Available online at: http://www.ijmtst.com/vol6issue02.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 02, February 2020

Speech Emotion Recognition Based on CNN Combined with Decision Tree Classifier

S Soundarya[1] | N Arumugam[1]

[1]Department of ECE, National Engineering College, Kovilpatti, Tamil Nadu, India.

To Cite this Article

S Soundarya and N Arumugam, "Speech Emotion Recognition Based on CNN Combined with Decision Tree Classifier", *International Journal for Modern Trends in Science and Technology*, Vol. 06, Issue 02, February 2020, pp.:100-104.

Article Info

Received on 26-January-2020, Revised on 06-February-2020, Accepted on 21-February-2020, Published on 24-February-2020.

ABSTRACT

Speech processing is developed as one of the significant application regions of digital signal processing. Different fields for speech processing work include Speech recognition, Emotion recognition, Speech synthesis, Speech coding, etc. The objective of speech emotion recognition is to extract, characterize and recognize the information about emotions. Extraction of speech features is the first step of Speech Emotion Recognition (SER). Many algorithms are developed by the researchers for feature extraction and classification process. In this work, the Decision tree classifier and Random forest classifier has been used for designing speech emotion recognition (SER) system. Some modifications to the existing technique of CNN for classification are also suggested to improve the speaker recognition accuracy. Performance analysis is done by using the confusion matrix and the accuracy obtained is 70%.

KEYWORDS: *Speech Emotion Recognition, MFCC, Decision tree classifier, CNN, Random forest classifier.*

I. Introduction

People are chatting with one another for thousands of years. In recent times, Human-machine interaction (HMI) has become a growing area of innovation in an industry also as an academic field [1]. Speech is one of the elemental ways of communication known to mankind. A speech signal may be a logical arrangement of sounds. For an efficient and natural HMI, emotion recognition plays an important role [3]. Emotions reflect the psychological state of the person through speech, facial expressions, body postures and gestures and also other physical parameters like blood heat, vital signs, muscle action, etc., The psychological state of the person indirectly affects the speech produced by the person. Speech recognition is the method of translating microphone or telephone recorded acoustic signal to a set of words. Recognized terms can be an end in themselves, as for applications such as commands & control, dataen try, and paper preparation.

The information in the speech signal is truly described by short term amplitude spectrum of the speech waveform, this enables us to extract features supported the short term amplitude spectrum from speech. The fundamental issue of speech recognition is that the speech signal is extremely variable due to completely different speakers, nt speaking rates, contents, and acoustic conditions. Speech emotion recognition is one of the newest challenges in the speech process.

Speech emotion recognition maybe quite

analyzing vocal behavior. Besides human facial expressions speech has proved together of the most promising modalities for the automated recognition of human emotions [2].

Particularly in the field of security systems, a growing interest will be discovered throughout the last year, also emotion Recognition is employed in the call center for classifying calls per emotions. Automatic recognition is usually studied in the sense of distinguishing emotion among some fixed set of classes[1].

The speech process involves 3 main steps i.e. preprocessing, feature extraction and pattern recognition. Feature extraction is the process of extract relevant features from speech signals regarding emotions, MFCC is the common method used for extraction purposes [1]. The classifiers are used to classify emotions such as happy, sad, neutral, anger, surprise, fear, etc. The simple architecture of the speech emotion recognition explained as in Figure.

.

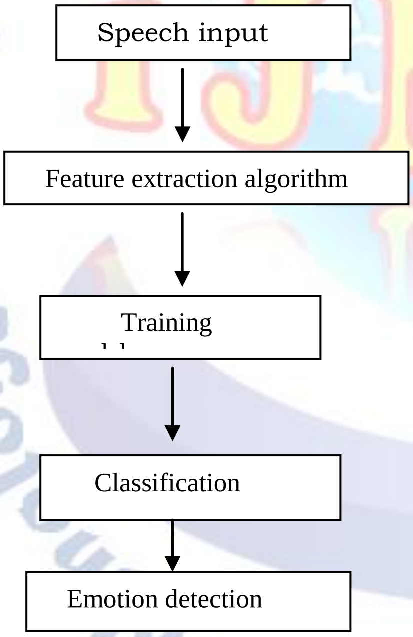


Fig 1.Simple SER System

II. METHODOLOGY

A. RAVDESS DATABASE

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 1440 files. Speech consists of calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with a further neutral expression. The dataset size is 563MB. The dataset description is tabulated in below:

Calm	192
Neutral	96
Happy	192
Sad	192
Angry	192
Fear	192
Disgust	192
Surprised	192
Total	**1440**

Table 1.Dataset description

B. MFCC-FEATURE EXTRACTION

In speech recognition systems, feature extraction and recognition are two essential modules [4]. The first objective of feature extraction is to determine robust and discriminative features within the acoustic data. The popularity module uses the speech features and also the acoustic models to decode the speech input and produces text results with high accuracy [6]. Variety of speech feature extraction methods are proposed, like linear predictive cepstral coefficients (LPCCs), Mel-frequency Cepstral coefficients (MFCCs) and perceptual linear predictive coefficients. Mel Frequency Cepstral Coefficients (MFCC) is usually used as a feature extraction technique. Mel scale may be a scale that relates to the perceived frequency of a tone to the particular measured frequency [5]. The formula to compute the Mel frequency for any given frequency f in Hz is given below:

$$Mel(f) = 2595*\log10(1 + f/700).$$

One of the main advantages of MFCC is a good recognition rate [6]. The Basic concept of MFCC method is shown in the block diagram of Fig. 2

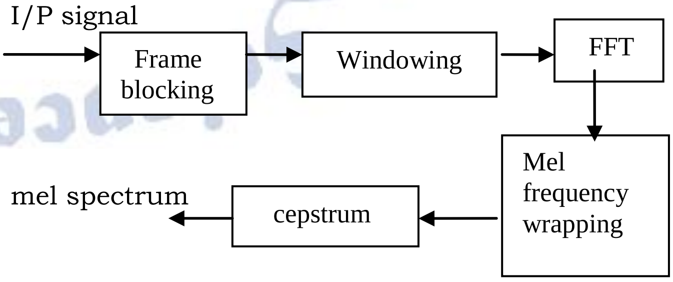


Fig 2.Flow diagram of MFCC

This first pre-emphasis used to boost the quantity of energy within the high frequency[4]. Then apply the hamming windowing technique for creating a section. For skills much energy requires for every band we require convert all time domain into the frequency domain by using DFT [6]. Then it gives to mel filter bank band log to make the feature less sensitive for variation of input. Cepstrum defines inverse DFT of a log of that signal or speech for 12-cepstral coefficients for every frame [3].

C. DECISION TREE CLASSIFIER AND CNN

After the feature extraction, it'll feed into the classifier. There are several types of classifiers wont to classify emotions from a speech like ANN, Gaussian Mixtures Model (GMM), K-nearest neighbors (KNN), Hidden Markov Model (HMM), CNN, and Support Vector Machine (SVM). Each classifier has some merits and demerits[1]. In this proposed method CNN with incorporated with the Decision tree classifier algorithm. Decision Trees are efficient and CNN is accurate[8]. So, we have to design a proposed SER system with a decision tree classifier and CNN for high efficiency and high state of art accuracy.

D. DECISION TREE CLASSIFIER

A decision tree classifier is a tree within which internal nodes are labeled by features. The classifier categorizes an object xi by recursively testing for the weights that the features labeling the inside nodes have in vector xi until a leaf node is reached[12]. The label of this node is then assigned to xi.

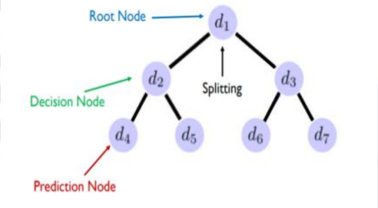


Fig 3.Decision tree classifier

The tree has three types of nodes:

- A root node that has no incoming edges and 0 or more outgoing edges.
- Internal nodes, each of which has exactly one incoming edge and 2 or more additional outgoing edges.
- Terminal nodes, each of which has exactly 1 incoming edge and no outgoing edges

E.CNN

A Convolutional neural network or CNN could be a subset of deep learning and neural networks that have proven very effective in areas of image recognition thus in most cases it's applied to speech processing and analyzes emotions of speech[8]. This network may be a great example of variation for multilayer perceptron for processing and classification. It's a deep learning algorithm during which it takes input as speech and put weights and biases effectively to its objects and finally ready to differentiate speech from one another.

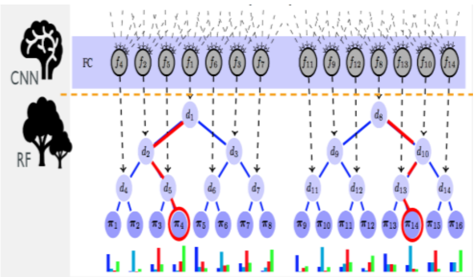


Fig 4.CNN combined with decision tree classifier

F.SPEECH EMOTION RECOGNITION SYSTEM

The speech emotion recognition system consists of three main modules emotional signal, feature extraction, classification, and output show the emotion. The Speech emotion recognition system aims to automatically identify the emotion of persons from the speaker's voice. It's supported the speech signal, extracting the features which contain emotional information from the speaker's speech, and using the appropriate method to recognize the emotion. This proposed method consists of 5 steps, namely 1. Emotional speech input 2.Feature extraction, 3.Training, 4.Classification, 5.Emotion recognition. A typical set of emotions contains 7 emotional states. Primary speech emotions are fear, joy, surprise anger, sadness, and disgust. The proposed speech emotion recognition system:

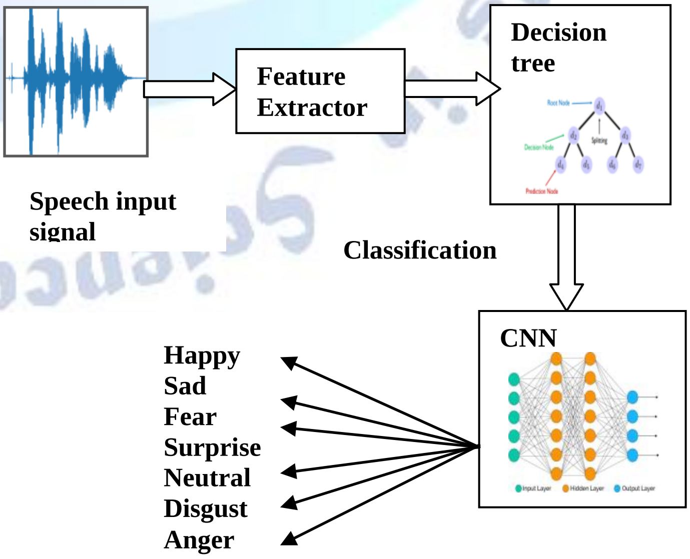


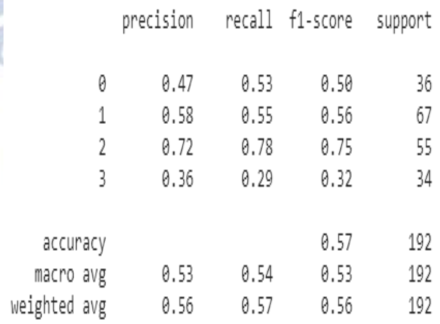
Fig 5.block diagram of proposed SER system

In that proposed methodology, speech input is converted into the spectrum. After that MFCC, extracting the features from the given spectrum and the necessary features are calculated then, it will feed into the decision tree classifier. Here that dataset is split into the classes of the label based on the features. The use of the decision tree is to achieve high efficiency [8]. Then the labeling dataset is dispatched in the convolutional neural network. After the analysis of speech signal CNN classify the different types of emotions.

III. EXPERIMENTAL RESULTS

Here four steps are utilized in the event and training of the feed-forward neural network, namely: (i) collect the training data, (ii) develop the network object,(iii) train the neural network, and (iv) Simulate the network response to check data. When using Decision Tree classifier and CNN classifier, RAVDESS dataset is employed by using 70% training and 30% testing, it gives an accuracy of 70% for 7 emotions that are happy, sad, disgust, angry, surprise, fearful, surprise.

MFCC has been employed as feature extraction methods and therefore the corresponding outputs are given as input to the classifier. Precision, accuracy, and F-score of Decision Tree classifier increases when decreasing the no. of emotions. The obtained results using DT classifier and RF Classifier as shown in figure1.

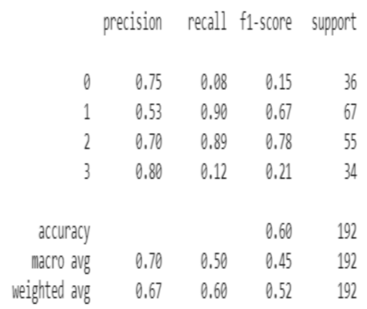


```
              precision    recall  f1-score   support

           0       0.47      0.53      0.50        36
           1       0.58      0.55      0.56        67
           2       0.72      0.78      0.75        55
           3       0.36      0.29      0.32        34

    accuracy                           0.57       192
   macro avg       0.53      0.54      0.53       192
weighted avg       0.56      0.57      0.56       192
```

Fig 6.The output of Decision tree classifier



```
              precision    recall  f1-score   support

           0       0.75      0.08      0.15        36
           1       0.53      0.90      0.67        67
           2       0.70      0.89      0.78        55
           3       0.80      0.12      0.21        34

    accuracy                           0.60       192
   macro avg       0.70      0.50      0.45       192
weighted avg       0.67      0.60      0.52       192
```

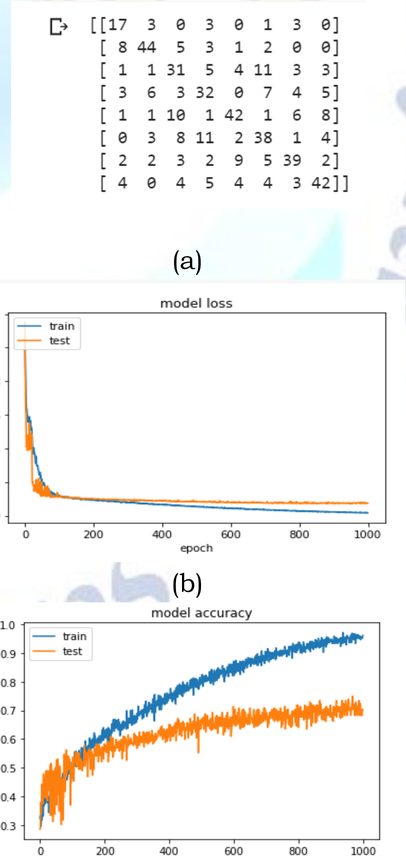
Fig 7.The output of Random forest classifier

When Convolution Neural Network (with DT) is applied to the RAVDESS dataset, the accuracy achieved for 7 emotions: happy, sad, angry, fear, surprise, neutral, disgust is 70.31% for 1000 epochs. This is the best accuracy achieved when compared to the accuracy achieved for other classifiers. Accuracy values are changed when classifiers are changed and it is tabulated in table 2.

Features &classification used	**Accuracy (%)**
MFCC &Decision tree	57%
MFCC &Random forest	60%
MFCC &CNN with decision tree	70.31%

Table 2.comparison of classifiers.

To check the model on different voices for emotion recognition that are completely different from the training and test data, an audio file for predicting different emotions along with the confusion matrix and model loss and model accuracy is shown below. The Accuracy of the proposed method is 70%. It is evident from the below figure that is the highest accuracy rate achieved



```
[[17  3  0  3  0  1  3  0]
 [ 8 44  5  3  1  2  0  0]
 [ 1  1 31  5  4 11  3  3]
 [ 3  6  3 32  0  7  4  5]
 [ 1  1 10  1 42  1  6  8]
 [ 0  3  8 11  2 38  1  4]
 [ 2  2  3  2  9  5 39  2]
 [ 4  0  4  5  4  4  3 42]]
```

(a)

(b)

(c)

Fig 8.Performance analysis of Confusion matrix (a), Model loss(b), Model accuracy(c)

IV. DISCUSSION

The speech recognition system performance is analyzed by recognition accuracy. The analysis accuracy changes for various classifiers. This paper presents the results for emotion recognition of a RAVDESS speech data using CNN with Decision Tree.CNN performs better in recognizing the emotion up to 70% whereas the Decision tree is added. To improve the emotion recognition system, combinations of the given classifications methods can be used. The Accuracy rate of about 70% is achieved. In the future, we will try to improve this system to be a text-independent speaker recognition system.

REFERENCES

[1] Ashish B. Ingale, D.S.Chaudhari, "Speech Emotion Recognition" International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307,Volume 2,Issue-1,March 2012.

[2] Nithya Roopa ,Prabhakaran M,Betty.P, "Speech Emotion Recognition using Deep Learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018.

[3] Akhilesh Watile,Vilas Alagdeve,Swapnil Jain, "Emotion Recognition in Speech by MFCC and SVM", International Journal of Science, Engineering and Technology Research (IJSETR) ISSN: 2278-7798 ,Volume 6, Issue 3, March 2017,.

[4] Vibha Tiwari ,"MFCC and its applications in speaker recognition", International Journal on Emerging Technologies1(1): 19-22(2010)ISSN : 0975-8364,10 February 2010.

[5] A.Milton, S.SharmyRoy, S.Tamilselvi, "SVM Scheme for speech Emotion Recognition using MFCC Feature", International Journal of Computer Applications,ISSN:0975 –8887,Volume 69–No.9, May 2013

[6] Rajiv Chechi,Reetu," Performance Analysis of MFCC And LPCC Techniques In Automatic Speech Recognition", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181,Vol. 2 Issue 9, September – 2013

[7] Chadawan Ittichaichareon,Siwat Suksri and Thaweesak Yingthawornsuk, "Speech Recognition using MFCC" ,International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012

[8] Navya Damodar, Vani ,Anusuya , "Voice Emotion Recognition using CNN and Decision Tree", International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075,Volume-8 Issue-12, October, 2019.

[9] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 Int. Conf. Platf. Technol. Serv., pp. 1–5, 2017.

[10] Tin Lay Nw, Say Wei Foo, Liyanage C. De Silva, "Speech emotion recognition using hidden Markov models", Elsevier Speech Communication journal volume-41 issue 4,pp-603623,November 2003.

[11] Rashmirekha Ram, Hemanta Kumar Palo, Mihir Narayan Mohanty, "Emotion Recognition with Speech for Call Centres using LPC and Spectral Analysis", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-3 Number-3 Issue-11 September-2013.

[12] Li Zheng ,Qiao Li, Hua Ban ,Shuhua Liu, "Speech emotion recognition based on convolution neural network combined with random forest", IEEE Explore- Chinese Control And Decision Conference (CCDC). ISSN: 1948-9447,11 June 2108.

[13] Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam,Vikas Gupta, "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,12 April 2014.

[14] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in Proc. ICASSP. IEEE, 2017, pp. 2227–2231.

[15] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Fifteenth annual conference of the International speech communication association, 2014.