International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Comparative Analysis of Emotion Recognition System

Chinmay Thakare[1], Neetesh Kumar Chaurasia[2], Darshan Rathod[3], Gargi Joshi[4], Santwana Gudadhe[5]

[1,2,3,4,5]*Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India*

Abstract - *Speech carries a lot more deal of information for emotion recognition. This paper discussed the implementation of emotion recognition system on SAVEE dataset. Emotion recognition system deals with speech input. Various features like MFCC, contrast, mel frequency, tonnetz and chroma are considered from input dataset. This paper presents a comparative analysis of speech emotion recognition system with CNN, RFC, XGBoost and SVM classifier. In all CNN model give highest accuracy for MFCC feature. The study is aimed at exploring the dependencies that exist with the human emotional state. We trained and tested the model with a CNN, RFC, XGBoost and SVM classifier and processed the audio clips to characterize them into 7 significant emotions.*

Keywords: *SAVEE, CNN, SVM, Speech Emotion Recognition, SER, MFCC*

1. INTRODUCTION

Emotion recognition using human speech input, is one of the most trending fields in speech analysis as well as emotion recognition. Speech signal is information rich and and non traditional approach to the problem. The speech signal is just as important as graphical in emotion recognition in nature. Features in speech are perceived scientifically and by humans quite differently. Several features like loudness ,accent, language are user dependent , contributes to how the signal is to be perceived. Emotion recognition is useful in breaching the gap in human machine interaction.

Recently a lot of work has been done in this area, using different techniques and features for the same. Several applications of the emotion recognition system can be taken into account, such as virtual personal assistants, automated call centres, automated feedback systems and other human machine interaction system(HCI) etc. previously, a lot of graphical methods have been used in emotion recognition, major work has been done in the facial emotion recognition system(FER). Also very few advancements have been made in the speech based system or the hybrid system.

A lot of literature on different ways to tackle the problem have been discussed. Features both in time and frequency domain have been looked into. Few features have found credible, though as the field is still been explored majority of the features used are similar to ones used for voice recognition. As there are a lot of features available to choose from a few have been discussed in the following been discussed here.

Automation of this process is contributed majorly by machine learning and signal processing. One of the most important parts of the system just like any machine learning is feature selection. The speech file is just sequence of sound wave, thus features like amplitude, frequency, power, etc. There are a lot of features to select from, in those mel frequency and mel-frequency cepstral coefficients have proven to be very useful. As both of these features are designed according to the human frequency perception.

Thus, most researchers prefer to use combining feature set that is composed of many kinds of features containing more emotional information[1]. While that been said, this increases the data to be processed, which is likely to cause over fitting and increase time required for result generation.

The categorization of emotions can be done into many classes, but so as to keep the problem approachable only the basic emotions are to be handled. Most popular is the Ekman model[2], with six emotions. The model used here has seven emotions classes, enlisted as happy, sad, angry, scared, surprised, neutral and disgust.

Speech processing has been an area of interest for the last four decades, but the last decade has witnessed significant progress in this area of research. This progress has been possible mainly due to the recent advances made in the powerful and efficient speech processing techniques such as vector quantization, hidden Markov Model etc.

In this paper, the first section discussed the introduction of emotion recognition system. Second section discovered the speech based emotion emotion recognition system. Followed by dataset used for analysis and last section discussed about result and conclusion.

1.1. THE EMOTION RECOGNITION SYSTEM

As Speech based emotion recognition has been in light for a few years, a lot of innovative methods have been developed. Even so, basic model has not changed much over time.

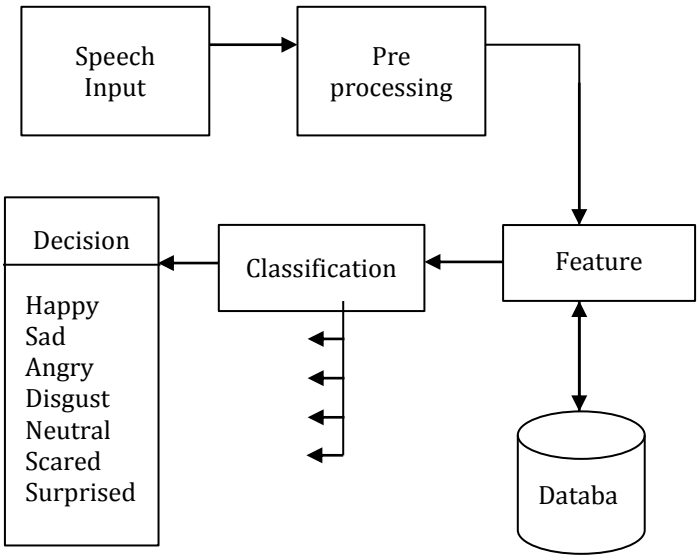

Figure 1: Speech based emotion recognition gives the fundamental flow of the system.

1. Speech Input

The audio file is to be taken in as input for the system to be processed. Standard encoding formats like .wav ,.mp3 ,ac3, and more for the audio file are expected to be supported by the system.

Audio Recording is preferred in '.wav' format for processing and best results. The recording can be from any device from the system user. If the audio file format is in other than ".wav"form, then it is converted to ".wav".

2. Pre-Processing:

The data recording has to be further processed before it can be worked on. If the features are extracted without cleaning and pre-processing of the accuracy will be hampered a lot causing the overall accuracy of the system to drop very low, and at times generate inconsistent results. This can also lead to recognition of non existent or incorrect pattern recognition by the classifier.

This module helps to suppress background noise from the audio clip to help increase accuracy. At the same time, in case of multiple speakers, it works out if there are more than one speaker in the audio clips it separates the data of the speakers. Silent part of the clips do not contribute to the final result but increases unnecessary computation, eliminating the silent parts of audio clip helps reduce the data to be computed. Further, transformation like, STFT, log and cosine have been proved to be accuracy boosters.

3. Feature Extraction

Speech emotion feature is the basis for speech emotion recognition. Extraction accuracy of speech emotion features in the original speech emotion samples can influence the final recognition rate of speech emotions directly.

Feature selection is very important and contributes to major difference in accuracy of the system. In order to get maximum accuracy, feature with highest accuracies have to be extracted and used for processing.

Features like MFCC, contrast, mel frequencies, tonnetz and chroma have been taken into account

1) MFCC:

Spectral property of voice signals is mostly represented by Mel-frequency cepstrum coefficient. For each frame, the Fourier transform and the energy spectrum are estimated and mapped into the Mel-frequency scale. The discrete cosine transform (DCT) of the Mel log energies is estimated.

2) Mel filterbank

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear. [3]

The formula for converting from frequency to Mel scale is:

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

3) Contrast

contrast is estimated by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy).[4]

4) Tonnetz

The Tonal Centroids (or Tonnetz) contain harmonic content of a given audio signal.

5) Chroma

The entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

Once the features are extracted , the values are stored in file(.csv), for future use.

4. Classification

Model represents the Machine Learning algorithm, many classifiers are available at hand after so many years of work in machine learning. Classifier taken in account are CNN, Random Forest Classifier, Support Vector Machine and XGBoost.

As each classifier has a different way of learning , the accuracies projected are different. With fine tuning the classifiers to gain maximum accuracy for each of the classifiers. Each feature is used as input for the classifiers and then fine tuned.

Train and testing to be done for models to learn using different standard datasets. These models once trained is stored and later it can be loaded, and used to predict the results.

5. Decision

The output of the system is detected emotion class of the seven emotions of Angry, happy, sad, neutral, scared, disgust, and fear.

1.2. DATASET

The dataset used in the experimentation here is SAVEE[5]. Surrey Audio-Visual Expressed Emotion (SAVEE) has a total of 480 audio sample in .wav format.

The dataset has been recorded by 4 actors (male british), with 7 emotions used to classify the audio clips. Every emotion has 15 audio clips for every actor, with the exception of neutral class with 30 clips. Each clip is of length ranging from 2 to 3 seconds.

1.3. Implementation

The SAVEE dataset's file structure is a hierarchy starting with folder Actor's name with seven folders for each emotion, and audio files in those folders. This has been changed to seven folders one for each emotion with a four folder one for each actor containing audio clips.

Dataset is in standard format, no background noise is observed. Even so butterworth filter has been used for noise elimination and the blank(silent) parts of the clip are eliminated for performance.

Five features are extracted from the cleaned data for each audio clip. The features are MFCC, Mel Filter bank(mel), Contrast, Tonnetz and Chroma. The extracted data is stored in form of .csv files, for training multiple classifiers, this saves the time required for feature extraction every time a classifier is to be trained.

Total of four classifiers are trained and tested on the dataset on the data namely, Random Forest Classifier, ConvNet(CNN), XGBoost and SVM.

Convolutional Neural Network is implemented with design of 5 layer, experiment with the number neuron in each layer was done. With SVM kernel function is played, all three kernel functions present in sklearn are experimented with.

In case Random Forest number of estimators are varied, and in XGBoost number of estimators, learning rate and random_state are experimented with. A lot of fine tuning is done in order to get the most accurate results. Results are discussed in the next section.

2. RESULTS AND DISCUSSIONS

After the features were extracted from the audio files, training and testing was done with ratio of 80:20.

In this paper, four different classifiers were used; namely Convolutional neural network(CNN), Random Forest classifier(RFC), XGBoost (implementation of gradient boosting) and Support Vector Machine(SVM).

CNN model was built using KERAS sequential model[6]. In this 5 layers were used. Activation layers were of relu, softmax. The layers included 256 neurons in first layer and 128 neurons each in next three layers. Last layer had seven filters corresponding to seven emotions. Maxpool was taken to be 4. The number of epochs was decided on the basis of model loss vs accuracy chart. Initially 1000 epochs were considered, but later changed to 400 epochs as the model showed the highest accuracy with this count.

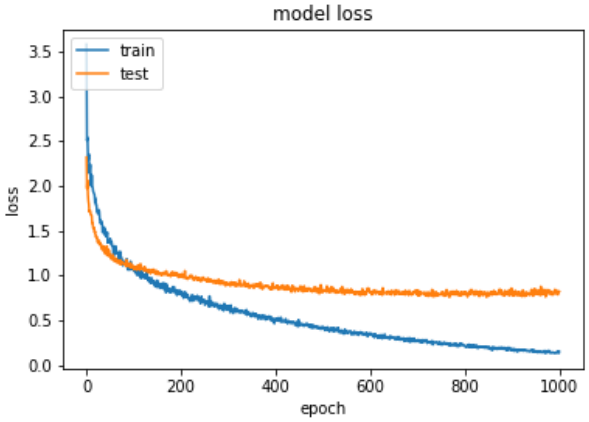

Chart-1: Model loss vs epochs with 1000 epochs.

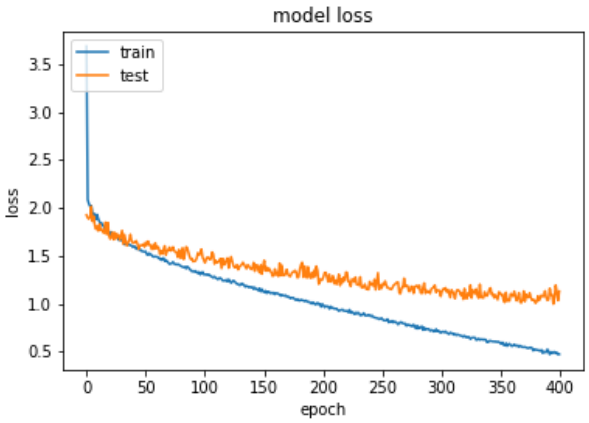

Chart-2: Model loss vs epochs with 400 epochs.

With this model, we got an accuracy of 72.92%, 27.08%, 53.12%, 32.29%, 36.46% for mfcc, tonnetz, mel, chroma and contrast respectively. MFCC gave the highest accuracy in CNN model.

For random forest classification[7], entropy criterion was chosen as it gave better accuracy. The number of n_estimator had to be changed according to the GridSearchCV cross validation with cv equal to 10. With n_estimator values in brackets the accuracies are 67.96(35), 29.16(81), 70.62(675), 38.54(80) and 52.08(70) for mfcc, tonnetz, mel, chroma and contrast respectively. In this classification technique, mel features had the highest accuracy.

In XGBoost classifier[8], the accuracies were similar for features compared to random forest classifier but lesser in values. The accuracies are 62.5, 21.87, 64.58, 36.45, 48.95 for mfcc, tonnetz, mel, chroma and contrast respectively. In this classifier also, the mel features had the highest accuracy. The results for CNN, RFC and XGBoost classifier is shown in chart-1 and the corresponding accuracy values in table-1.

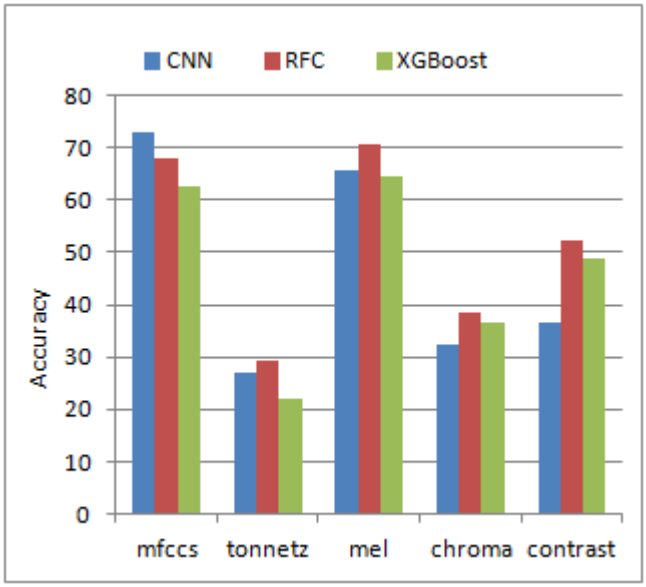

Chart-3: Accuracy comparison for CCN, RFC and XG Boost

Table-1: Table for values of accuracy in CCN, RFC and XGBoost

classifier\ feature	MFCC	Tonnetz	Mel Filter bank	Chroma	Contrast
CNN	72.92	27.08	65.63	32.29	36.46
RFC	67.96	29.16	70.62	38.54	52.08
XGBoost	62.5	21.87	64.58	36.45	48.95

In Support Vector Machine classifier[9], kernels of linear, rbf and poly were used taking regularization parameter c equal to 1. For linear and kernel, mfcc had the highest accuracy of 68.75%. For rbf kernel, contrast gave the most accuracy for 49.47% for c=7. Rbf kernel did not show response to other features. The results for SVM classifier with linear, rbf and poly kernel is shown in chart-4 and the corresponding accuracy values in table-2.

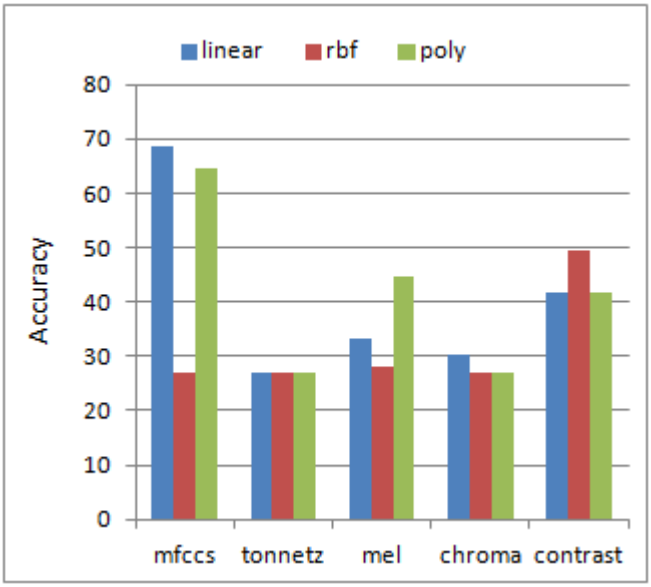

Chart-4: Accuracy comparison for SVM classifier using linear, rbf and poly kernel.

Table-2: Table for values of accuracy in SVM classifier using linear, rbf and poly kernel.

	MFCC	Tonnetz	Mel Filter bank	Chroma	Contrast
SVM kernel= poly	64.58	27.08	44.79	27.08	41.66
SVM kernel= linear	68.75	27.08	33.3	30.2	41.66
SVM kernel= rbf	27.08	27.08	28.12	27.08	49.47

3. CONCLUSION

In this paper, multiple emotion recognition features were presented. This paper presents a comparative analysis of emotion recognition system using CNN, RFC, XGBoost and SVM classifier on SAVEE dataset. The emotions were

classified into happy, sad, fear, neutral, angry, disgust and surprise. CNN model gives the highest accuracy with 72.92% for mfcc features. When trained with Random Forest Classifier 70.62% accuracy is achieved for mel features. XGBoost had an accuracy of 64.58% with mel features and SVMs accuracy was 68.75% for linear model with mfcc features. It is observed that MFCC and mel features give highest accuracy and can be used for emotion recognition depending on the machine learning model used.

REFERENCES

[1] Koolagudi, S.G. & Rao, K.S. Int J Speech Technol (2012) 15: 99. https://doi.org/10.1007/s10772-011-9125-1 (Last referred on 5-Nov-2019)

[2] Scherer, K. R. (2005). What are emotions? And how can they be measured? Social Science Information, 44(4), 695–729. https://doi.org/10.1177/0539018405058216 (Last referred on 5-Nov-2019)

[3] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs (Last referred on 5-Nov-2019)

[4] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenbergk , Oriol Nieto, "PROC. OF THE 14th PYTHON IN SCIENCE CONF. (SCIPY 2015)"

[5] S. Haq and P. J. B. Jackson, "Machine Audition: Principles, Algorithms and Systems," Hershey PA, 2010, pp. 398-423.

[6] Chollet, F., & others. (2015). Keras. https://keras.io. (Last referred on 5-Nov-2019)

[7] Andy Liaw, & Matthew Wiener (2002). Classification and Regression by randomForestR News, 2(3), 18-22.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. ISBN: 978-1-4503-4232-2

[9] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and Their Applications, 13(4), 18–28. doi:10.1109/5254.708428 (Last referred on 5-Nov-2019)