RESEARCH ARTICLE OPEN ACCESS

Human Emotion Recognition From Speech

Miss. Aparna P. Wanare*, Prof. Shankar N. Dandare
*(Department of Electronics & Telecommunication Engineering, Sant Gadge Baba Amravati University, Amravati)
**(Department of Electronics & Telecommunication Engineering, Sant Gadge Baba Amravati University, Amravati)

ABSTRACT

Speech Emotion Recognition is a recent research topic in the Human Computer Interaction (HCI) field. The need has risen for a more natural communication interface between humans and computer, as computers have become an integral part of our lives. A lot of work currently going on to improve the interaction between humans and computers. To achieve this goal, a computer would have to be able to distinguish its present situation and respond differently depending on that observation. Part of this process involves understanding a user's emotional state. To make the human computer interaction more natural, the objective is that computer should be able to recognize emotional states in the same as human does. The efficiency of emotion recognition system depends on type of features extracted and classifier used for detection of emotions.

The proposed system aims at identification of basic emotional states such as anger, joy, neutral and sadness from human speech. While classifying different emotions, features like MFCC (Mel Frequency Cepstral Coefficient) and Energy is used. In this paper, Standard Emotional Database i.e. English Database is used which gives the satisfactory detection of emotions than recorded samples of emotions. This methodology describes and compares the performances of Learning Vector Quantization Neural Network (LVQ NN), Multiclass Support Vector Machine (SVM) and their combination for emotion recognition.

***Keywords* -** Emotion recognition; Feature extraction; Mel-scale Frequency Cepstral Coefficients; Neural Network; Support Vector Machines;

I. INTRODUCTION

Emotion recognition through speech is an area which increasingly attracting attention within the engineers in the field of pattern recognition and speech signal processing in recent years. Emotion recognition plays an important role in identifying emotional state of speaker from voice signal. Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. The emotional and physical states of a speaker are known as emotional aspects of speech and are included in the so-called paralinguistic aspects [1]. Accurate detection of emotion from speech has clear benefits for the design of more natural human- machine speech interfaces or for the extraction of useful information from large quantities of speech data. It is also becoming more and more important in computer application fields as health care, children education, etc. In speech-based communications, emotion plays an important role [2].

The proposed system aims at identification of basic emotional states such as anger, joy, neutral and sadness from human speech. While classifying different emotions, features like MFCC (Mel Frequency Cepstral Coefficient) and Energy is used. In this paper, Standard Emotional Database i.e. English Database is used which gives the satisfactory results .This methodology describes and compares the performances of Learning Vector Quantization Neural Network (LVQ NN), Multiclass Support Vector Machine (SVM) and their combination for emotion recognition. The overall experimental results reveal that combination of LVQ NN-SVM has greater accuracy than LVQ NN and SVM.

II. BASIC ARCHITECTURE

The block diagram of the emotion recognition system through speech considered in this study is illustrated in Fig. 1. The block diagram consists of the emotional speech as input, feature extraction, feature selection, classifier and detection of emotion as the output [3].

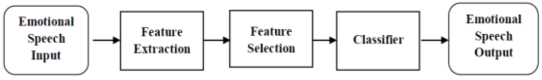

Fig. 1: Basic Block Diagram of Emotion Recognition

a. Emotional Speech Input: A suitable emotional speech database is important requirement for any emotional recognition model. The quality of database determines the efficiency of the system. The emotional database may contain collection of acted speech or real data world.

b. Feature Extraction and Selection: An important step in emotion recognition system through speech is to select a significant feature which carries large emotional information about the speech signal. After collection of the database containing emotional speech proper and necessary features such as prosodic and spectral features are extracted from the speech signal. The commonly used features are pitch, energy, MFCC, LPCC, formant. The steps involved in calculation of MFCC are shown below.

A. MFCC

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound. Fig. 2 shows the MFCC feature extraction process [4] [5]. As shown in Fig. 2 feature extraction process contains following steps:

- Pre-processing: The continuous time signal (speech) is sampled at sampling frequency. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This pre-emphasis is done by using a filter.
- Framing: It is a process of segmenting the speech samples obtained from analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 msec. Framing enables the non-stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behaviour within the short time period of 20-40 msec.
- Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame.
- FFT: Fast Fourier Transform (FFT) algorithm is widely used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain.
- Mel Filter bank and Frequency wrapping: The mel filter bank consists of overlapping triangular filters with the cut-off frequencies determined by the centre frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale.
- Take Logarithm: The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition.
- Take Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components.

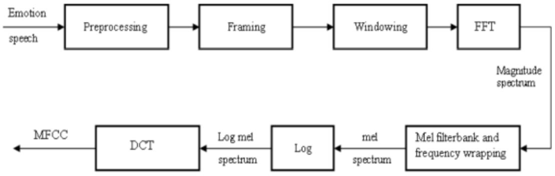

Fig. 2: Block Diagram of the MFCC Feature Extraction

B. Energy

The Energy is the basic and most important feature in speech signal. Energy frequently referred to the volume or intensity of the speech, where it is also known to contain valuable information. Energy provides information that can be used to differentiate sets of emotions, but this measurement alone is not sufficient to differentiate basic emotions. Joy and anger have increased energy level, where sadness has low energy level. Mean of energy is taken into consideration in proposed emotion recognition system [6] [7].

$$E_n = \sum_{n=1}^{N} x(n).x^{*}(n)$$

C. Classifier

The most important aspect of emotion recognition system through speech is classification of an emotion. The performance of system is dependent on proper choice of classifier. There are many types of classifier such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Artificial Neural Network (ANN) and Support Vector Machine (SVM).

III. PROPOSED SYSTEM

The ultimate goal of system design should be its simplicity and efficiency. The Fig. 3 shows architecture of a speech emotion recognition system. Generally the speech files are in .mp3 or in .wav format. For proposed system .wav format files are used. The English Emotion database contains files in .wav format. As shown in Fig. 3 system is separated in two parts. The left hand side represents training part of system while right hand side represents testing part. System contains following major blocks:

1. Input speech emotion
2. Feature Extraction
3. Training

4. Classification

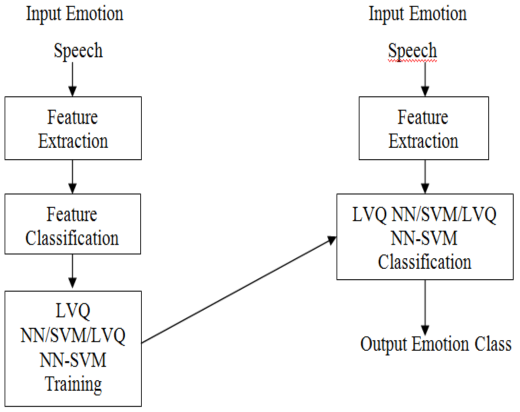

Fig. 3: Structure of Speech Emotion Recognition System

3.1 Input

As shown in above Fig the input is a .wav file containing emotional speech utterances from English Emotion Database. An English Emotion database is used for training and testing the SVM, LVQ NN and LVQ NN- SVM. The English Emotional Speech (EES) Database expresses four emotional states (happy, sad, angry and neutral) is used for the conduction of the experiment. The basic material of the database consists of 'clips' extracted from the selected recordings. Clips ranged from 3-8 sec's in length. An audio file (.wav format) contains speech alone, edited to remove sounds other than the main speaker. Here some samples are used for training the system and then remaining samples are used for testing purpose.

3.2 Feature Extraction

It is an important step in emotion recognition system to extract the features which contains maximum information related to human emotions. A proper selection of set of features can increase efficiency of system. In this paper two features are taken into consideration and extracted from audio samples they are MFCC and Energy.

The steps involved in calculation of MFCC are show above.

3.3 Classifier

Support Vector Machine

The Support Vector Machine is used as a classifier for emotion recognition. The SVM is computer algorithm used in pattern recognition for data classification and regression. The classifier is used for classifying or separating the features from other features. Support Vector Machine performs classification by constructing an N-dimensional hyper-plane that optimally separates the data into categories. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.The classification is achieved by a linear or nonlinear separating surface in the input space of the dataset [8] [9]. The SVM is a binary classifier but with some approaches it can be used as multiclass classifier. Two common methods to build binary classifiers are where each classifier distinguishes between (i) one of the labels to the rest (one-versus-all) or (ii) between every pair of classes (one-versus-one). Classification of new instances for one-versus-all case is done by winner takes- all strategy, in which the classifiers with the highest output function assign the class.

The classification of one-versus-one case is done by max-wins voting strategy, in which every classifier assign the instance to one of the two classes, then the vote for the assigned class is increased by one vote. Finally the class with most votes determines the instance classification. The Fig. 4 for one-versus-all and Fig. 5 for one-versus-one is shown below.

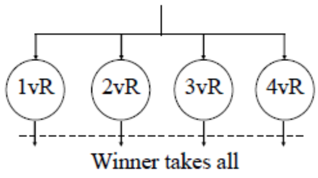

Fig. 4: One-versus-all Approach

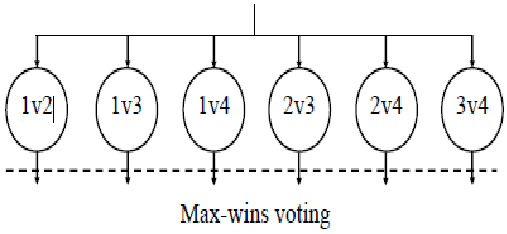

Fig. 5: One-versus-one Approach

In this paper classification is carried out with help of Multiclass classifier i.e. one-to-one multiclass approach, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification. The one-versus-one (1v1) classifier uses a 'max-wins' voting strategy. It constructs $m(m - 1)/2$ binary classifiers, one for every pair of distinct classes. Each binary classifier *Ci j* is trained on the data from the *i* th and *j* th classes only. For a given test sample, if classifier *Ci j* predicts it is in class *i* , then the vote for class *i* is increased by one; otherwise the vote for class *j* is increased by one. Then the 'max-wins' voting strategy assigns the test sample to the highest scoring class.

Learning Vector Quantization

LVQ can be understood as a special case of an artificial neural network, more precisely; it applies a winner-take-all Hebbian learning based approach. It is a precursor to Self-organizing map and LVQ was invented by Kohonen. The LVQ network architecture is shown below in Fig. 6.

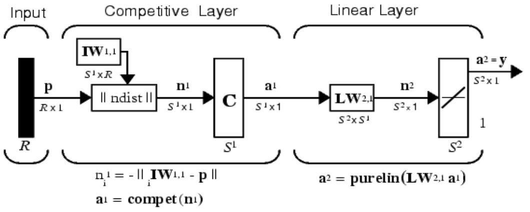

Fig. 6: LVQ Network Architecture

Where R= number of elements of input vector
S^1= number of competitive neurons
S^2= number of linear neurons

An LVQ network has a first competitive layer and a second linear layer. The competitive layer learns to classify input vectors in much the same way as the competitive layers of Self-Organizing Map. The linear layer transforms the competitive layer's classes into target classifications defined by the user. Learning Vector Quantization (LVQ) is a neural net that combines competitive learning with supervision. It can be used for pattern classification. Learning Vector Quantization (LVQ) is a method for training competitive layers in a supervised manner. A competitive layer automatically learns to classify input vectors. However, the classes that the competitive layer finds are dependent only on the distance between input vectors. If two input vectors are very similar, the competitive layer probably will put them in the same class. There is no mechanism in a strictly competitive layer design to say whether or not any two input vectors are in the same class or different classes. LVQ networks, on the other hand, learn to classify input vectors into target classes chosen by the user. In the competitive layer, neuron in the first layer learns a prototype vector which allows it to classify a region of the input space.

Hybrid Method (SVM-LVQ NN)

In this paper, a combination of Support vector Machine (SVM) and Learning Vector Quantization (LVQ) is proposed. This paper clears that the result can be enhance by combining the properties of SVM and LVQ NN. An advantage of LVQ is that it creates prototypes that are easy to interpret for experts in the respective application domain. LVQ systems can be applied to multi-class classification problems in a natural way. A competitive layer automatically learns to classify input vectors. LVQ networks, on the other hand, learn to classify input vectors into target classes chosen by the user.Support Vector Machine performs classification by constructing an N-dimensional hyper-plane that optimally separates the data into categories. It is one of the best methods for classification of emotions from speech. Hence we proposed a system based on combination of LVQ NN and SVM.A comparative analysis of result of single NN, single SVM and hybrid model of both is done.

IV. RESULT AND CONCLUSION

Result

For comparison of results using three different approach graphical representations is shown in following figure. The result obtained using hybrid combination is comparatively superior to remaining two methods for each emotion. The overall accuracy of proposed method is boosted by 6-10%.

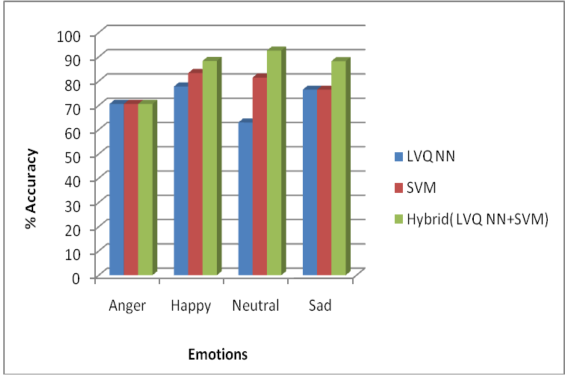

Fig. 7: Percent Accuracy of each Emotion in NN, SVM and Hybrid Model

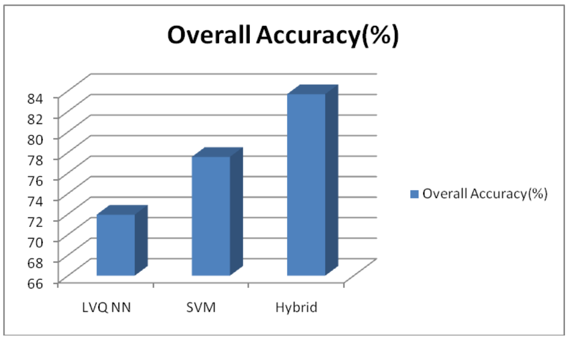

Fig. 8: Percent Overall Accuracy of each Model

V. CONCLUSION

Recognition of emotional states from speech is a current research topic with wide range. Emotion recognition through speech is particularly useful for applications in the field of human machine interaction to make better human machine interface. It is gaining a lot of importance due to its wide range of application in day to day life.

In this paper, the features to be extracted are MFCC and Energy from English Emotion Database. The emotion recognition accuracy using LVQ NN is 71.94% whereas by multiclass SVM is 77.57%. The proposed method i.e. hybrid model is designed and implemented. The hybrid model i.e. LVQ NN-SVM yields better accuracy i.e. 83.68% than other two methods. From the result it can be concluded that the proposed method LVQ NN-SVM yields better result than LVQ NN and SVM and has been successfully implemented.

REFERENCES

[1] Prof. Sujata Pathak, Prof. Arun Kulkarni, *"Recognizing Emotions from Speech"*, 3rd International conference, Vol.:6, pp.107-109, IEEE 2011.

[2] Vaishali M. Chavan, V. V. Gohokar, *"Speech Emotion Recognition by using SVM-Classifier"*, International Journal of Engineering and Advanced Technology (IJEAT), Vol: 1, Issue: 5, pp.11-15, ISSN: 2249-8958, June 2012.

[3] Dipti D. Joshi1, Prof. M. B. Zalte, "*Speech Emotion Recognition: A Review*" IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), Volume 4, Issue 4 ,Pp 34-37, ISSN: 2278-2834, (Jan. - Feb. 2013)

[4] Bhoomika Panda, Debananda Padhi, Kshamamayee Dash, Prof. Sanghamitra Mohanty*, "Use of SVM Classifier & MFCC in Speech Emotion Recognition System"*, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, Issue 3, pp.226-230, ISSN:2277-128X, March 2012.

[5] Sujata B. Wankhade, Pritish Tijare, Yashpalsing Chavhan, *"Speech Emotion Recognition System Using SVM AND LIBSVM"*, International Journal of Computer Science And Applications Vol. 4, No. 2, pp.89-96, ISSN: 0974-1003, July 2011.

[6] Yixiong Pan, Peipei Shen and Liping Shen *,"Speech Emotion Recognition Using Support Vector Machine"*, International Journal of Smart Home Vol. 6, No. 2, pp.101-108, April 2012.

[7] Mohammad Masoud Javidi and Ebrahim Fazlizadeh Roshan, *"Speech Emotion Recognition by Using Combinations of C5.0, Neural Network (NN), and Support Vector Machines (SVM) Classification Methods"*, Journal of mathematics and computer Science, Vol.: 6, Issue: 3, pp.191-200, 2013.

[8] A. Milton, S. Sharmy Roy, S. Tamil Selvi, *"SVM Scheme for Speech Emotion Recognition using MFCC Feature"*, International Journal of Computer Applications, Vol. 69 – No. 9, pp. 34-40, May 2013.

[9] Thapanee Seehapoch, Sartra Wongthanavasu*, "Speech Emotion Recognition Using Support Vector Machines"*, 5th International Conference on Knowledge and Smart Technology, pp. 86-91, 2013.