IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 1, January 2020

Speech Emotion Recognition for Malayalam Language

Ashly K John[1]

Department of Computer Science & Engineering, LBS Institute of Technology for Women, Poojappura, Thiruvananthapuram[1]

Abstract: Automatic speech emotion recognition is an active research area in the field of Human Computer Interaction (HCI) with wide range of applications. The proposed work is Speech Emotion Recognition for Malayalam language by using gradient boosted tree classifiers in python. Speech Emotion Recognition is extraction of the emotional state of the speaker from his or her speech signal. In this first we collect different speeches from individuals in Malayalam language. The emotion in one input audio will be found out by extracting features in that audio by MFCC (Mel Frequency Cepstral Co-efficient) and then classified by Gradient boosted tree classifiers. Four types of emotions such as Angry, Happy, Neutral and Sad are identified by this approach. Applications of SER are in the field of Medicine, Counselling, Autism patients, Music therapy, Law and Entertainment – Recognize mood & emotions of user.

Keywords: Speech Emotion Recognition, MFCC, Gradient Boosted Tree Classifiers

I. INTRODUCTION

Recent years have seen increasing interest in automated, and particularly vision-based, methods for establishing the emotional state of a human subject. Many applications exist in artificial intelligence and machine interaction. Effective Automatic Speech Emotion Recognition is an interesting area for Human Computer Interaction. The system must be able to recognize the user's emotion and perform actions. The proposed work includes various modules performing like feature extraction, feature selection, classification and identify the emotion. The major motive of speech emotion recognition system is to identify the emotional state of the person speaking. This can also be used in call centre applications where the support staff can handle the conversation in a more adjusting manner if the emotion of the caller is identified earlier. The system also finds application in intelligent spoken tutoring systems where the computer tutors can adapt to the student's emotion. The research in emotion recognition greatly amplifies the efficiency of people in their work and study and upgrades the quality of life.

The proposed work is to find emotion in Malayalam speech. Audio samples are collected from different individuals. Features are extracted from the input signal by MFCC. Then classify the emotion of it by comparing values in the training dataset by Gradient Boosted Tree Classifier.

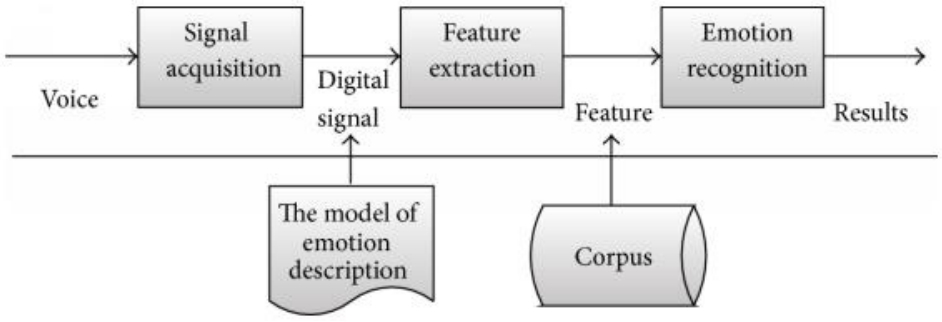

Figure 1: General flowchart for speech emotion recognition

All programs are written by using PYTHON. In this article, there are 6 sections. Section 1. Introduction, 2. LSB method, 3. Skin tone detection, 4. Randomized steganography, 5. Comparison and 6. Conclusion.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 1, January 2020

II. OVERVIEW OF SER

The steps of Speech Emotion Recognition are

- Collection of Emotional Database (Malayalam)
- Feature Extraction - MFCC
- Classification of Emotion
  – Gradient boosted tree classifier
- Recognizing Emotion

III. ALGORITHM

1 Read the audio file
2 Compute the length of the signal
3 Preprocessing the signal by $y(t)=x(t)-\alpha x(t-1)$
4 Compute the frequency points and filter bank co-efficient
Mel Scale, Mel(f) = 2595 * ln(1+f/700)
Filterbank co efficient, f=700(10^(m/2595)−1)
5 Segmenting the values to frames
6 For each short-term window getting from hamming window processing go to steps 7 to 10
7 Get the current window $w[n]=0.54-0.46\cos(2\pi nN-1)$
8 Update window position by incrementing
9 Get FFT magnitude $P=|FFT(xi)|2N$
10 Normalize FFT
11 Compute the MFCC co efficient for final feature and it is stored it in an array

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

12 Compare the predicted features with the trained data and classify the emotion by gradient boosting tree classifier,do steps 13 to 16 otherwise go to step 17
13 Absolute loss function is calculated at each stage by L(y,F) = |y-F|
14 Calculate the true probability distribution y(x) for each datapoint
15 Calculate the predicted probability distribution based on the current function F(x,y)
16 Calculate the difference between true probability and predicted probability and classify the emotion as angry, happy, sad & neutral
17 Stop

IV. METHODOLOGY

4.1 COLLECTION OF EMOTIONAL DATABASE

Speech Samples from individuals are collected in malayalam language. Different sounds of different energy values, frequencies and pitches are collected from male and female speakers

4.2 FEATURE EXTRACTION

Features of an input signals are extracted by MFCC (Mel Frequency Cepstral Co-efficients). Identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc. Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since.

The steps are

Step 1: Pre-emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y(n) = X(n) - a*X(n-l)$$

Lets consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 1, January 2020

Step 2: Framing

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. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N).

$$\text{Frame} = \text{floor}((1- N)\ 1M) + 1$$

Typical values used are M = 100 and N= 256.

Step 3: Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

$$w[n]=0.54-0.46\cos(2\pi nN-1)$$

N = number of samples in each frame

Step 4: Fast Fourier Transform

In FTT convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$Y(w)= FFT[h(t)* X(t)]= H(w)*X(w)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Step 5: Triangular Band pass Filters

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale . Multiply the magnitude frequency response by a set of 20 triangular band pass filters to get the log energy of each triangular band pass filter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation:

$$\text{Mel}(f) =2595*\ln(1 +f/700)$$

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

4.3 CLASSIFICATION OF EMOTIONS

Gradient boosted (GB) tree classifier is used for classifying emotions. GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n classes regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced. It works in a stage wise manner. A loss function to be optimized at each stage. A weak learner (Decision trees) to make predictions. An additive model to add weak learners to minimize the loss function. Deviance loss function is minimized by adding new trees at each stage. By this, additional copies of classifier are fitted on the initial dataset and weight of samples that have been incorrectly classified are adjusted.

Steps for calculating loss function for each data point

- Turn the label y_i into a true probability distribution $Y_c(xi)$

  For example : $y_5 = G$,
  $Y_A(x_5)=0$, $Y_B(x_5) = 0,\ldots,Y_G(x_5) = 1,\ldots..Y_Z(x_5)=0$
- Calculate the predicted probability distribution $P_c(x_i)$ based on the current model $F_A, F_B,\ldots F_Z$

$P_A(x_5) = 0.03$, $P_B(x_5) = 0.05,\ldots,P_G(x_5) = 0.3,\ldots,P_Z(x_5) = 0.05$

- Calculate the difference between the true probability distribution and the predicted probability distribution

By this value, emotions are recognised as Angry, Happy,Neutral and Sad.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 1, January 2020

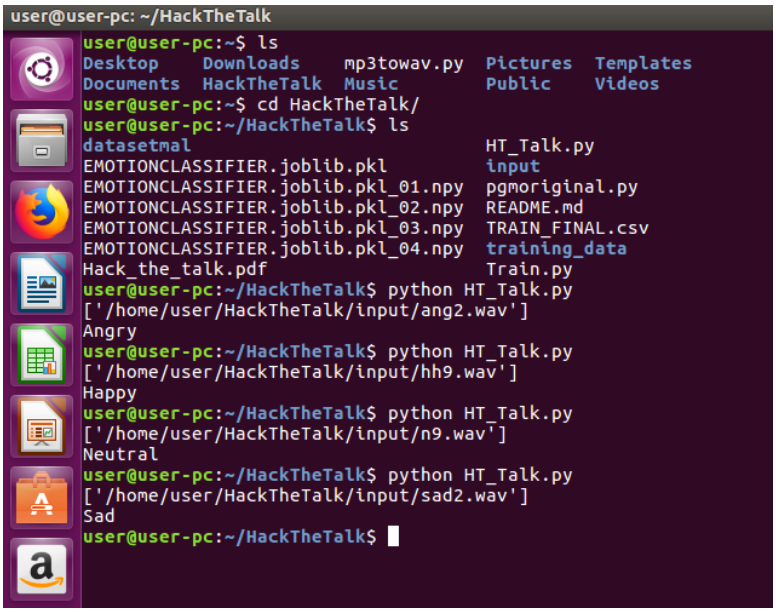

Figure 2: Output of proposed system

V. CONCLUSION

In this paper emotions in Malayalam language is recognised as Angry, Happy, Neutral and Sad by using gradient boosted tree classifiers. In this work emotions are efficiently recognised. We can use this method for different persons in different situations. Adding new emotions to this is its future work.

REFERENCES

[1]. Carlos Busso, Soroosh Mariooryad, Angeliki Metallinou, and Shrikanth Narayanan, (2013) "Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech", *IEEE Transactions On Affective Computing, Vol. 4, No. 4.*

[2]. Peng Song, Yun Jin, Cheng Zha and Li Zhao, (2015), "Speech emotion recognition method based on hidden factor analysis*", Electronics Letters, Vol. 51 No. 1 pp. 112–114*

[3]. Yuan Zong, Wenming Zheng, Zhen Cui and Qiang Li, (2016), "Double sparse learning mode for speech emotion recognition" Electronics Letters, Vol. 52 No. 16 pp. 1410–1412

[4]. Emily Mower, Student Member, IEEE, Maja J Mataric´, Senior Member, IEEE, and Shrikanth Narayanan, Fellow, (2011). "A Framework for Automatic Human Emotion Classification Using Emotion Profiles", *IEEE, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 5.*

[5]. C.Sunitha Ram Department of CSE, Principal, SCSVMV University, Dr.R.Ponnusamy, Madha Engineering College, Kanchipuram,India Chennai,India, (2014) "An Effective Automatic Speech Emotion Recognition for Tamil language based on DWT and MFCC using Stability-Plasticity Dilemma Neural Network", *ICICES.*

[6]. Mohan Ghai, Shamit Lal, Shivam Dugga l and Shrey Manik, Jagvir Kaur1, Er. Rakesh Singh Delhi Technological University,(2014),"Emotion Recognition On Speech Signals Using Machine Learning", *978-1-5090-6399-4/17/$31.00c, IEEE4.*

[7]. M.S. Likitha,1 Sri Raksha R. Gupta,2 K. Hasitha3 and A. Upendra Raju4 Dept. of Electronics, Mount Carmel College, Autonomous, Bangalore,(2017) "Speech Based Human Emotion Recognition Using MFCC", *IEEE WiSPNET conference.*

[8]. Anuja Pawar ME Student Department of Computer Engineering D. Y. Patil College of Engineering, pune, India,(2017), "Recognition and Classification of Human Emotion from Audio".*IJESC.*

BIOGRPAHIES

Ashly K John, received her B. Tech degree in Computer Science and Engineering from University of Kerala. She is now pursuing her M. Tech degree in Computer Science and Engineering from LBS Institute of Technology for Women, Thiruvananthapuram affiliated to APJ Abdul Kalam Technological University. Her areas of interest are Information Systems, Networking etc.