International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Voice Emotion Recognition using CNN and Decision Tree

Navya Damodar, Vani H Y, Anusuya M A

Abstract*: *This paper presents the use of decision tree and CNN as classifier to classify the emotions from the English and Kannada audio data. The performance of CNN and DT are potential for various emotions. Comparative study of the classifiers using various parameters is presented. The performance of CNN has been identified as the best classifier for emotion recognition. Emotions are recognized with 72% and 63% accuracy using CNN and Decision Tree algorithms respectively. MFCC features are extracted from the audio signals and Model is trained, tested and evaluated accordingly by changing the parameters. Speech Emotion Recognition system is useful in psychiatric diagnosis, lie detection, call centre conversations, customer voice review, voice messages.

***.Keywords: Emotion Recognition(ER),* Convolution Neural NetworkCNN), Mel Frequency Co-efficient (MFCC), Decision Tree (DT).**

I. INTRODUCTION

Speech is one of the simple and alternate method for interaction between human and machines. Speech recognition involves identifying the word and sentences uttered by a speaker. With the invention of latest technologies like convolutional neural networks, Long short term memory (LSTM) made speech recognition [1] a possible area as similar to other recognition methods (like image processing, hand written recognition). In affective computing, speech emotion recognition [2] is one of major research field for emotion recognition human computer interaction. The applications of the speech emotion recognition system include the psychiatric diagnosis, smart toys, lie detection, call centre conversations. The different classifiers available are k-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM), Decision Tree (DT), Convolution Neural Network (CNN). This paper throws light on performance of Decision Tree and CNN for emotion recognition.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Navya Damodar*, Information Science and Engg, JSS science and Technological University, Mysuru, India. Email: navya.damodar55@gmail.com
VANI H Y, Information Science and Engg, JSS science and Technological University, Mysuru, India. Email: vanihy@sjce.ac.in
Anusuya M A, Computer Science and Engg, JSS science and Technological University, Mysuru, India. Email : anusuya_ma@sjce.ac.in

To achieve this work in this paper, features are extracted using Mel-frequency cepstrum coefficients (MFCC) and classified using Decision Tree and Convolutional Neural network. The remaining part of the paper is organized as follows: Section 2 discusses the related literature in this field. Database creation and Methodologies used for evaluation is explained in Section 3 and 4. Result analysis is presented in section 5. Conclusions and future enhancements are discussed in section 6.

II. RELATED WORK

A. Speech Emotion Recognition System

The figure shows a simple block diagram for speech emotion recognition system. The speech emotion recognition system comprise three main modules emotional signal, feature extraction, classification and output shows the emotion [2] The simple architecture of the speech emotion recognition illustrated as in Figure 1

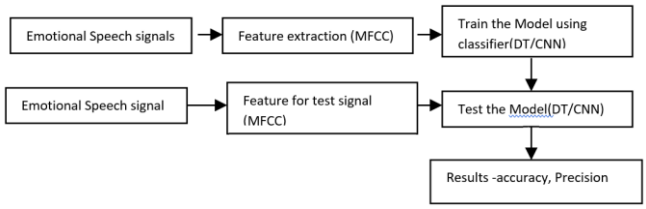


Fig. 1. Simple architecture of speech emotion recognition system

The first step is to extract the features from speech signal uttered by the speaker. The features will become the basic unit for classifying. The emotions like happy', 'neutral', 'sad', 'calm', 'angry', 'disgust', 'fearful', 'surprised' are trained and tested in our system. The signals from the same users are tested and verified with CNN and DT for the required output. The evaluation of the speech emotion recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system [12] [15].

B. Feature Extraction (MFCC) [3]

The goal of feature extraction and processing is to extract relevant features from speech signals with respect to emotions, MFCC is the common method used to extract spectral features of speech signal. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since.

The phase starts with framing, for each frame periodogram estimate of the power spectrum by applying the Mel scale, log for filter bank energies finally DCT is applied to get the discretized real values of a speech signal as features.

C. Decision Tree

A Decision Tree is a simple and common supervised method used for classification. The method used for predicting class or value for target variables. The decision tree solves by using tree representation where each node represents attribute, leaf node as the label and the branch represents a decision rule. It partitions the tree in recursive manner called recursive partitioning. This flowchart-like structure helps in decision making. It's visualization like a flowchart diagram easily mimics the human level thinking. That is why decision trees are easy to understand and interpret. Its training time is faster compared to the neural network algorithm **[4]**.

D. CNN classifier [5]

Convolutional Neural Networks are very similar to ordinary Neural Networks. The CNN are made of neurons that have learnable weights and biases. The neurons on each layer perform dot product. The layers in CNN are a sequence of layers with input conv, Relu pool, and fully connected layer. Through differentiable function each layer transforms one volume of activation to another. The whole network represents the information with a single differentiable score function from converting raw information to a class on the other side. There are three main types of layers to build ConvNet architectures: **Convolutional Layer**, **Pooling Layer**, and **Fully-Connected Layer** (exactly as seen in regular Neural Networks) forming a full ConvNet **architecture**

III. LITERATURE SURVEY

In literature there are two methods for Speech Emotion Recognition.

1) Classifying the signal features received either from Time Domain or Frequency Domain
2) Raw Signals are fed to the auto encoders or CNN [13]model

In the first method several classification methods such as, such as SVM [6], Hidden Markov Method [7], Random Forest [8], DNN [9], or RNN [10] are applied.

The authors Semiye Demircan and Humar kahramanli [14] have performed preprocessing for emotion recognition from speech data. The authors used MFCC for feature extraction and KNN for classification. The accuracy achieved was 50% with training and testing percentage is 80 and 20.

Sawit Kasuriya, Nattapong Kurpukdee and et al.[17] have done comparison of SVM and BSVM algorithms in utterance based recognition of emotion. Acoustic features such as energy, MFCC, PLP, FBANK, pitch and their I^{st} and 2^{nd} derivatives are utilized as frame based attributes. They have selected 4 emotions such as anger, neutral, happiness and sadness in an IEMOCAP dataset. BSVM algorithm shows enhancement of accuracy in few emotions such as sadness and happiness. Accuracy achieved is 58.40% in this method. H.M. Fayek and L. Cavedon [9] have presented Speech Recognition application based on deep learning. Using DNN to identify emotions from speech spectograms of one second frame. They have presented a pipeline which is easier than other complicated systems. Achieved an accuracy of 60.53% for eNTERFACE dataset and 59.7% for SAVEE dataset.

The authors YAWEI MU et.al [10] proposed a new method using distributed Convolution Neural Networks (CNN)[11] to automatically learn affect-salient features from raw spectral information, and then applying Bidirectional Recurrent Neural Network (BRNN) to obtain the temporal information from the output of CNN, but even with this method accuracy achieved is 64.08% .

The authors Jayashree and D J Ravi proposed KNN[18] and NN [19]using Frequency Domain features and Time domain fetaures for Kannada data set and they obtained 70 to 90 for speaker dependent data.

IV. DATASETS

To conduct the simulation the following data sets are considered.

RAVDESS: Ryerson Audio Visual free Database of Emotional Speech and Song totally 7356 files are considered. In this , audio only files are used. The signals are recorded from 24 users out of which 12 male and 12 female[20].

Customized Kannada Dataset: Customized Kannada Dataset has been prepared with the voices of 6 Users, in which 4 Users are females and 2 Users are male. Users have expressed their emotion via speech and then consists of 30 audio files of each user. These 30 audio files contains 3 emotions of Happy, sad and angry emotions. Totally there are 180 audio files maintained for each emotion ten files are taken. These audio files are in wave format.

.

V. METHODOLOGY

The following figure2 illustrates the methodology applied for emotion recognition. The process starts with collecting the audio files which are pre-processed, and features are extracted with MFCC. The obtained features are applied with different classifiers like CNN and DT. The performance is tabulated in the following tables. Table 1 illustrates the results for DT.

and table 2 shows the results obtained from CNN

Training: Testing with 80:20 and 70:30 is considered

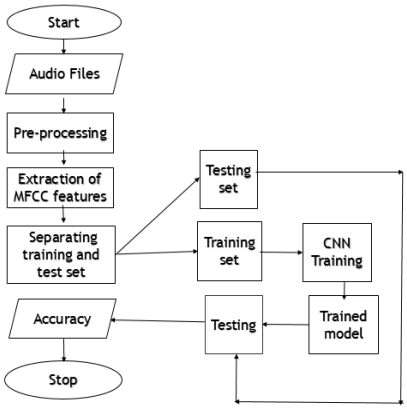


Fig. 2. Flowchart of a Voice Emotion Recognition System

A. ALGORITHM

Algorithm to obtain MFCC features:

- Audio signal is framed into shorter frames for 20ms.
- Periodogram estimate of the power spectrum is found for each and every short frame with an overlap of 20ms.
- To the power spectra, mel-filterbank is applied and in each and every filter, energy is summed.
- Log of all filter bank energies are taken with DCT.

B. DECISION TREE

The first step is to convert these audio samples from wav format into trainable data which is done as follows:

- Using MFCC generate the Melcepst coefficients
- Randomize the data
- Allocate the best feature to the root node of the decision tree
- Making best decisions at each internal node traversing from the root node.
- Routing back to the starting step and repeating the steps until input data is allocated to a class.

C. *CNN*

- The first step is to convert these audio samples from wav format into trainable data.
- Using MFCC generating the Melcepst co-efficients.
- Normalizing the Co-efficients.
- Randomizing the data
- Training the data on a Convolution Neural Network with the input being the coefficients from MFCC predicting the correct emotion of a signal.
- The architecture of the network contains one Convolution Layer of kernel size 8, stride 2 and number of feature maps being 128. This is followed by a pooling layer of size 6.
- There are two fully connected layers each containing 1024 elements. Finally, a softmax layer of size 6 (for 6 emotions) follows responsible for classification.

VI. RESULTS AND DISCUSSION

When using Decision Tree classifier RAVDESS dataset is used by using 70% training and 30% testing, it gives accuracy of 33% for 8 emotions that are 'happy', 'neutral', 'sad', 'calm', 'angry','disgust', 'fearful', 'surprised'. 52% when 5 emotions are considered, that is emotions such as 'happy','calm', 'angry','sad' and 'fearful' and 60% for 3 emotions: happy', 'sad', 'angry'. When train and test ratio is changed to 80% and 20% respectively, it gives accuracy of 38% for 8 emotions. 52% for 5 emotions and 71% for 3 emotions. The obtained results using Decision Tree classified as shown in the table 1.

The accuracy, precision and F-score of Decision Tree classifier increases when decreasing the number of emotions, it proves that Decision Tree classifier is not suitable for multi-class classification problems and is more efficient for binary class problems. Also, 80% training set and 20% testing set is the best when compared to other to 70% and 30% training set and testing test. From this it is clear that DT requires more samples for training the data.

Table 1: Emotion recognition using DT

Dataset	Emotions	Train_Test	Precision	Recall	f-score	Accuracy
RAVDESS	8	70-30	0.34	0.34	0.34	33
	5		0.53	0.53	0.53	52
	3		0.71	0.71	0.71	71
Kannada	3		0.52	0.51	0.51	51
RAVDESS	8	80-20	0.40	0.38	0.39	38
	5		0.53	0.53	0.53	52
	3		0.72	0.71	0.71	71
Kannada	3		0.64	0.64	0.64	63.8

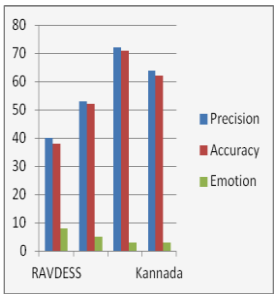


Fig 3: Emotion recognition using DT 70:30

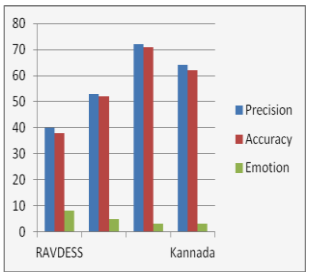


Fig 4: Emotion recognition using DT 80:20

CNN is applied to RAVDESS dataset, with train and test data set with a ratio 70 and 30 respectively. Accuracy is checked for different epochs and the best result is selected. Also, epochs and accuracy values are tabulated. Best accuracy achieved for 5 emotions: 'calm', 'happy', 'sad', 'angry', 'fearful' is 69.42% when epochs are 4000. When changing the training and testing data set proportions to 80 and 20, accuracy 72.40% is achieved for 3500 epochs. This is the best accuracy achieved when compared to the accuracy achieved for other epochs.

Accuracy is checked by varying epoch values. Epoch and Accuracy values are tabulated in the table 3. Accuracy value changes when epochs are increased but at some particular value of epoch, accuracy starts decreasing when epochs are increased When Convolution Neural Network is applied to Customized Kannada dataset, accuracy achieved for 3 emotions: 'happy', 'sad', 'angry' is 63.89% for 2000 epochs. This is the best accuracy achieved when compared to the accuracy achieved for other epochs. Accuracy values are changed when epochs are changed and it is tabulated in the table 2.

Table 2: Emotion recognition using CNN

Dataset	Train-Test	Epochs	Accuracy
RAVDESS	70-30	1000	54
		1500	63.5
		2000	66
		2500	68
		3000	68.73
		3500	69.42
		4000	69.42
	80-20	1000	61
		1500	65
		2000	67
		2500	70
		3000	71.88
		3500	72
		4000	68
Kannada	80-20	1500	55
		2000	63.89
		2500	61
		3000	55

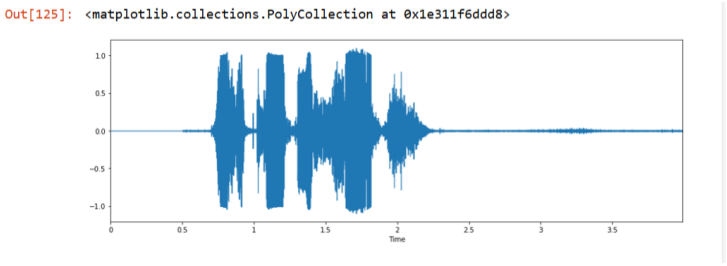
Varying epoch values epochs and Accuracy values are tabulated in the table 3.4. Accuracy changes when epochs are increased but at some particular value of epoch, accuracy starts decreasing when epochs are increased.

From the test data actual and predicted values are generated. Actual and predicted values of 10 files are shown. It is shown in table 3 that, emotions of 1st, 6th and 7th voice samples are predicted incorrectly, however, emotions of 2nd, 3rd, 4th, 5th, 8th and 10th voice sample is predicted correctly. So, among 10 audio samples, emotions of 3 audio samples are predicted incorrectly and emotions of 7 audio samples are predicted correctly.

A. *Testing the Model with different audio file:*

In order to test the model on different voices for emotion recognition that are completely different from the training and test data, audio file for predicting different emotion along with equivalent image is as shown below. The audio signal contains a male voice who says"I hate this coffee" in an angry tone.

Plotting the wave file:



Predicting the class of the new wave file:

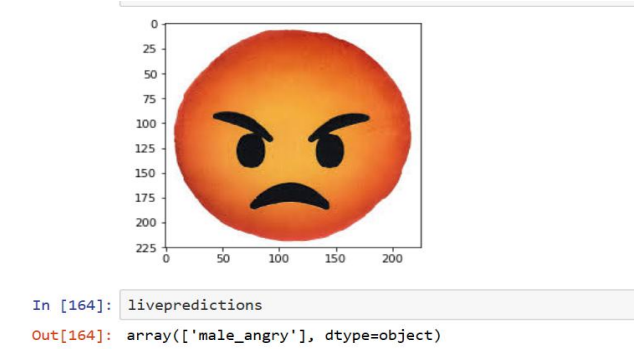


Fig 5: Emotion recognition using CNN

Table 3: Emotion recognition using CNN

Out[149]:

	actualvalues	predictedvalues
0	male_angry	male_sad
1	female_happy	female_happy
2	female_happy	female_happy
3	female_calm	female_calm
4	female_calm	female_calm
5	female_angry	female_fearful
6	female_fearful	female_happy
7	male_calm	male_calm
8	male_sad	male_sad
9	female_sad	female_sad

CNN classifier predicting the emotion of the audio file as angry which is correct.

VII. CONCLUSIONS

This paper presents the results for emotion recognition for a speech data using CNN and Decision Tree. Among two classifiers, CNN performs better in recognizing the emotion upto 72% where as Decision tree gives 38% and 52 % accuracy for less number of emotions.
Hence it shows that CNN is best for emotion classification of speech data. However, Decision Tree classifier accuracy is improved when class labels are reduced. It shows that Decision Tree classifiers are best for binary class problems than multi-class problems. Also, model would perform better if there's more data available for training.

VIII. FUTURE WORK

The system can be further applied to total mood swings and depression identification. These can also be applied for verifying RNN and other clustering techniques.

REFERENCES

1. Ossama Abdel-Hamid, Abdel-rahman Mohamed, Convolutional Neural Networks for Speech Recognition, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 10, OCTOBER 2014
2. Ramakrishnan, S. & El Emary, I.M.M Speech emotion recognition approaches in human computer interaction. Telecommun Syst (2013) 52: 1467. https://doi.org/10.1007/s11235-011-9624-z.
3. Vani H Y and Anusuya M A Isolated speech recognition using FCM and K-means Technique 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), DOI: 10.1109/ERECT.2015.7499040
4. Enes Yunc, Huseyin Hacıhabibo ,Automatic Speech Emotion Recognition using Auditory Models with Binary Decision Tree and SVM, 2014 22nd International Conference on Pattern Recognition, DOI 10.1109/ICPR.2014.143
5. Yafeng Niu , Dongsheng Zou etal,A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks, https://arxiv.org › pdf
6. Dahake, Prajakta P., Kailash Shaw, and P. Malathi. "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine". Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016. DOI: 10.1109/ICACDOT.2016.7877753
7. Tin LayNwe Say WeiFoo Speech emotion recognition using hidden Markov models, Speech Communication, Volume, November 2003, Pages 603-623
8. Vocal-based emotion recognition using random forests and decision tree February 2017 International Journal of Speech Technology DOI: 10.1007/s10772-017-9396-2
9. H.M. Fayek and L. Cavedon "Towards Real-time Speech Emotion Recognition using Deep Neural Networks", 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS).
10. YAWEI MU, LUIS A et al, Speech Emotion Recognition Using ConvolutionalRecurrent Neural Networks with Attention Model.
11. Huang, Zhengwei, et al. "Speech emotion recognition using CNN", Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.
12. Zheng, W. Q., J. S. Yu, and Y. X. Zou.An experimental study of speech emotion recognition based on deep convolutional neural networks.Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE, 2015
13. Eduard FRANŢ, I1, Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots, ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY Volume 20, Number 3, 2017, 222–240
14. Semiye Demircan and Humar kahramanli, "Feature Extraction from Speech data for Emotion Recognition", Journal of Advances in Computer Networks, Vol. 2, No. 1, March 2014.
15. Kim, Yelin, and Emily Mower Provost. "Emotion classification via utterance-level dynamics" A pattern-based approach to characterizing affective expressions. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference
16. Aishwarya Murarka, Kajal Shivarkar and et al."Sentiment Analysis of Speech", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 11,November 2017.
17. Sawit Kasuriya, Nattapong Kurpukdee and et al,"A Study of Support Vector Machines for Emotional Speech Recognition", 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES).
18. J Pallavi,Geethashree,D.J Ravi, EMOTIONAL ANALYSIS AND EVALUATION OF KANNADA SPEECH DATABASE ,IAEME Publication 2014
19. Speaker Dependent Emotion Recognition from Speech for Kannada language NCRACES – 2019 (Volume 7, Issue 10)
20. https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-auo

AUTHORS PROFILE

Navya Damodar
MTech Data Science
Department of Information Science and Engg
JSS Science and Technological University Mysuru
Email: navya.damodar55@gmail.com

Vani H Y
Information Science and Engg, JSS science and Technological University, Mysuru, India. Email: vanihy@sjce.ac.in

Anusuya M A
Computer Science and Engg, JSS science and Technological University, Mysuru, India. Email :anusuya_ma@sjce.ac.in