IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Speech Emotion Recognition Using Support Vector Machine

K Mahesh Raj*1, Dr.Suwarna Gothane *2, Chandrashekar Aitha*3, Maligi A Akshay*4 , Bharath Vanaparthi*5

*1Assistant Professor,Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

*2Associate Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

*3JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

*4JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

*5JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

ABSTRACT

Speech Emotion Recognition is the act of attempting to recognize human emotion from the input audio file and the associated affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone, frequency and pitch. Here we use python modules and machine learning algorithms for detecting the emotion from the audio file. Here we are predicting the emotion and display it to the user on the screen. We have mainly three different steps to detect the emotion and they are like training the model, extracting feature sets and providing input test audio. We use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non-linearly separated. We use python modules like librosa, pyaudio, sklearn, numpy to detect the emotion from audio file.

Keywords: Feature subset extraction, MLP Classifier, Emotion recognition.

I. INTRODUCTION

Traditionally, machine learning (ML) involves the calculation of feature parameters from the raw data (e.g., face emotion, speech, images, video). In case of blind people, it is very much helpful for them to detect the emotion of opposite person just by hearing the voice. The features are used to train a model that learns to produce the desired output labels. A common issue faced by this approach is the choice of features. In general, it is not known which features can lead to the most efficient clustering of data into different categories (or classes). Some insights can be gained by testing a large number of different datasets, combining different features into a common feature vector, or applying various feature selection techniques. Here we train the model using different datasets. The quality of the resulting hand-crafted features can have a significant effect on classification performance.

Speech Emotion is generated by giving a test audio to the trained model and after that using various python libraries the model detects the feature sets and compare them to give the output emotion to the user on screen.

II. LITERATURE SURVEY

Thiang, et al. (2011) presented speech recognition using Linear Predictive Coding (LPC) and Artificial Neural Network (ANN) for controlling movement of mobile robot. Input signals were sampled directly from the microphone and then the extraction was done by LPC and ANN [39]. Ms.Vimala.C and Dr.V.Radha (2012) proposed speaker independent isolated speech recognition system for Tamil language. Feature extraction, acoustic model, pronunciation dictionary and language model were implemented using HMM which produced 88% of accuracy in 2500 words [29]. Cini Kurian and Kannan Balakrishnan (2012) found development and evaluation of different acoustic models for Malayalam continuous speech recognition. In this paper HMM is used to compare and evaluate the Context Dependent (CD), Context Independent (CI) models and Context Dependent tied (CD tied) models from this CI model 21%. The database consists of 21 speakers including 10 males and 11 females [7]. Suma Swamy et al. (2013) introduced an efficient speech recognition system which was experimented with Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ), HMM which recognize the speech by 98% accuracy. The database consists of five words spoken by 4 speakers at ten times [35]. Annu Choudhary et al. (2013) proposed an automatic speech recognition system for isolated and connected words of Hindi language by using Hidden Markov Model Toolkit (HTK). Hindi words are used for dataset extracted by MFCC and the recognition system achieved 95% accuracy in isolated words and 90% in connected words [3]. Preeti Saini et al. (2013) proposed Hindi automatic speech recognition using HTK. Isolated words are used to recognize the speech with 10 states in HMM topology which produced 96.61% [31]. Md. Akkas Ali et al. (2013) presented automatic speech recognition technique for Bangla words. Feature extraction was done by, Linear Predictive Coding (LPC) and Gaussian Mixture Model (GMM). Totally 100 words recorded in 1000 times which gave 84% accuracy [25]. Maya Moneykumar, et al. (2014) developed Malayalam word identification for speech recognition system. The proposed work was done with syllable based segmentation using HMM on MFCC for feature extraction [24]. Jitendra Singh Pokhariya and Dr. Sanjay Mathur (2014) introduced Sanskrit speech recognition using HTK. MFCC and two state of HMM were used for extraction which produces 95.2% to 97.2% accuracy respectively [16]. In 2014, Geeta Nijhawan et al. developed real time speaker recognition system for Hindi words. Feature extraction done with MFCC using Quantization Linde, Buzo and Gray (VQLBG) algorithm. Voice Activity Detector (VAC) was proposed to remove the noisy.

III. PROPOSED SYSTEM

In proposed system we are using traditional mathematics to detect the emotion like detecting the frequency and amplitude of the audio to provide exact prediction of emotion. Because the frequency of voice will be differ based on the emotion which we are using to deliver the audio and we are using python modules like numpy, librosa ro classify the feature sets and train the model.

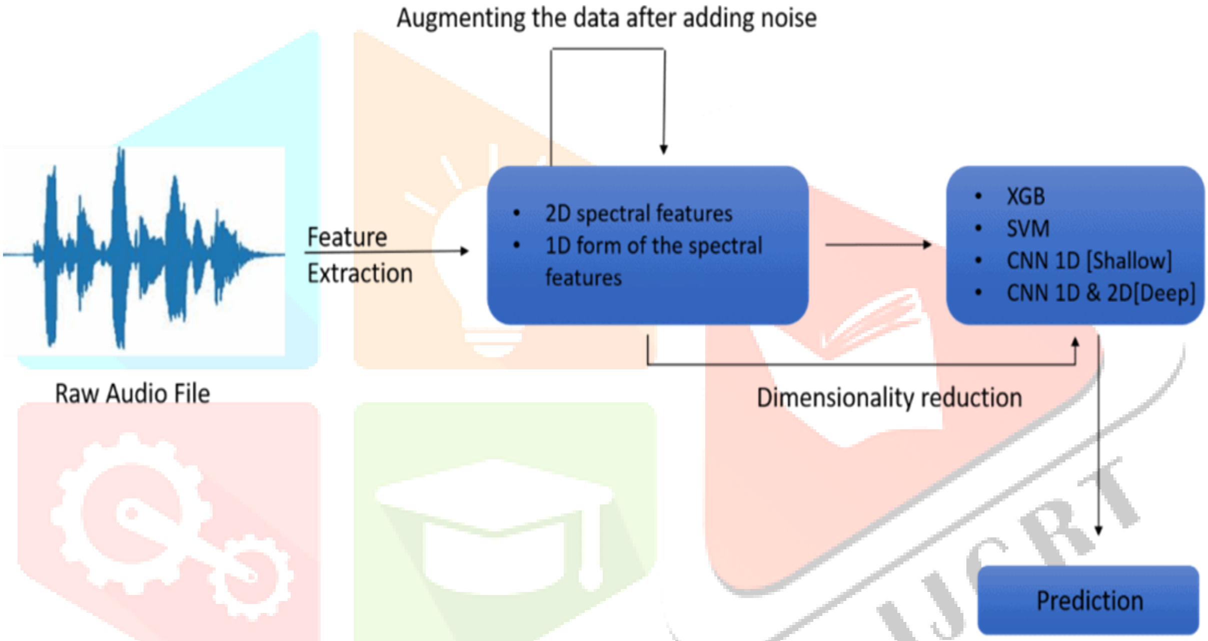

Fig 1: Structure of Speech Emotion recognition System.

Emotion Distribution Bar Graph:

Regarding the distribution of gender, the number of female speakers was found to be slightly more than the male speakers, but the imbalance was not large enough to warrant any special attention.

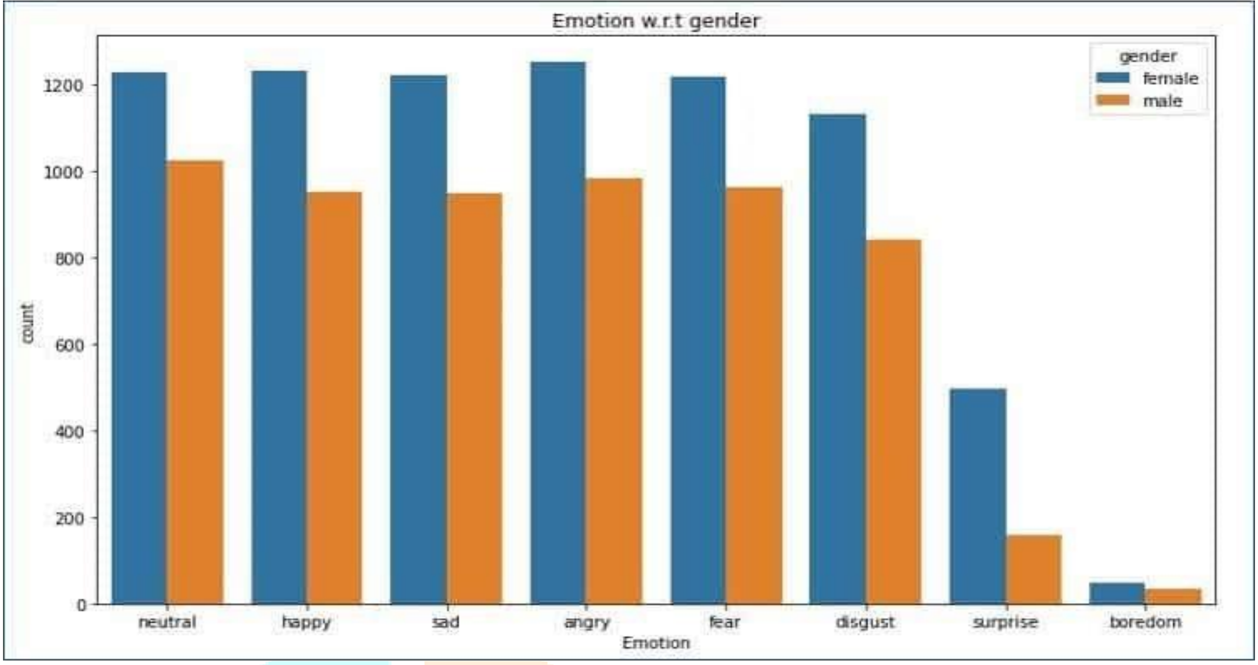

Fig 2: Distributions of emotion with respect to gender bar graph representation

Grouping Similar Emotions

Since the model was confusing between similar emotions like anger-disgust and sad-bored, we tried combining those labels and training the model on 6 classes which were neutral, sadness/boredom, happy, anger/disgust, surprise and fear. The accuracies certainly improved on reducing the number of classes, but this introduced another problem with regards to class imbalance. After, combining anger-disgust and sad-boredom, the model developed a high bias towards the anger-disgust. This may have happened because the number of instances of anger-disgust became disproportionately more than the other labels. So, it was decided to stick with the older model.

METHODOLOGY:

Mel-Frequency Cepstral Coefficients (MFCC) The Mel-frequency cepstral coefficients (MFCC) is one of the most popular audio feature. It is a representation of the speech signals where a feature called the cepstrum of a windowed short-time signal is derived from the FFT of that signal. Afterwards the signal goes to the frequency axis of the melfrequency scale using a log-based transform, and then decorrelated using a modified Discrete Cosine Transform. The steps to extract MFCC features are including pre-emphasis, frame blocking and windowing, FFT magnitude, Mel filterbank, log energy, and DCT. MFCC utilizes the mel-scale, which is tuned to the human’s ear frequency response. Due to this, MFCC has been proven to be invaluable in the speech recognition field and has been attempted to be integrated with emotion recognition. According to Spectral audio features such as MFCC is best suited for a N-way classifier.

Here we are using concept of amplitude and frequency to detect the emotion from the audio file and frequency calculation can be done using numpy library where we have lot of formulas and in this way, emotion is detected. In proposed system we are trying to reduce the background noise which can lead us to the wrong detection of emotion.

For the training, we store the numerical values of emotions and their respective features correspondingly in different arrays. These arrays are given as an input to the MLP Classifier that has been initialized. The Classifier identifies different categories in the datasets and classifies them into different emotions. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. For testing the performance of the model, if we enter the unknown test dataset as an input, it will retrieve the parameters and predict the emotion as per training dataset values. The accuracy of the system is displayed in the form of percentage which is the final result of our project.

Variation in Energy Across Emotions

To ensure uniformity in our study of energy variation as the audio clips in our dataset were of different lengths, a power which is energy per unit time was found to be a more accurate measure. This metric was plotted with respect to different emotions. From the graph See Fig. 2) it is quite evident that the primary method of expression of anger or fear in people is a higher energy delivery. We also observe that disgust and sadness are closer to neutral with regards to energy although exceptions do exist.

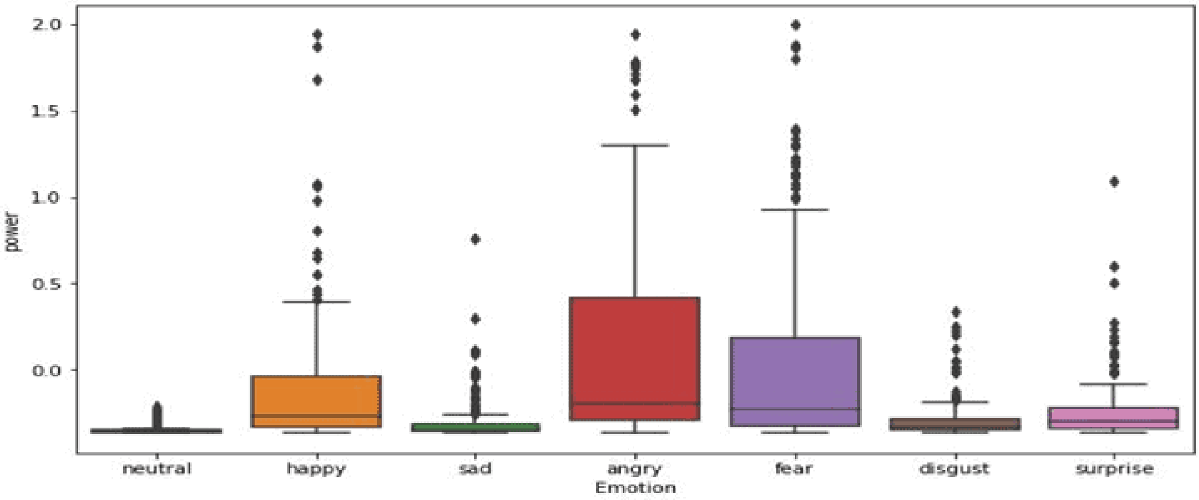

Fig 3: Distributions of emotion with respect to gender

Variation of Relative Pace and Power with respect to Emotions

A scatter-plot of power vs relative pace of the audio clips was analysed and it was observed that the 'disgust' emotion was skewed towards the low pace side while the 'surprise' emotion was skewed more towards the higher pace side. As mentioned before, anger and fear occupy the high power space and sadness and neutral occupy the low power space while being scattered pace-wise. Only, the RAVDESS dataset was used for plotting here because it contains only two sentences of equal length spoken in different emotions, so the lexical features don't vary and the relative pace can be reliably calculated.

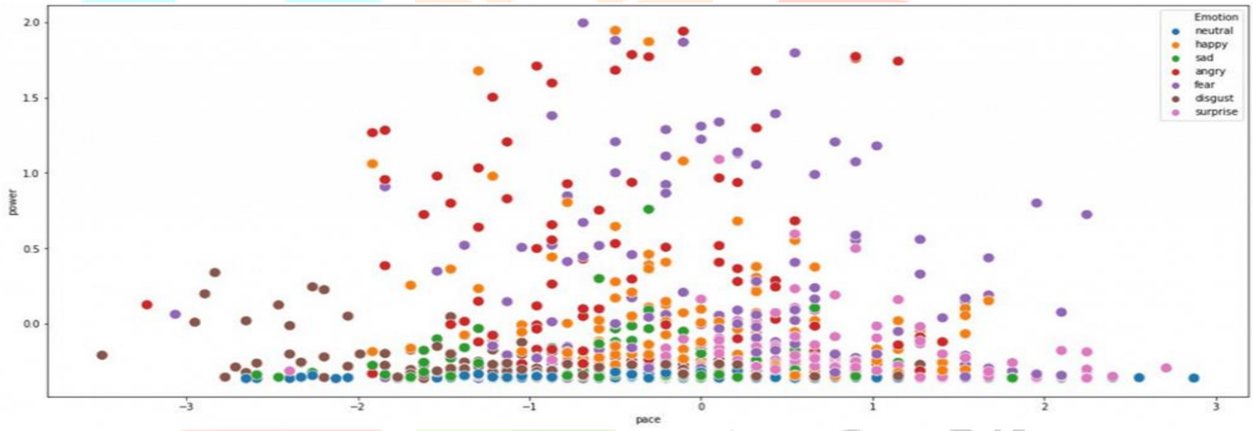

Fig 4: Scatter of power Vs relative pace of audio clips

Calculation of Accuracy

The result is based on the accuracy metrics in which there is a comparison between predicted values and the actual values. A confusion matrix is created which consists of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From confusion metrics, we have calculated accuracy as follows:

$$\text{accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

CONCLUSION

We proposed a speech emotion recognition system that extracts audio emotion features effectively using a DNN model. Unlike conventional approaches, the proposed method will learn pattern features automatically and minimise incompleteness induced by python libraries and extracts features. The Speech emotion recognition is a very challenging problem. They require a heavy effort for enhancing the performance measure of speech emotion recognition. This area of emotion recognition is gaining attention owing to its applications in various domains such as gaming, software engineering, and education. The proposed approach uses training sample audio data to directly input the audio frequency value. Autonomous learning can gain more perfection by cancelling background noise function expressions implicitly.

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of paper depends largely on encouragement and guidelines of many others. We take this opportunity to express our profound gratitude to CMR Technical Campus College Management for motivating us and for providing all the facilities required for this work. We are deeply indebted to Chairman Shri C. Gopal Reddy, Secretary Smt. C. Vasantha Latha , Director Dr . A. Raji Reddy, HOD CSE Dr. K. Srujan Raju, PRC Co-ordinator Dr. Suwarna Gothane, Project Guide K. Mahesh Raj who always has been a constant source of inspiration for us.

REFERENCES

[1] Z. Yongzhao and C. Peng, “Research and implementation of emotional feature extraction and recognition in speech signal,” *Joural of Jiangsu University*, vol. 26, no. 1, pp. 72–75, 2005.

[2] L. Zhao, C. Jiang, C. Zou, and Z. Wu, “Study on emotional feature analysis and recognition in speech,” *Acta Electronica Sinica*, vol. 32, no. 4, pp. 606–609, 2004.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.

[4] Z. Li, “A study on emotional feature analysis and recognition in speech signal,” *Journal of China Institute of Communications*, vol. 21, no. 10, pp. 18–24, 2000.

[5] T. L. Nwe, S. W. Foo, and L. C. de Silva, “Speech emotion recognition using hidden Markov models,” *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[6] L. Zhao, X. Qian, C. Zhou, and Z. Wu, “Study on emotional feature derived from speech signal,” *Journal of Data Acquistion & Processing*, vol. 15, no. 1, pp. 120–123, 2000.