Emotion Recognition in Speech by MFCC and SVM

Akhilesh Watile, Vilas Alagdeve, Swapnil Jain

***Abstract* —Speech emotion recognition is one of the latest challenges in speech processing and Human Computer Interaction (HCI) in order to address the operational needs in real world applications. Besides human facial expressions, speech has proven to be one of the most promising modalities for automatic human emotion recognition. Speech is a spontaneous medium of perceiving emotions which provide in-depth information related to different cognitive states of a human being. In this context, we introduce a novel approach using a combination of prosody features (i.e. pitch, energy, Zero crossing rate), quality features (i.e. Formant Frequencies, Spectral features etc.), derived features ((i.e.) Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding Coefficients (LPCC)) and dynamic feature (Mel-Energy spectrum dynamic Coefficients (MEDC)) for robust automatic recognition of speaker's emotional states. Multilevel SVM classifier is used for identification of seven discrete emotional states namely angry, disgust, fear, happy, neutral, sad and surprise[1].**

***Index Terms*— Linear Predictive Coding Coefficients, Mel Frequency Cepstral Coefficients, Prosody features, Quality features, Speech Emotion Recognition, Support Vector Machine**

I. INTRODUCTION

Emotion recognition system from speech is one of most advanced topics in the electronic media. Emotion detection helps the security system to prevent the data from various attacks at the cyber world. A lot of research work has already been done into this contrast but the problem of accuracy is always there. This work has been done to categorize three emotions namely HAPPY, FEAR AND SAD using the EEMD, SVM and ANN algorithms. In this work, noise levels are taken so that the emotion can be identified even though if the voice signal is highly noised.

The aim of this work is to check the accuracy of the EEMD algorithm with noisy signals in contrast to the emotion detection. We proceed as detecting the noise level and segmenting the signal for the further processing. There are two segments: first part is the training part in which the system is trained to identify the further proceedings. In this part, samples of each voice category are taken and their features are fetched after successful segmentation of the voice file and further on saved into the database. The second part is the testing part in which a voice sample is taken and all the required properties are fetched and matched with the saved database values. The closest match comes out as the category of the voice file.

***Akhilesh Watile**, Dept.of ElectronicsEngg. YCCE Nagpur,India, +91 7776065239.*

***Vilas Alagdeve**, Dept.of ElectronicsEngg. YCCE Nagpur, India,+91 7768842506.*

***Swanil Jain** , Dept.of Electronics&Telecomm Engg. DMIETR Wardha, India,+91 9423421866*

II. DATABASE SELECTION

Wherever In speech emotion recognition system selection of proper database is a critical task. The efficiency of speech emotion recognition is highly depends upon the naturalness of the database selected. Different databases are implied by different researches. Generally, there are two types of databases that are used in emotion recognition – acted and real[10]. As the name suggests, in acted emotional speech corpus, a professional actor is asked to speak in a certain emotion. In real databases, speech databases for each emotion are obtained by recording conversations in real-life situations such as call centers and talk shows. But it has been observed that there is a difference in the features of acted and real emotional speeches. This is because acted emotions are not felt while speaking and thus come out more strongly. In this work, both acted as well real emotional speech have been considered for classification.

III. BLOCK DIAGRAM

Speech emotion recognition system is a system which recognizes emotion from user's speech. The structure of speech emotion system is illustrated in Figure 3.1. It has speech with emotions as input, then features are extracted and then classification is done by taking algorithm for classification. The proposed human emotion recognition system is of five components: input speech signal, pre-processing, feature extraction and selection, classification and finally emotions recognition. The Accuracy of the Emotional speech recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations. As we change the emotion it can be change with pitch, energy, speech and spectrum. First pre-processing done for removing the silent part from the speech which doesn't containing information [5].

Speech emotion recognition system aims to automatically identify the emotion of human beings from speaker's voice. It is based on the speech signal, extracting the features which contain emotional information from the speaker's speech, and using appropriate method to recognize the emotion. This system consists of 5 steps, namely:-

1. Emotional speech input 2.Feature extraction and selection, 3.Training, 4.Classification, 5.Emotion recognition

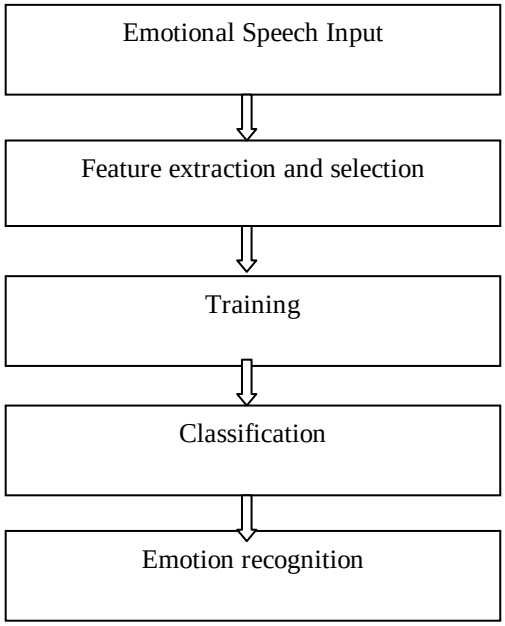

Figure 1: Basic Block Diagram of system

IV. Feature Extraction

Different emotional states can be recognized using certain speech features which can be either prosody features or quality features. Some Prosody features which can be extracted directly; includes pitch, intensity and energy are the most widely used features in the emotion recognition domain. Though it is possible to distinguish some emotional states using only these features, but it becomes very inconvenient when it comes to emotional states with same level of stimulation

A.MFCC

Mel-frequency cepstral coefficients (MFCCs) are a parametric representation of the speech signal, mainly used in emotion recognition system, but they have proved to be successful for other purposes as well, among them speaker identification and emotion recognition. They consider as robust take part in any feature for any other speech task. Mel is the unit of the frequency which different by considering for the human ears. Both MFCC and FP feature extracted for speaker independent emotion recognition for continues feature section [1].

In this first pre-emphasis used to boost the amount of energy in the high frequency. Then apply hamming windowing technique for create section. For know how much energy require for each band we require convert all time domain into frequency domain by using DFT.Then it gives to mel filter bank band log make the feature less sensitive for variation of input .Cepstrum define inverse DFT of log of the that signal or speech for 12 cepstral coefficients for each frame

V. Training

As we can take the real time database so we should require training compulsory. But for our trained acted database 70% of database require training. We used different database combine different feature to build different training model and analyse their recognition accuracy. First off we want to analyze and feature extract small collection of audio sample storing there feature data as training data. In this project we trained our own speech wave sample trained it by adding one or more database at time[6] .It shows its sound extraction with database accuracy. In the training section we would be taking fifty voice samples of each and every category taken for the classification. In this scenario, we would-be fetching properties of each voice sample and after putting them into an array. We would be storing the average of each property of each section into the database.

VI. Testing

Like any other recognition systems, emotion recognition systems also involve two phases namely, training and testing. Training is the process of familiarizing the system with the emotions characteristics of the speakers. Testing is the actual recognition task. The speech emotion recognition system has the emotional speech as an input and the classified emotion as an output. At the time of testing we would be using a combinational algorithm using the SVM and NEURAL feed forward method. In this part, we would be finalizing the saved data of the database and would provide it to the NEURAL classifier [7]. On the basis of the saved database, the neural classifier would match the properties with the uploaded file and would produce a result. The classification would be done on the basis of 5 categories: happy, sad, angry and neutral.

VII. Classification Scheme

A speech emotion recognition system consists of two stages: (1) afront-end processing unit that extracts the appropriate features from the available (speech) data, and (2) a classifier that decides the underlying emotion of the speech utterance. In fact, most current research in speech emotion recognition has focused on this step since it represents the interface between the problem domain and the classification techniques. On the other hand, traditional classifiers have been used in almost all proposed speech emotion recognition systems.

A.SVM

An important example of the general discriminate classifiers is the support vector machine SVM classifiers are mainly based on the use of kernel functions to nonlinearly map the original features to a high-dimensional space where data can be well classified using a linear classifier[3]. SVM classifiers are widely used in many pattern recognition applications and shown to outperform other well-known classifiers. They have some advantages over GMM and HMM including the global optimality of the training algorithm, and the existence of excellent data-dependent generalization bounds. However, their treatment of non separable cases is somewhat heuristic[2]. In fact, there is no systematic way to choose the kernel functions, and hence, reparability of the transformed features is not guaranteed. In fact, in many pattern recognition applications including speech emotion recognition, it is not advised to have a perfect of the training data so as to avoid over-fitting.

VIII. Result And Discussion

A.MFCC

First we take one wave file sample with some frame rate at the section part. As MFCC feature extracted in the five most important part. For cepstrum about 12-13 feature extracted in any wave sample. Large numbers of features are extracted in the frequency response and frequency reconstruction at the section end. MFCC features extracted into following five parts:-

- o Cepstrum :13
- o Frequency response:256
- o Frequency Bandwidth:40
- o FB Reconstruction:40
- o Frequency Reconstruction:256

Total: 605

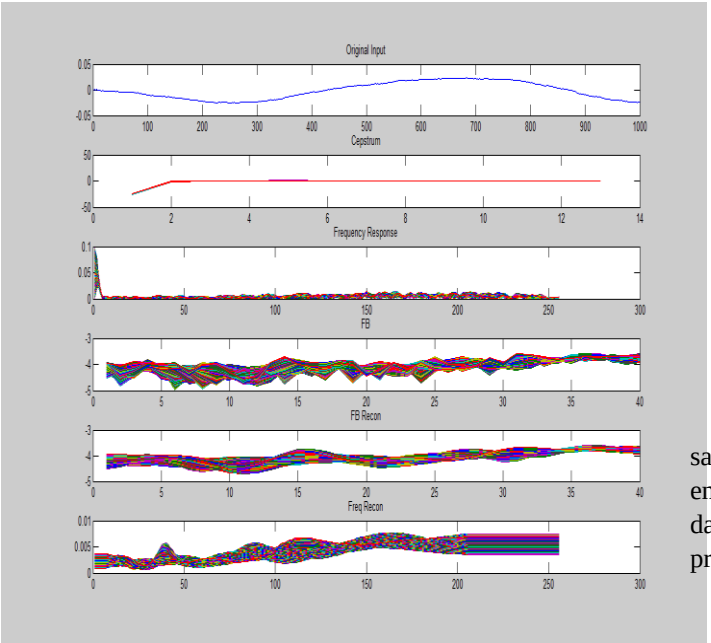

Fig 2. Extracted feature by MFCC

B.TRAINING

The Berlin Emotion database contains 406 speech files for five emotion classes. We can elaborate our own speech wave sample for the feature extraction in the condition for that. Other part of wave sample which is in other format we can converted in .wave format.We use both database, combine different features to build different training models, and analyse their recognition accuracy. First off, we want to analyze and feature extract a small collection of audio sample storing there feature data as training data.

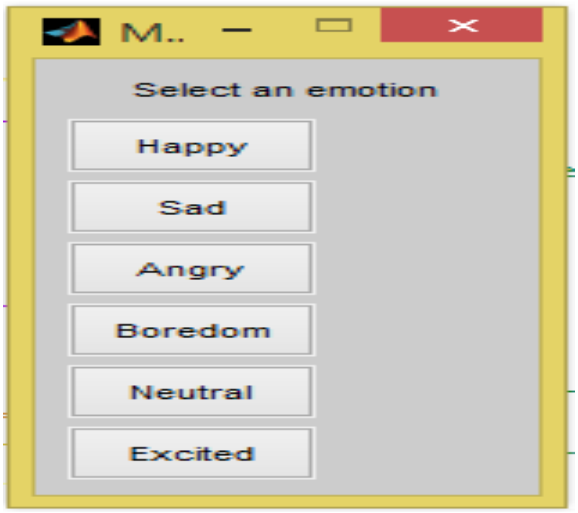

Fig. 3 Selection of emotion from database

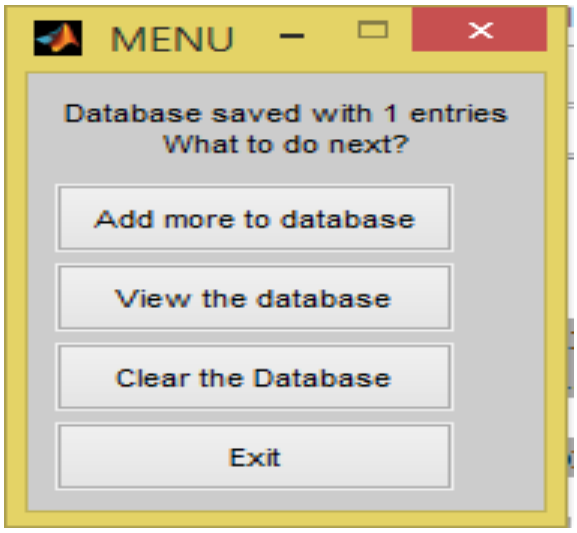

Fig.4 Creating database and add entries

As example we create database for 4 wave samples and saved it by adding entries. Then it show the window i.e. for enter database entry number from 1to 4.If we select the database number as 2 then it shows the emotion for that proper database number.

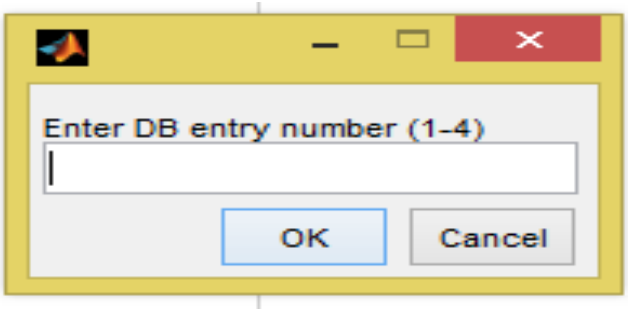

Fig.5 Database entry number for emotion saved

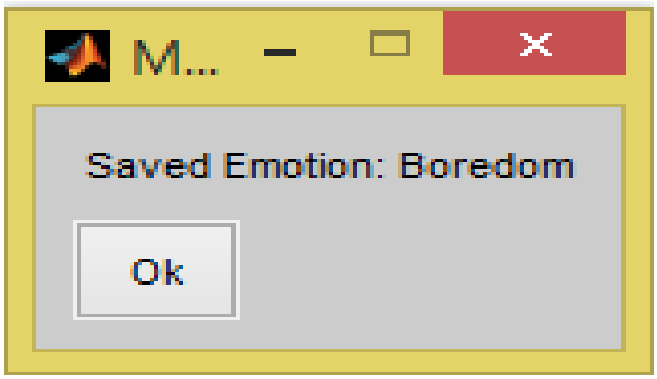

Fig.6 Emotion for enter database number

C.EVALUTION OF DATABASE BY SVM

The main idea of SVM classification is to transform the original input set to a high dimensional feature space by using kernel function. In Classification, training examples are used to learn a model that can classify the data samples into known classes.We use two class SVM in Mat lab.

15 SVM inputs are classified as below:-

- HAPPY(H)=HS,HA,HB,HE,HN
- SAD(S)=SA,SB,SE,SN
- ANGER(A)=AB,AE,AN
- BOREDOM(B)=BE,BN
- EXICTED(E)=EN
- NEUTRAL(N)

Load the one wave sample data base Then it shows the MFCC feature extraction. And calculate 15 classes of SVM as input for it that separate out into number of classes 3 to 4 depend on which is higher in two class SVM.

- Take one wave Sample as DB, as example it is for sad emotion.
- Then its show its MFCC extraction and classify into classes as
  - Class 1(H), Count:2
  - Class 2(A), Count:3
  - Class 3(S), Count:4
  - Class 4(N), Count:2
- Then it shows SAD (class3) as recognized emotion for that and time required for that.

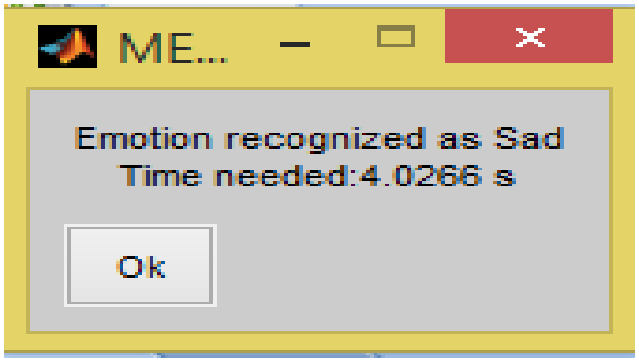

Fig.7 Recognized Emotion and Time Needed

IX. CONCLUSION

In this work, speech emotion recognition system acoustic part of speech carries important info about emotions. MFCC are used for the feature extraction .Algorithm with the SVM's overall performance is tested. As training is done for the learning module section to create database selection.

ACKNOWLEDGMENT

This research was supported by publisher of this paper. We thank our colleagues from YCCE Nagpur who provided expertise that greatly assisted the research. Also, for sharing their pearls of wisdom with us during the course of this review paper.

REFERENCES

[1] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li, "Speech Emotion Recognition Using Fourier Parameters," IEEE Transaction on affective Cnew omputing, vol. 6, no. 1, January -march 2015

[2] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 4, pp. 582–596, May 2009

[3] E.Vayrynen,J.Kortelainen,and T. Seppanen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody," IEEE Trans. Affective Comput ., vol. 4, no. 1, pp. 47–56, Jan.-Mar. 2013.

[4] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. New- man, "Stress and emotion classification using jitter and shimmer features," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2007, vol. 4, pp. IV– 1081–IV-1084.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech Signal Process., vol. 28, no. 4, pp. 357–366, Aug. 1980.

[6] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw., vol. 13, no. 2, pp. 415–425, Mar. 2002.

[7] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 201–216, Mar. 2001.

[8] M. Y. You, C. Chen, J. J. Bu, J. Liu, and J. H. Tao, "Emotion recognition from noisy speech," in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2006, pp. 1653– 1656.

[9] M. Hayat and M. Bennamoun, "An automatic framework for textured 3D video-based facial expression recognition," IEEE Trans. Affective Comput., vol. 5, no. 3, pp. 301–313, Jul.-Sep. 2014.

[10] K. X. Wang, Q. L. Zhang, and S. Y. Liao, "A database of elderly emotional speech," in Proc. Int. Symp. Signal Process. Biomed. Eng Informat. , 2014, pp. 549–553.

[11] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intel. Syst. Technol. (TIST), vol. 2, no. 3, p. 27, 2011.

[12] M. Lugger and B. Yang, "Psychological motivated multi-stage emotion classification exploiting voice quality features," in Speech Recognition, F. Mihelic and J. Zibert.