ISSN:0975-9646

Non Linear Classification for Emotion Detection on Telugu Corpus

B.S.Yalamanchili[1], Anusha.K.K[2], Santhi.K[3], Sruthi.P[4], SwapnaMadhavi.B[5]

[1,2,3,4,5] *Department of Information Technology, VR Siddhartha Engineering College*

***Abstract*—Speech is the most desirable medium of communication between humans. Emotion recognition from speech has emerged as an important research area in the recent past. It plays a vital role in the field of Human Computer Interaction (HCI). An Emotion is a mental and physiological state associated with a wide variety of feelings, thoughts and behaviour. This paper deals with two modules of Speech Emotion Recognition (SER), named Feature extraction and Classification of Emotions. Feature extraction is based on partitioning speech into small intervals known as frames. Selection of features plays an important role in classifying emotions based on which the performance of speech emotion recognition depends. The major features extracted here: Prosodic features including energy, pitch, formants and Spectral features including MFCC and LPCC. To classify emotions several classifiers are available like SVM (Support Vector Machine), k-nearest neighbour, HMM (Hidden Markov Model). This paper discusses the results of SVM classifier on the features extracted from the IITKGP-SESC. Also the comparative study among different classifiers is performed.**

Keywords—Speech Emotion Recognition (SER) system, Feature extraction, MFCC, LPCC, Pitch, Energy, Support Vector machine (SVM)

I. Introduction

Emotional speech recognition is an area of great interest for human-computer interaction. The system must be able to recognize the user's emotion and perform the actions accordingly. It is essential to have a framework that includes various modules performing actions like speech to text conversion, feature extraction, feature selection and classification of those features to identify the emotions. The classifications of features involve the training of various emotional models to perform the classification appropriately. Another important aspect to be considered in emotional speech recognition is the database used for training the models. Then the features selected to be classified must be salient to identify the emotions correctly. [7]The integration of all the above modules provides us with an application that can recognize the emotions of the user and give it as input to the system to respond appropriately. Emotions play an extremely important role in human mental life. It is a medium of expression of one's perspective or his mental state to others. Speech Emotion Recognition (SER) can be defined as the extraction of the emotional state of the speaker from his or her speech signal.

There are few universal emotions [2] including: Neutral, Anger, Surprise, Fear, Happiness, Sadness.

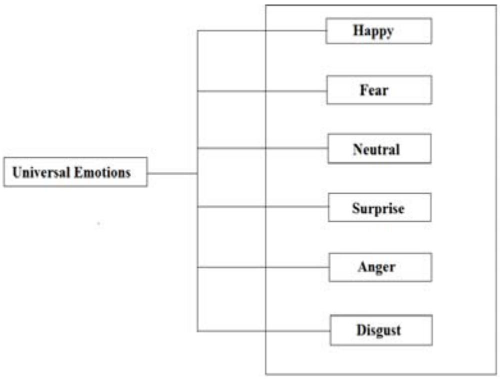

Fig 1. Universal Emotions

The importance of automatically recognizing emotions in human speech has grown with increasing role of spoken language interfaces in the field of human machine interaction to make the human machine interface more efficient. It can also be used for in-car board system where information of the mental state of the driver maybe provided to initiate his/her safety. In automatic remote call centre, it is used to timely detect customers dissatisfaction [2]. In E-learning field, identifying student's emotion timely and making appropriate treatment can enhance the quality of teaching.

Both spectral and prosodic features can be used for speech emotion recognition because both of these features contain the emotional information. Linear predictive cepstrum coefficients (LPCC) and Mel-frequency cepstrum coefficients [3] (MFCC) are some of the spectral features. Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features used to model the different emotions.

II. State Of Art

Speech emotion recognition is an important part in emotion recognition. 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

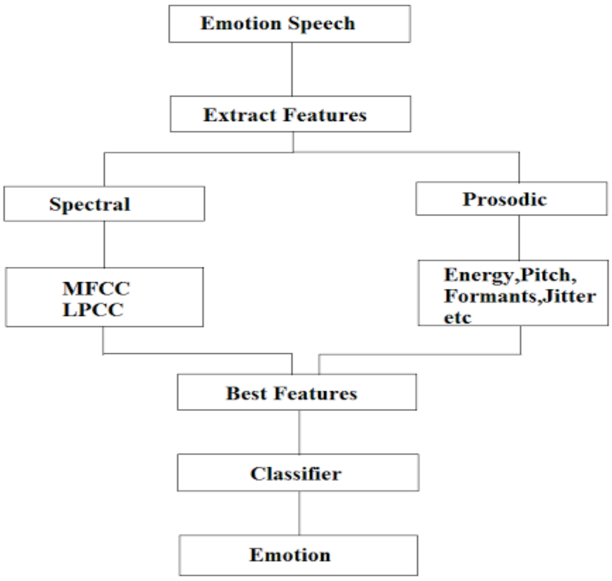

Fig 2. Emotion Recognition from Speech

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, emotions play an important role. Research has long been done on [1] emotion in the fields of psychology and physiology many persons like Vishali M. Chavan, V. V. Gohokar, Aastha Joshi and Rajneet Kaur (Punjab, India). More recently it is the subject of attention by engineers. It is the most important application is in intelligent human-machine interaction. In today's human-machine interaction systems, machines can recognize "what is said" and "who said it" using speech recognition and speaker identification techniques. If it is equipped with emotion recognition techniques, machines can also know "how it is said" to react more appropriately, and make the interaction more natural. Other applications of automatic emotion recognition include psychiatric diagnosis, intelligent toys, and lie detection [1]. Today there are many different algorithms which were used for various signal processing applications.

a. Speech Emotion Recognition System

The general architecture for SER system has three steps shown in Fig. 2 [2]:

i. A speech processing system extracts some appropriate quantities from signal, such as pitch or energy

ii. These quantities are summarized into reduced set of features.

iii. A classifier learns in a supervised manner with example data how to associate the features to the emotions.

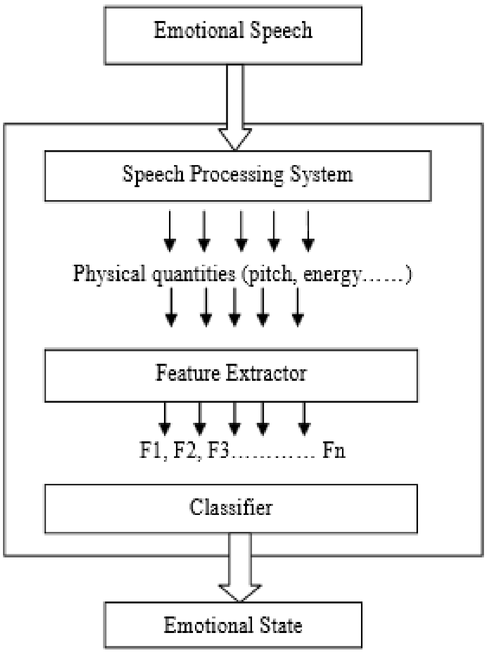

Fig 3. A Basic Outline of Speech Emotion Recognition System

b. Feature Extraction

The speech signal contains a large number of information which reflects the emotional characteristics. So in the research of speech emotion recognition, the most important thing is that how to extract and select better speech features with which most emotions could be recognized. In recent researches, many common features are extracted, such as speech rate, energy, pitch, formant, and some spectrum features,[3] for example Linear Prediction Coefficients (LPC), Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative.

1)Energy and Related Features: The Energy is the basic and most important feature in speech signal. We can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max value, variance, variation range, contour of energy.

2)Pitch and Related Features: We calculate the value of pitch frequency in each speech frame, and obtain the statistics of pitch in the whole speech sample. These statistical values reflect the global properties of characteristic parameters. Each Pitch feature vector has the same 19 dimensions as energy.

3)*Linear Predictive Cepstrum Coefficients (LPCC):* LPCC embodies the characteristics of particular channel of speech, and the same person with different emotional speech will have different channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech. The computational method of LPCC is usually a recurrence of computing the linear prediction coefficients (LPC), which is according to the all-pole model

4)*Mel Frequency Cepstrum Coefficients (MFCC):* MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages.

5)*Jitter:* It is defined as varying pitch in the voice which causes rough sound. It is a measure of period-to-period fluctuations in fundamental frequency. It adds a small amount of noise to a numeric vector.

6)*Shimmer:* It describes varying loudness in the voice. It is a measure of the period-to-period variability of the amplitude value.

7)*Formants:* The peaks that are observed in a sound spectrum are called formants.

8)*Mel Energy Spectrum Dynamic Coefficients (MEDC):* MEDC extraction process is similar with MFCC. The only one difference in extraction process is that the MEDC is taking logarithmic mean of energies after Mel Filter bank and Frequency wrapping, while the MFCC is taking logarithmic after Mel Filter bank and Frequency wrapping. After that, we also compute 1st and 2nd difference about this feature.

III. DATABASES

The Telugu Database named IITKGP-SESC developed by IIT Kharagpur is used in our experiment. This database is widely used in emotional speech recognition. The database is well annotated and provided with different seasons. The sample taken for experimenting consists of totally 360 speech samples. The Telugu Database contains speech samples spoken by different speakers in different emotions. We use happy, disgust, surprise, anger, surprise and neutral emotions in our recognition system.

IV. METHODOLOGY

a. *MFCC Feature Extraction:*

The algorithm used for obtaining mel frequency cepstrum coefficients (MFCCs) from a speech signal is as follows:

1. Pre-emphasize the speech signal.
2. Divide the speech signal into a sequence of frames with a frame size of 20ms and a shift of 5ms. Apply the hamming window over each of the frames.
3. Compute the magnitude spectrum for each windowed frame by applying DFT.
4. Mel spectrum is computed by passing the DFT signal through a mel filter bank.
5. DCT is apply to the log mel frequency coefficients to derive the desired MFCCs.

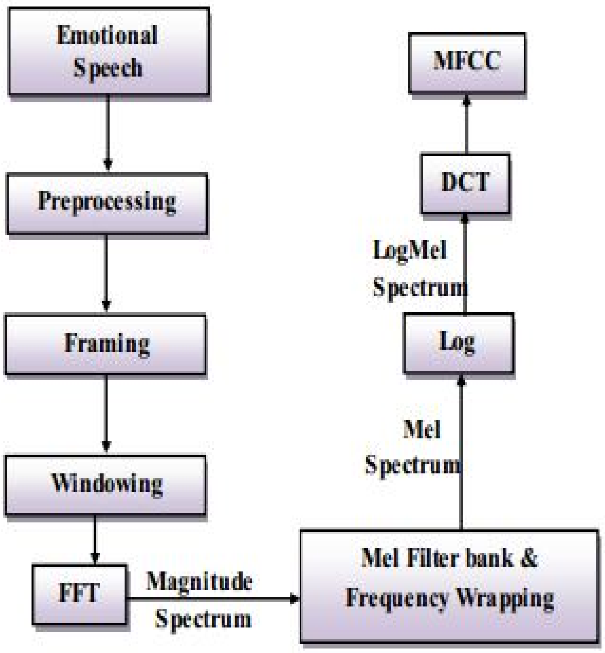

Fig 4. Block Diagram of MFCC Feature Extraction

As shown in Fig. 4, the MFCC feature extraction process [5] consists of the following steps:

1) *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.
2) *Framing:* It is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. 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 ms.
3) *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
4) *FFT:* Fast Fourier Transform (FFT) algorithm is ideally used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain.
5) *Mel Filterbank and Frequency wrapping:* The mel filter bank consists of overlapping triangular filters with the cutoff frequencies determined by the center frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale.
6) *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

7) *Take Discrete Cosine Transform:* It is used to orthogonalize 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.

b. *Mel Energy Spectrum Dynamic Coefficients(MEDC) Feature Extraction[5]*

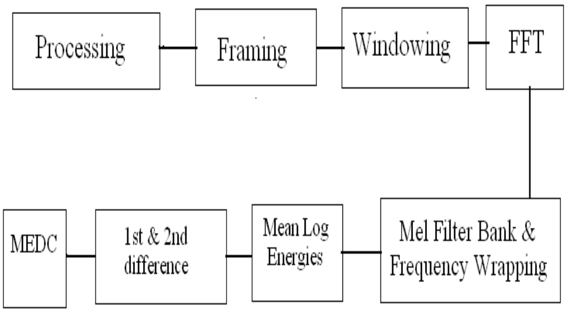

Fig 5. MEDC Feature Extraction

c. *SVM Classification Algorithm*

The Support Vector Machine is used as a classifier for emotion recognition. The SVM is used for classification and regression purpose. It performs classification by constructing an N- dimensional hyper planes that optimally separates the data into categories. [6]The classification is achieved by a linear or nonlinear separating surface in the input feature space of the dataset. Its main idea is to transform the original input set to a high-dimensional feature space by using a kernel function, and then achieve optimum classification in this new feature space. Since SVM is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers. Thus we adopted the support vector machine to classify the speech emotion in this paper.

The performance of speech emotion recognition system is influenced by many factors, especially the quality of the speech samples, the features extracted and classification algorithm. The results obtained for MFCC Feature Extraction are as shown in Fig 6 and 7. The Results obtained for Formants Feature as shown in Fig 8.

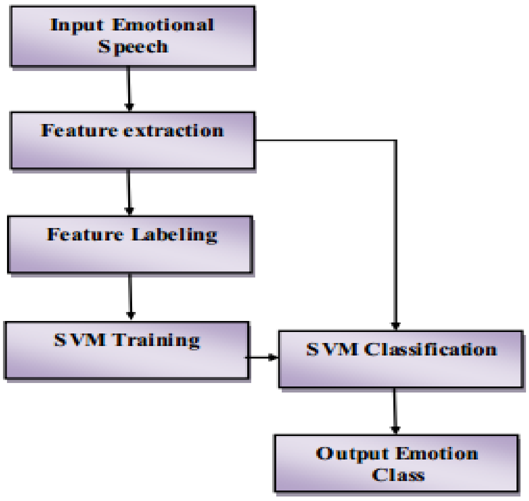

Fig 5. Block Diagram of Speech Emotion Recognition System using SVM classification.

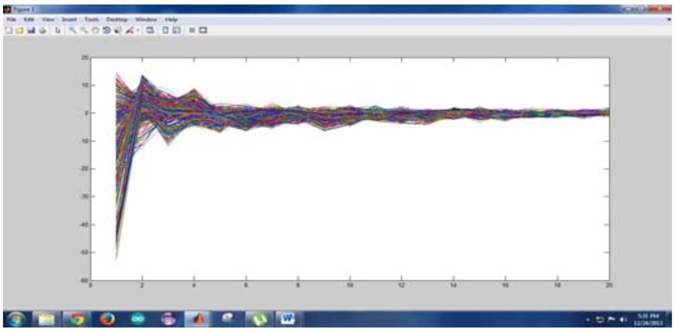

Fig 6. MFCC waveform for Anger

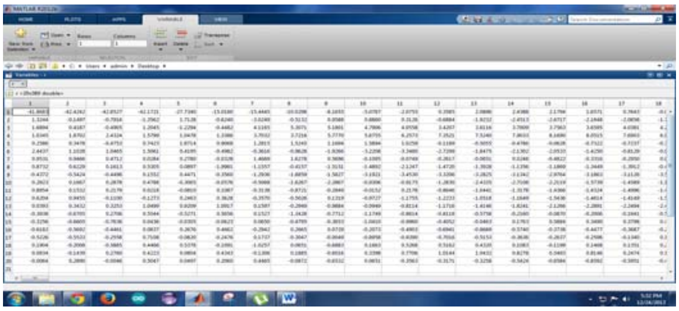

Fig 7. MFCC coefficients for Anger

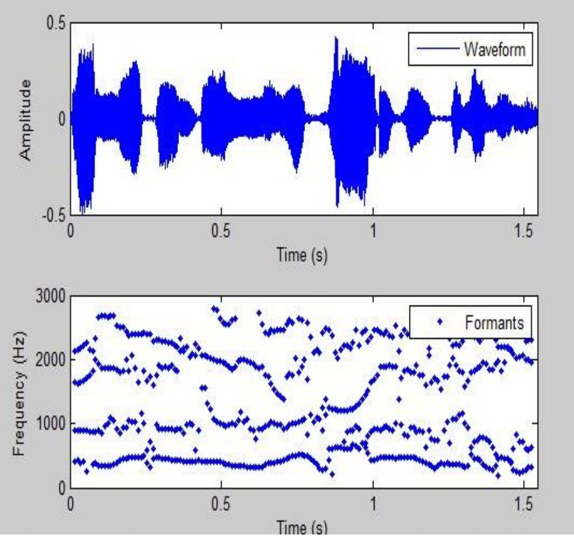

Fig 8. Formants

V. Training Models

The Telugu Emotion database contains 360 speech files for five emotion classes. We choose three from it. Emotion classes sad, happy, neutral are having 62, 71, and 79 speech utterance respectively. We combine different features to build different training models, and analyse their recognition accuracy. Table1 shows different combination of the features for the experiment.

TABLE I
Different Combination of Speech Feature Parameters

Training model	Combination of feature parameters
Model 1	Energy+Pitch
Model 2	MFCC+MEDC
Model 3	MFCC+MEDC+LPCC
Model 4	MFCC+MEDC+Energy
Model 5	MFCC+MEDC+Energy+Pitch

VI. Experimental Results

We use libsvm tool in Matlab to do the cross validation of models and analyse results. With the experiment, we pick pitch, energy, MFCC, its first-order difference, second-order difference, and MEDC as well as its first-order and second-order difference and their combination to extract features. For each emotion, we divide these speech utterances into two subsets as training subset and testing subset. The number of speech utterances for emotion as the training subset is 90%, and 10% as the test subset. Table2 shows the models cross validation rate and recognition rate based on Telugu database.

TABLE II
The Recognition Rate and Cross Validation Based on German Model

Training Model	Features Combination	Cross Validation Rate	Recognitio n Rate
Model 1	Energy+Pitch	66.6667%	33.3333%
Model 2	MFCC+MEDC	90.1538%	86.6667%
Model 3	MFCC+MEDC+L PCC	72.5275%	86.6667%
Model 4	MFCC+MEDC+E nergy	95.0549%	91.3043%
Model 5	MFCC+MEDC+E nergy+Pitch	94.5055%	90%

TABLE III
The Recognition Rate and Cross Validation Based on man Model

Training Model	Features Combination	Cross Validation Rate	Recognition Rate
Model 2	MFCC+MEDC	88.6168%	80.4763%
Model 4	MFCC+MEDC+Energy	95.1852%	95.0874%

As is shown at Table 2 and Table 3, different features combination results in different recognition accuracy rate. To the Telugu Database, the feature combination of Energy and Pitch has the worst recognition rate, which can only recognize one emotional state. That may because these two are simple prosodic features with few numbers of dimensions. The accuracy rate for the feature combination of MFCC and MEDC is higher compared with Model 1. It can better recognize three standard emotional states. We also add the LPCC feature, but the performance of the model becomes lower which may result from the feature redundance. [4]The best feature combination is MFCC+MEDC+Energy, for which the cross validation rate can be as high as 95% for nonreal-time recognition. The reason for this high performance is that it contains prosodic features as well as spectrum features, and the features have excellent emotional characters. The cross validation rate is as high as 95%, and the recognition accuracy rate is also around 95%.

VII. Conclusions And Future Work

We can conclude that, different combination of emotional characteristic features can obtain different emotion recognition rate, and the sensitivity of different emotional features in different languages are also different. So we need to adjust our features to different various corpuses. As can be seen from the experiment, the emotion recognition rate of the system which only uses the spectrum features of speech is slightly higher than that only uses the prosodic features of speech. And the system that uses both spectral and prosodic features is better than that only uses spectrum or prosodic features. Meanwhile, the recognition rate of that use energy, pitch, LPCC MFCC and MEDC features is slightly lower than that only use energy, pitch MFCC and MEDC features. This may be accused by feature redundance. To extract the more effective features of speech and enhance the emotion recognition accuracy is our future work. More work is needed to improve the system so that it can be better used in real-time speech emotion recognition.

REFERENCES

[1] Chavan, V.V. Gohokar, Speech Emotion Recognition by using SVM-Classifier, International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-1, Issue-5, June 2012

[2] Aastha Joshi, Rajneet Kaur, A Study of Speech Emotion Recognition Methods, International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 2, Issue. 4, April 2013

[3] Yixiong Pan1, Peipei Shen2, Liping Shen3, Feature Extraction and Selection in Speech Emotion Recognition , Department of Computer Technology ,Shanghai JiaoTong University,Shanghai, China.

[4] Yixiong Pan, Peipei Shen and Liping Shen, Speech Emotion Recognition Using Support Vector Machine, Department of Computer Technology Shanghai JiaoTong University, Shanghai, China.

[5] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, Speech Emotion Recognition Using Support Vector Machine , VIT, Pune, India.

[6] http://www.ukessays.com/essays/computer-science/speech-emotion-recognition-using-support-vector-machine-computer-science-essay.php

[7] Bhoomika Panda, Debananda Padhi2, 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, Volume 2, Issue 3, March 2012.