion speech emotion recognition based on Deep learning approach. In this multi layers convolutional neural network along with simple K nearest neighbor classifier is applied for the classification of areas emotions such as happy, sad, neutral, disgust and surprise. Extensive experimentation on the real-time database collected from open source social media platform YouTube has shown that combination of MFCC-CNN along with KNN classifier performs better than existing MFCC algorithm.

Keywords: *Speech Emotion Recognition, Deep Learning, Convolutional Neural Network, Mel Frequency Cepstrum Coefficients (MFCC)*

I INTRODUCTION

Speech processing is science of study of speech signal and its processing methods [1]. In the recent year speech emotion recognition has greater demand because of evolution of Technology. Emotion recognition systems can be used in voice best personal assistant and in audio conferencing. Speech emotion recognition plays a vital role in the audio conferencing, call center customer evaluation, speech based personal profiling etc[2].

Various machine learning techniques have been adopted over the past decades for the speech emotion recognition. Traditional emotion recognition systems are dependent on the hand crafted features and highly sensitive to the variability to the speaker, language, reverberation and environmental conditions [3][4].

Emotion recognition has been the subject of research for years. Detection of emotion from facial expressions and biological measurements such as heart beats or skin resistance formed the preliminary framework of research in emotion recognition [5]. More recently, emotion recognition from speech signal has received growing attention. The traditional approach toward this problem was based on the fact that there are relationships between acoustic features and emotion. In other words, emotion is encoded by acoustic and prosodic correlates of speech signals such as speaking rate, intonation, energy, formant frequencies, fundamental frequency (pitch), intensity (loudness), duration (length), and spectral characteristic (timbre) [6].

In this paper, we presented the emotion recognition based on CNN which is used for boosting the discriminative power of generalized MFCC features. It is applied on the in-house database consisting of emotions such as happy, sad, neutral, disgust and surprise.

This paper is organized as follow: Section II gives the detailed about previous work carried out on emotion recognition and algorithms. Section III describes the proposed methodology in details. Further, experimental results and discussions are given in given in section IV. Section V concludes the paper.

II RELATED WORK

Qirong Mao et al. [7] presented the CNN for the salient emotion feature detection. Sonawane Anagha et. al [8] presented speech emotion recognition based on the MFCC and the multiple SVM. They have investigated that non-linear SVM Kernel performs better than the linear SVM. MFCC is popular algorithm for speech recognition tasks. . K. Bhanagle et al. [9] have presented the MFCC for the synthetic speech spoofing detection along with radial basis function SVM classifier. In [10], Hidden Markov Model has been used for the speech emotion recognition with acted and spontaneous speech of German and English language and resulted in 86% recognition rate. Further, Kun Han [11] presented deep learning based extreme learning machine for the speech emotion recognition which extracts the high level features.

Yixiong Pan et al. [12] present support vector machine for

the speech emotion recognition that used energy, pitch, linear predictive spectrum coding (LPCC), mel-frequency spectrum coefficients (MFCC), and mel-energy spectrum dynamic coefficients (MEDC) as the speech features.

III PROPOSED METHODOLOGY

In this work, we have employed the convolutional neural network for the emotion speech feature extraction. The flow diagram of the proposed system is shown in Fig. 1. Generally CNN accept the two dimensional input therefore we have converted the input speech signal in to the MFCC spectrum. Rather than taking the simple spectrogram of the input speech signal, MFCC is employed so that it will represent the signal better than the spectrogram. CNN increases the discriminative and representative power of the MFCC features. It also models the frequency and time domain characteristics of the emotion signal [13][14]. For the classification we have used simple KNN which is very simple to implement and takes lower time for training. The whole system is trained for the different speech signal such as happy, anger, disgust, fear, surprise and neutral.

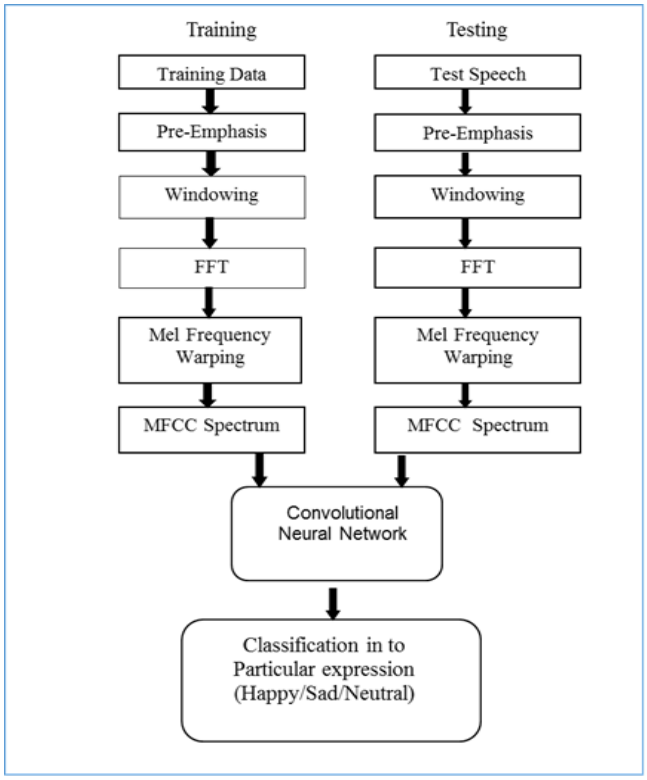

Fig. 1 Flow diagram of the proposed methodology

IV EXPERIMENTAL RESULTS & DISCUSSION

We implemented the system on a system having 6GB RAM with Intel (R) Core(TM) i5-4200U CPU @ 1.60GHz 2.30Ghz processor. The proposed work is implemented using MATLAB R2013b having following system specifications.

Table 1: implementation system details

System	: Personal Laptop with core i5 with 2.64 GHz
Environment (Operating System)	: Windows 8.1
Software	: MATLAB 2013b
Toolbox	: Digital Image Processing, Digital Signal Processing
RAM	: 4 GB

We have created the Neutral emotions sound database from the voice samples of BBC news in male and female voices in British accent. For the Happy and Sad samples, we have collected the samples of male, female, children and infants with the variety of emotion sounds. Out of total samples 70 % samples are selected for the training and 30 % samples are selected for the testing. For the training purpose we have taken 175 samples of happy, sad, surprise, disgust and Neutral emotion sound samples and for testing purpose for every group 75 samples are selected. The database sound is recorded in clean and noise free environment to ensure good results.

Table 2: Database Details

Database	**Trainig samples**	**Testing Samples**	**Total Images**
Happy	175	75	250
Sad	175	75	250
Neutral	175	75	250
Surprise	175	75	250
Disgust	175	75	250
Total data	875	360	**1250**

Recording Time of the all sound is of 4 second with a recording frequency of 8000 samples per second. The Happy sound is shown in fig. 2.

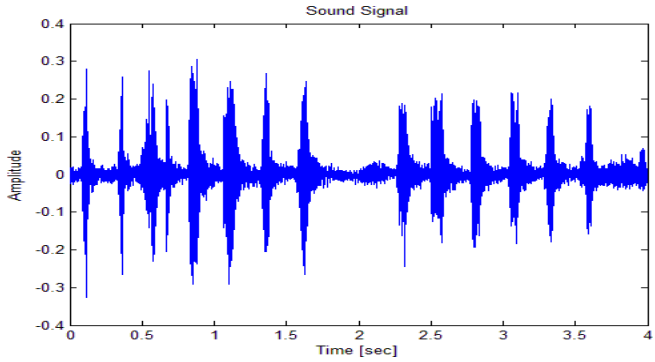

Fig 2 . Original Happy sound signal

In the pre-emphasis step , low frequency components of original happy sound file is removed. Low frequency components are assumed as noise in the sound.

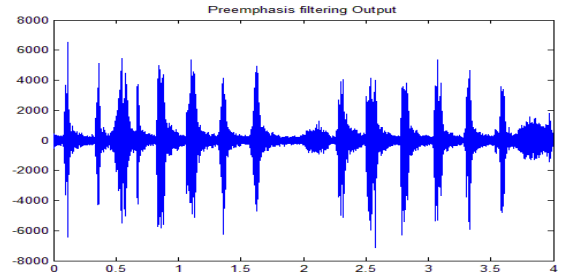

Fig. 3 Pre-emphasis filter output of happy sound

The effect of pre-emphasis is shown in fig.3 which shows significant improvement in the input speech quality.

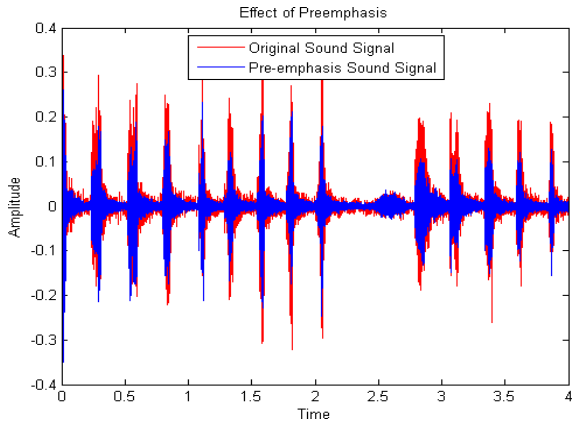

Fig. 4 Effect of pre-emphasis

It becomes difficult and time consuming while accessing all the samples of sound file at a time, therefore it is divided in to the frames. The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec.

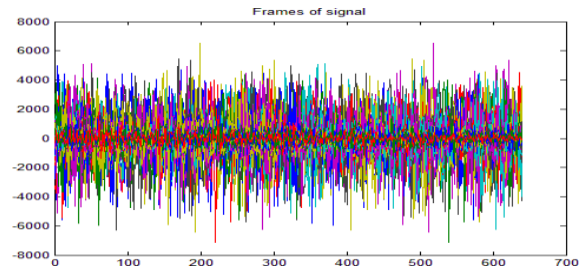

Fig. 5 Frames of Happy signal

This figure 6 shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters.

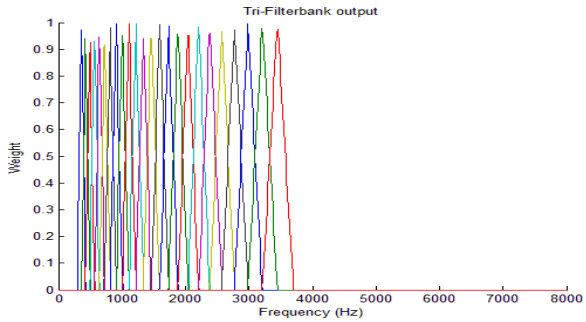

Fig. 6 Triangular filter bank output of happy sound

Therefore, the speaker dependent harmonics are suppressed by taking the lower order cepstral coefficients for further processing. The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time . 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added.

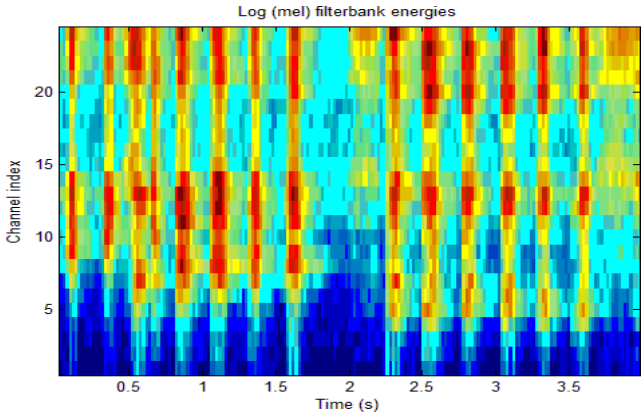

Fig. 7 Log filterbank energies of happy sound

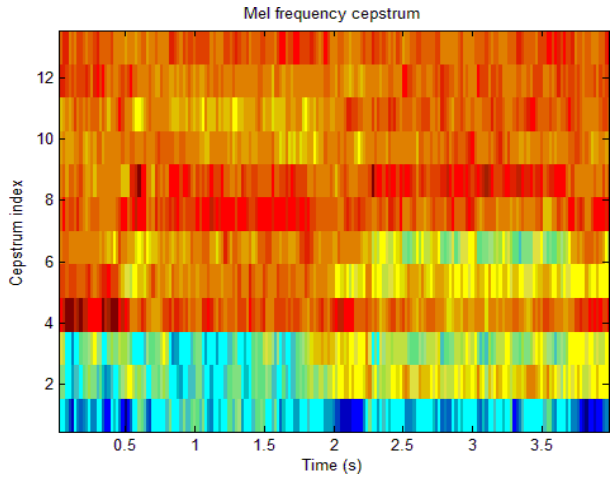

Fig.8 Mel frequency cepstrum coefficients of Happy sound

The output of convolution layer is shown in fig.9, ReLU layer output is shown in fig.10, Max pooling layer output is shown in fig.11.

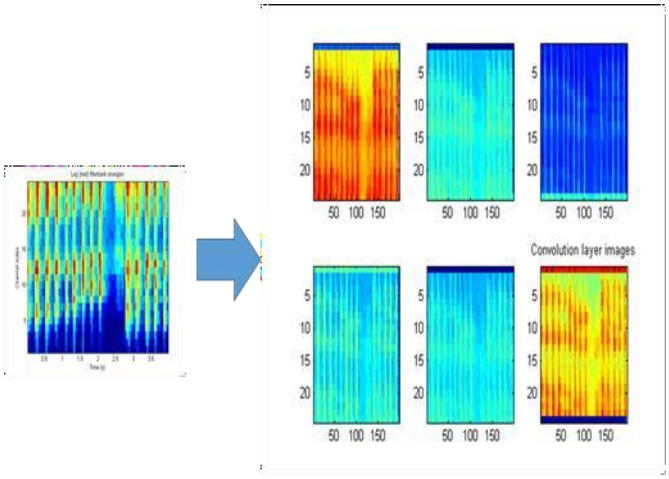

Fig 9. Convolution layer output

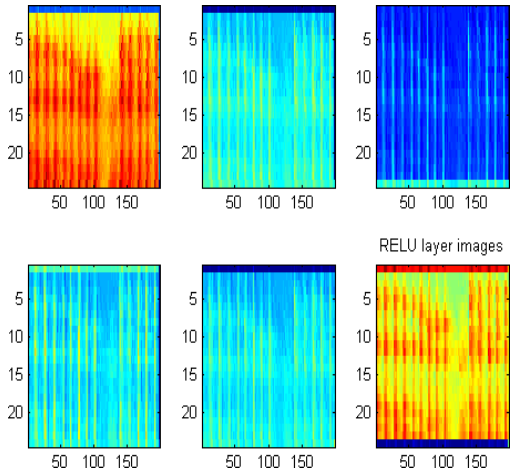

Fig 10. ReLU layer output

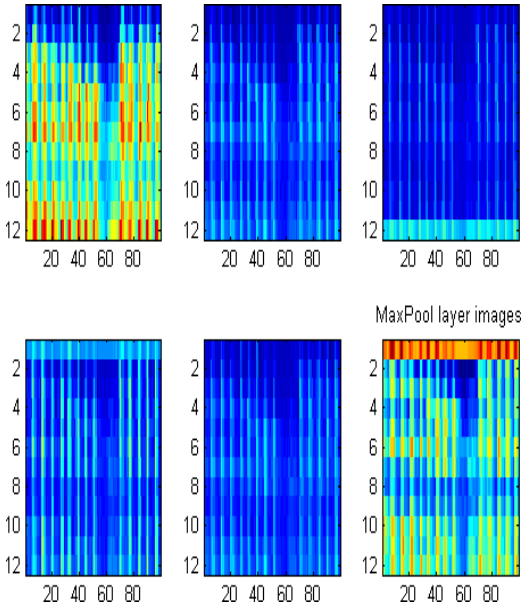

Fig.11 Max-pooling layer output

Visualization of CNN layer 2 is given in figure 12.

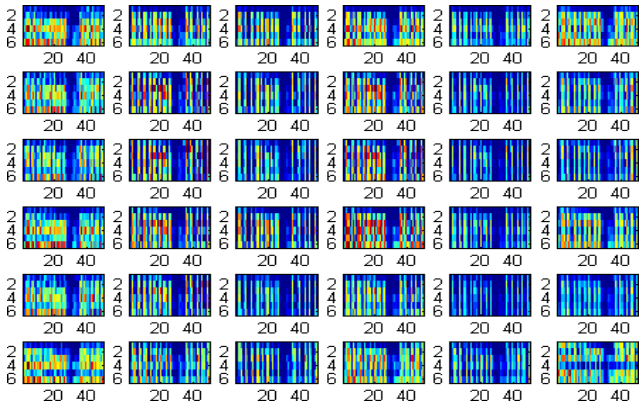

Fig.12 Visualization of CNN layer 2

We have tested the algorithm on the various samples of the happy, neutral, and sad sound database and checked the performance of algorithm on the basis of cross validation accuracy. The performance of classification algorithm is measured on the basis of total cross validation accuracy.

$$Accuracy = \frac{\text{True Positive (TP}}{\text{True Positive (TP)+ True Negative (TN)}}$$

Table 3 shows the experimental results for the MFCC-CNN with the 3x3 filter kernel, 6 filter feature maps, 2x2 maximum pooling and mini-batch gradient learning method. Experimental results shows that CNN performs better than the traditional MFCC features and gives the average accuracy of 88.66 % for the all types of speech emotion.

Table 3. % Recognition Accuracy for MFCC+CNN

Speech Emotion	% Recognition Accuracy			
	MCCC	MFCC+ CNN (Layer 1)	MFCC+ CNN (Layer 2)	MFCC+ CNN (Layer 3)
Happy	57.71	70	68	83.33
Sad	95.45	96.66	98	98.66
Neutral	73.71	82.66	85.53	84
Surprise	75.34	81	84	87
Disgust	76	85.2	83.44	90.32
Average % Recognition Accuracy	75.64	83.10	83.79	**88.66**

V CONCLUSION

In this work, we identified the problem definition as Human emotion recognition using speech processing with MFFC-CNN feature extraction technique and KNN classifier. The

robustness and discriminative power of the MFCC features is increased using CNN. Thus, our system aims to give better performance in noisy environment for larger database. It is suitable for real world application also. In CNN layer increases accuracy also increases. For 3 layered CNN it achieved 88.66% accuracy for emotion recognition.

REFERENCES

1. Benesty, Jacob, M. Mohan Sondhi, and Yiteng Huang, eds. *Springer handbook of speech processing*. Springer, 2007.
2. Gray, Augustine, and John Markel. "Distance measures for speech processing." IEEE Transactions on Acoustics, Speech, and Signal Processing 24, no. 5 (1976): 380-391.
3. Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards, "Artificial intelligence: a modern approach," volume 2. Prentice hall Upper Saddle River, 2003.
4. Lanjewar, Rahul B., and D. S. Chaudhari. "Speech emotion recognition: a review." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2, no. 4 (2013): 68-71.
5. Germine, Laura Thi, and Christine Hooker. "Face emotion recognition is related to individual differences in psychosis-proneness." Psychological medicine (2010).
6. Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion recognition from speech: a review." International journal of speech technology 15, no. 2 (2012): 99-117.
7. Mao, Qirong, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. "Learning salient features for speech emotion recognition using convolutional neural networks." IEEE transactions on multimedia 16, no. 8 (2014): 2203-2213.
8. Sonawane, Anagha, M. U. Inamdar, and Kishor B. Bhangale. "Sound based human emotion recognition using MFCC & multiple SVM." In 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), pp. 1-4. IEEE, 2017.
9. Bhangale, Kishor B., Prashant Titare, Raosaheb Pawar, and Sagar Bhavsar. "Synthetic Speech Spoofing Detection Using MFCC And Radial Basis Function SVM." IOSR Journal of Engineering (IOSRJEN), Vol. 8, Issue 6, pp.55-62, 2018.
10. Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41, no. 4 (2003): 603-623.
11. Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." In Fifteenth annual conference of the international speech communication association. 2014.
12. Pan, Yixiong, Peipei Shen, and Liping Shen. "Speech emotion recognition using support vector machine." International Journal of Smart Home 6, no. 2 (2012): 101-108.
13. Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network." In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5200-5204. IEEE, 2016.
14. Huang, Zhengwei, Ming Dong, Qirong Mao, and Yongzhao Zhan. "Speech emotion recognition using CNN." In Proceedings of the 22nd ACM international conference on Multimedia, pp. 801-804. 2014.