Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 49 (2015) 50 – 57

Procedia Computer Science

ICAC3'15

Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) techniques

Rahul B. Lanjewar[a], Swarup Mathurkar[b], Nilesh Patel[c]

[a]*Lecturer, Department of Electronics, Dr. Babasaheb Ambedkar College of Engineering and Research, Nagpur-441110, Maharashtra, India*
[b]*Assistant Professor, Department of EXTC, Government College of Engineering, Amravati-444604, Maharashtra, India*

Abstract

The kinship between man and machines has become a new trend of technology such that machines now have to respond by considering the human emotional levels. The signal processing and machine learning technologies have boosted the machine intelligence that it gained the capability to understand human emotions. Incorporating the aspects of speech processing and pattern recognition algorithms an intelligent and emotions specific man-machine interaction can be achieved which can be harnessed to design a smart and secure automated home as well as commercial application. This paper emphasizes on implementation of speech emotion recognition system by utilizing the spectral components of Mel Frequency Cepstrum Coefficients (MFCC), wavelet features of speech and the pitch of vocal traces. The different machine learning algorithms used for the classification are Gaussian Mixture Model (GMM) and K- Nearest Neighbour (K-NN) models for the recognition of six emotional categories namely happy, angry, neutral, surprised, fearful and sad from the standard speech database Berlin emotion database (BES) followed by the comparison of the two algorithms for performance analysis which is supported by the confusion matrix.

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)

Keywords: Speech Features; Emotion; MFCC; wavelet; pitch; K-NN; GMM; Database

1. Introduction

The demanding trends of automated systems have pushed the extent of recognition system to consider the concise commands rather to run only on the set of instructions. The idea to design a speech emotion recognition system is one of the scopes of the speaker identification. The acoustic features can not only identify the speaker but

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)
doi:10.1016/j.procs.2015.04.226

also the nature of utterances with the intent to achieve the maximum natural interaction in an automated system. This generalized idea can also be used in the spoken dialogue system e.g. at call centre applications where the support staff can provide better service by adjusting their temperament to the emotion of the caller.

It is a human instinct to recognize the emotions by analyzing the psycho-visual appearances as well as by the vocal traces. The same artificial function can be embedded in the machines to emulate this natural tendency of human by employing speech processing and pattern recognition concepts. The earlier research on speech signals open the doors to exploit the acoustic properties that can deal with the emotions. On the other side the signal processing toolbox available in MATLAB software and the algorithms (e.g. GMM, K-NN, HMM) on pattern recognition at helm may turn fruitful to achieve the goal of recognizing emotions from speech. This paper focuses on implementation of a speech emotion recognition system by extracting the different speech features from the Berlin Emotion Speech Database (BES) with intent to perform a comparative analysis on the performance of classifiers Gaussian Mixture Model (GMM) and the K nearest neighbour (K-NN).

2. Literature

In the two experiments conducted by Busso *et al*. the emotional pitch contours were compared with neutral speech while in the second experiment the measure of discriminative power of pitch features for the emotions was derived that lead to conclusion that the pitch plays a salient role to understand the emotions. It is also observed in an work that sentence level pitch features outperforms the voiced-level pitch statistics both in accuracy and robustness [5]. Farooq *et al.* analysed the wavelet packet transform's multi-resolution capabilities can derive superior spectral features as compared to MFCC features sets and which can be used to improve in recognition of unvoiced phonemes and stop statements[12]. The biomedical research approach by Yao *et.al* developed the Bionic Wavelet transform (BWT) which dealt with concentrating on the energy distribution to retain and introduce an active control mechanism of auditory system to adjust the wavelet transform to enhance the sensitivity and selectivity [13]. The spectral coefficients derived from Linear Prediction Cepstrum Coefficients (LPCC), Mel Frequency Cepstrum Coefficients (MFCC) and formants represent the vocal tract information. MFCC are robust features for any speech oriented tasks, Koolagudi *et.al*. The more clinical research by Wu *et al.* on recognition system gave a large variety of classifiers to map the test data by using a new approach of multiple classifiers Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) in a single recognition system. But the results of GMM is more discriminative than HMM for Berlin Emotional Speech Database (BES) and it measure up to 76% than of 71% of HMM, 67% of K-NN and 55% of FFNN. The classifier GMM was used in two stages: the first stage dealt with the classification of high, low and neutral emotions, while the second stage classifies the emotions of same categories. The experiment resulted in a novel result which adamants the improvement in recognition [22]. A natural database can boost the results of recognition at the same speaker specific information always plays an important role in such a way that usage of same speaker for the training and testing reduces the generality. Thus, for the discriminative results the database should have large speaker text prompts and natural [14]. The review by Ververidis et al. on the available 32 database derived that not more than 50% classification can be achieved for the four basic emotions in automated emotion recognition system, simulated emotions are easy to classify as compared to natural emotions and the results usual of emotion recognition in the descending order of their easier classification are anger, sadness, happiness, fear, disgust, joy, surprise and boredom. The purpose to implement emotion recognition systems is to blend the machines with emotion-related knowledge so that human computer communication can be enhanced and furthermore the user's experience will become more satisfying. The paper is organized in the following manner, first the system development and its resources of speech features and their extraction has been discussed in literature followed by the two classifiers GMM and K-NN for selection based on the emotional relevance to compare their performance.

3. Speech Emotion Recognition System

The technique of Speech emotion recognition is very much similar to the pattern recognition task. The modular flow of signal data in various stages helps to understand the stages involved in the speech emotion recognition called the pattern recognition cycle Fig. 1.The pre-processed speech signals are used for feature extraction in which the

training data is cut into overlapping frames. The features extracted are based on extraction of pitch, MFCCs and Wavelet domain information. The next stage is the feature selection process which limits the coefficients of feature set to reduce the curse of dimensionality while the relevant features are retained as it is. The next stage is to pass the selected features to the classifiers. In our study we used two types of classifiers in the form of Gaussian Mixture Model (GMM) and K-Nearest Neighbour (K-NN) about which more details are discussed in the following sections.

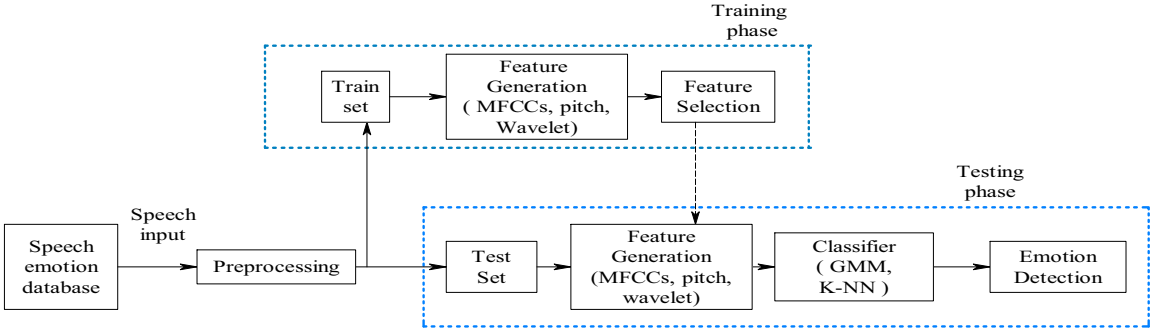

Fig.1. Speech Emotion Recognition System

3.1 Berlin Emotion Speech Database (BES)

The Berlin emotion database is most often used database in the speech processing community. It is an acted emotional content database created by the audio recordings of ten actors, five male and five female. The recordings portray emotions from the following set: happiness, anger, disgust, fear, sadness, surprise and neutral. The database contain approximately 800 sentence records of y 20 participants, the amount of sentences on the database is reduced to around 500 samples which have a human recognition rate better than 80% and naturalness scores of more than 60%.

3.2 Features Extraction

3.2.1 Mel Frequency Cepstrum Coefficients (MFCC)

Mel is a unit to measure the perception of pitch or frequency of a tone. The Mel-scale is a mapping of the real frequency scale (Hz) to the perceived frequency scale (mels). This mapping is virtually linear below 1 KHz and logarithmic above. Early research on the MFCCs indicate that they are less sensitive to noise compared to other currently used parameters and provide better recognition performance. The speech signal is pre-processed by windowing techniques after which the Discrete Fourier Transform of each framed speech signal is taken to obtain its magnitude spectrum and later frequency wrapping is performed to transform the spectrum into Mel scale in which the triangular filter bank of uniformly spaced is obtained. These filters are multiplied with the magnitude spectra are taken to obtain the MFCCs. In this work 20 filter banks and 22 MFCCs are used for the simulations [7].

3.2.2 Wavelet Features

The detail discussion on wavelet analysis is beyond the scope of this paper and the more complete discussion is presented in [8]. The continuous wavelet transform is defined here. Let $f(t)$ be any square integrable function. The CWT or continuous-time wavelet transform of $f(t)$ with respect to a wavelet $\psi(t)$ is defined as:

$$W(a,b) \equiv \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{a}}\psi^*\left(\frac{t-b}{a}\right)dt \quad (1)$$

Where a and b are real and $*$ denotes complex conjugation. Thus the wavelet transform is a function of two variables. Both the $f(t)$ and $\psi(t)$ belongs to the set of energy signals. The equation (2.1) can be written in compacform by defining $\psi_{a,b}(t)$ as:

$$\psi_{a,b}(t) \equiv \frac{1}{\sqrt{|a|}}\psi^*\left(\frac{t-b}{a}\right) \quad (2)$$

Wavelet functions comprise an infinite set. The different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are. The Haar wavelet is discontinuous

and resembles a step function. It represents the same wavelet as Daubechies db1. In this paper we considered db1 family of wavelets for the speech features extraction.

3.2.3 Pitch Features of speech

To extract the pitch features Subharmonic-to-Harmonic Ratio (SHR) is used. The magnitude of subharmonics with respect to harmonics reflects the degree of deviation from modal voice. The SHR reflects the ratio of amplitudes of harmonics and subharmonics. The algorithm is based on the pitch perception study to determine the perceived speech and SHR. The technique involves synthesis of vowels with alternate cycles through amplitude and frequency modulation, which generates subharmonics with lowest frequency of $0.5F_0$. Generally, when the ratio is smaller than 0.2, the subharmonics do not have effects on pitch perception. As the ratio increases approximately above 0.4, the pitch is mostly perceived as one octave lower that corresponds to the lowest subharmonic frequency. When SHR is between 0.2 and 0.4, the pitch seems to be ambiguous. These findings suggest that pitch could be determined by computing SHR and comparing it with the pitch perception data. The procedure for computing SHR falls in the general category of spectrum compression technique [9].

4. Classification

4.1 Gaussian Mixture Model (GMM)

In this study, a Gaussian Mixture Model approach is proposed where speech emotions are modelled as a mixture of Gaussian densities. The use of this model is motivated by the interpretation that the Gaussian components represent some general emotion dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities.

The Gaussian Mixture Model is a linear combination of M Gaussian densities, and given by the equation,

$$P(\vec{x}|\lambda)= \sum_{i=1}^{M} p_i \, b_i(\vec{x}) \quad (3)$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$, i=1,.....,M are the component densities and p_i, i=1,.....M are the mixture weights. Each component density is a D-dimensional Gaussian function of the form

$$b_i(\vec{x})=\frac{1}{(2\pi)^2|\sum i|^2} \exp\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \sum_i^{-1}(\vec{x}-\vec{\mu}_i)\} \quad (4)$$

where $\vec{\mu}_i$ denotes the mean vector and $\sum_i$ denotes the covariance matrix. The mixture weights satisfy the law of total probability, $\sum_{i=1}^{M} p_i$=1. The major advantage of this representation of speaker models is the mathematical tractability where the complete Gaussian mixtures density is represented by the mean vectors, covariance matrices and mixture weights from all component densities.

The probability density functions of distorted features caused by different emotions are different. As a result, we can use a set of GMMs to estimate the probability that the observed utterance from a particular emotion. This technique involves Maximum Likelihood Estimation by constructing a Bayesian classifier the class-conditional probability density functions need to be determined. The initial model selection can be done for example by visualizing the training data, but the adjustment of the model parameters requires some measure of goodness, i.e., how well the distribution fits the observed data. Data likelihood is a goodness value [10].

Assume that there is a set of independent samples X = $\{x_1, x_2, \ldots \ldots x_N\}$ drawn from a single distribution described by a probability density function p(x;θ) where θ is the PDF parameter list.

The likelihood function

$$\mathcal{L}(\mathcal{X};\theta)=\prod_{x=1}^{N} P(x_N:\theta) \quad (5)$$

tells the likelihood of the data X given the distribution or, more specifically, given the distribution parameters θ. The goal is to find $\hat{\theta}$ that maximizes the likelihood:

$$\hat{\theta}=\arg max_{\theta}= \mathcal{L}(X;\theta) \quad (6)$$

Usually this function is not maximized directly but the logarithm

$$\mathcal{L}(X;\theta)=\ln \mathcal{L}(X;\theta)=\sum_{n=1}^{N} lnp(x_N:\theta) \quad (7)$$

called the log-likelihood function which is analytically easier to handle. Because of the monotonicity of the logarithm function the solution to Eq. 8 is the same using $\mathcal{L}(X;\theta)$.

Thus the implementation steps for GMM classification are described as follows:

1. Initialize parameters
2. Expectation step: Compute the posterior probability for i=1,..... n, k=1,....K.

$$P_{i,k}= \frac{a_k^{(r)} \emptyset\,(x_i\, \mu_k^{(r)}, \Sigma_k^{(r)})}{\sum_{k=1}^{(r)} a_k^{(r)} \emptyset\,(x_i\, \mu_k^{(r)}, \Sigma_k^{(r)})} \quad (8)$$

3. Maximization step

$$a_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k}}{n} \quad (9)$$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k}\, \mathrm{X}}{\sum_{i=1}^{n} P_{i,k}} \quad (10)$$

$$\mu_k^{(r+1)} = \frac{\sum_{k}^{(r+1)} P_{i,k}(x_i\, \mu_k^{(r+1)})(x_i\, \mu_k^{(r+1)})\mathrm{t}}{\sum_{i=1}^{n} P_{i,k}} \quad (11)$$

4. Repeat steps 2 and step3 until convergence.

4.2 K- Nearest Neighbour (K-NN)

A general version of the nearest neighbour technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbour rather than on only it's on single nearest neighbour. If the costs of error are equal for each class, the estimated class of an unknown sample is chosen to be the class that is most commonly represented in the collection of its K nearest neighbours [11].

Let the *k* neighbours nearest to *y* be $N_k(\mathrm{Y})$ and c(*z*) be the class label of *z*. The cardinality of $N_k(\mathrm{Y})$ is equal to *k* and the number of classes is *l*. Then the subset of nearest neighbours within class $j \in \{1, \ldots, l\}$ is:

$$N_k^j(\mathrm{Y}) = \{\mathrm{Z} \in N_k(\mathrm{Y}) : c(z) = j\} \quad (12)$$

$$j^* \in \{1, \ldots . l\} \quad (13)$$

The classification result $\{1,...,\} * j \in l$ is defined as the majority vote:

$$j^* = \arg max_j \left| N_k^j(\mathrm{Y}) \right| \quad (14)$$

1. Generate the class for each emotion and store the features in the database.
2. Take the emotion input for each class and train the machine for each class.
3. Sum all the class numbers to assign a number class for unknown class.
4. Find the K neighbour nearest to unlabelled data from training space based on selected distance measured using Euclidian distance.

Let the *k* neighbours nearest to *y* be $N_k(\mathrm{Y})$ and c (*z*) is the class label of *z*. The cardinality of $N_k(\mathrm{Y})$ is equal to *k* and the number of classes is *l*. Then the subset of nearest neighbours within class $j \in \{1, \ldots, l\}$ is:

$$N_k^j(\mathrm{Y}) = \{\mathrm{Z} \in N_k(\mathrm{Y}) : c(z) = j\} \quad (15)$$

$$j^* \in \{1, \ldots . l\} \quad (16)$$

The classification result $\{1,...,\} * j \in l$ is defined as the majority vote:

$$j^* = \arg max_j \left| N_k^j(\mathrm{Y}) \right| \quad (17)$$

5. Experimental Analysis

The paper focuses on the detection of happy, fear, angry, neutral, sad and surprise emotions from the extracted speech features (MFCC, Pitch, Wavelet). The extent of results is totally dependent on the type of speech database employed in training and testing phase. In this work GMM and K-NN are used as classifiers with Wavelet features, spectral MFCC features and pitch features of voice has been used as speech features. Thus from the results obtained during the testing phase after utilizing resources of speech features (MFCC, wavelet, pitch), database

(BES) and classifiers (GMM, K-NN) a comparative as well as cross validation approach is followed by adopting the evaluation parameters in the form of recognition accuracy, precision rate and F-measure on BES for all the two classifiers supported by the confusion matrix.

5.1 Recognition accuracy

This measure signifies the recognition accuracy in percentage for each known test speech input to the total trained emotional speech data [6].

$$Accuracy = \frac{Correctly\ detected\ Emotions\ inputs}{Total\ trained\ emotions\ inputs} \text{ X } 100\% \quad (18)$$

The accuracy for each classifier for the six emotions was calculated on the basis of above relation. It was calculated for both of Berlin emotional database (BES) and a recorded non-standard database.

Table 1: Recognition Accuracy (%) for Berlin Emotional Speech Database (BES)

Emotion / Classifier	Angry	Happy	Sad	Neutral	Surprise	Fear
GMM	92	67	89	73	25	50
K-NN	72	90	44	54	25	25

5.2 Confusion Matrix

The confusion matrix of the classifier for speech emotion recognition system describes the confusion to choose the best correct emotion in the testing phase. It depicts the confusion faced by a particular classifier in selection the correct pattern among the trained patterns of emotion features which have similarities in their respective feature pattern.

Table 2: Confusion Matrix for K-NN Classifier

Responded / Presented	Angry	Happy	Sad	Neutral	Surprise	Fear
Angry	72	-	-	28	-	-
Happy	-	90	-	10	-	-
Sad	12	22	44	-	-	22
Neutral	28	-	-	54	4	14
Surprise	50	-	-	-	25	25
Fear	41	-	25	-	-	34

Table 3: Confusion Matrix for GMM Classifier

Responded / Presented	Angry	Happy	Sad	Neutral	Surprise	Fear
Angry	92	-	-	-	8	-
Happy	-	67	-	-	33	-
Sad	-	-	89	11	-	-
Neutral	-	19	-	73	-	8
Surprise	25	25	-	-	25	25
Fear	-	-	-	-	50	50

5.3 Precision Rate

It is defined as the ratio of correctly recognized emotions for each class to the correctly recognized emotions for all the classes [6].

$$Precision\ Rate\ (\%) = \frac{Correctly\ recognized\ emotions\ for\ each\ class}{Correctly\ recognized\ emotions\ for\ all\ the\ classes} \quad (19)$$

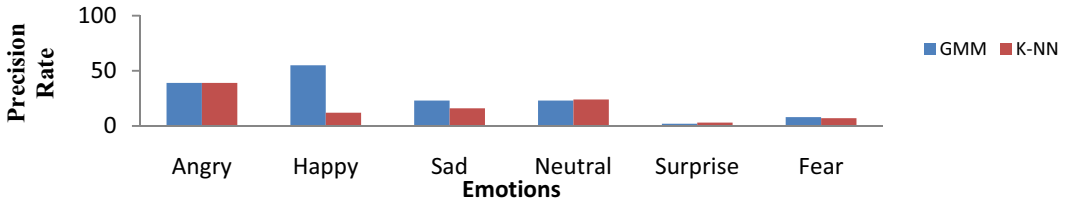

Fig.2. Precision Rates (%) for each Class of Emotions under Different Classifiers

5.4 F-Measure

The F-Measure is the merit of combination of precision rate and accuracy. It is adopted from the work of Natalampiras *et al*. in which the performance of the implementation was evaluated from this factor to obtain the overall performance of the system in terms of correct results *i.e.* by not considering the wrong recognition observations and is given by [6]:

$$F - Measure = \frac{2*Accuracy*Precision}{Accuracy+Precision} \quad (20)$$

The F-Measure of the proposed system as shown in Fig. 3 for the two classifiers displays the overall performance of the system in terms of its correctness.

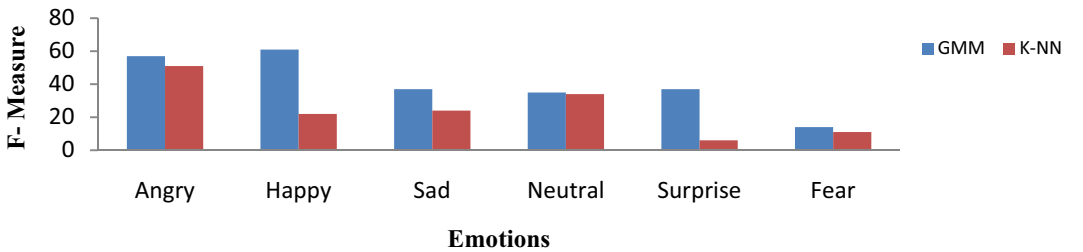

Fig.3. F-Measure of the system for the GMM and K-NN

All the evaluation merits depict the superiority of GMM classifier for all the emotions as compared to K-NN classifier. The performance of K-NN classifier has slumped in its results.

Discussion and Conclusions

From the practical perspective and the obtained results it can be said that the implemented model is appropriate to use in machine learning techniques which is justified by the statistics at the same the results promising in recognizing the emotions for the various test speech features of database which helps to draw a comparative conclusion on their performance. The GMM technique has shown best results of the two classifiers by recognizing ‘angry’ with highest rate of 92% for and minimum recognition rate of 25% recognition rate for ‘surprise’ emotion. The result is also supported by its confusion matrix which shows that minimum confusion occurred to detect between ‘surprise-happy’ emotions. The speed of computation as well as the recognition rates for

K-NN has driven the result to a new direction. The recognition rates for the K-NN technique in detection of 'happy' emotion with 90% rate and lowest rate for "fear' and for the 'surprise' emotion it is 50% justified by the confusion matrix. However the confusion matrix results of K-NN technique to recognize between happy and neutral emotions adds similar relevance to natural techniques since there are circumstances in which even humans get confuse to judge between 'happy' and 'normal' emotions. The GMM technique erred in recognizing 'angry' emotion among the rest of the five emotions. However, the evaluation merits of Precision and F-Measure shows dominance of GMM technique and its robustness in speech emotion recognition system. Last but not least to implement more robust and efficient emotion recognition technique on real time applications then the two classification techniques can be fused together to recognize the emotions very effectively because of their dominance for particular type of emotions as such GMM for 'angry' and 'sad' emotions detection, K-NN for 'happy' as well as 'angry' emotions. The speed of computation is fast for K-NN classifier makes it as one of the optional techniques that can be used widely if time constraint is critical. The time of computation increased for GMM classifier when the number of speech features increased in training phase.

Future Scope

The trial and analysis approach of this work leads us to different ideas for future work and system improvements especially on the database. The biggest challenge will be to create a real life emotion database so that the research will get closer to the real-life. If the recognition process for the emotion detection will get improved then it could be useful in day to day life for the various applications. Also there is a need of a new technique in which the more classifiers can fused together to serve the best recognition rates.

References

1. C. Busso, S. Lee and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", *IEEE Trans. on Audio, Speech and Language processing*, Vol. 17, No. 4,2009, pp 582-596.
2. O. Farooq and S. Datta, 'Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition', *IEEE Signal Processing Letters*, Vol. 8, No. 7, 2001, pp.196-198.
3. Jun Yao and Yuan-Ting Zhang, 'Bionic Wavelet Transform: A New Time–Frequency Method Based on an Auditory Model', *IEEE Trans*. on *Biomedical Engineering*, Vol. 48, No. 8, 2001, pp.856-863.
4. Shashidhar G. Koolagudi ·K. Sreenivasa Rao, 'Emotion recognition from speech: a review', Int'l Journal on Speech Technology, 2012, pp.99–117.
5. Chang-Hsein Wu and Wei-Bin Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic and Semantics Labels, *IEEE Trans. on Affective Computing*, Vol 2, No.1, 2011, pp.567-569.
6. Stavros Ntalampiras and Nikos Fakotakis,' Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition', *IEEE Trans. on Affective Computing,* Vol. 3, No. 1, 2012, pp. 116-125.
7. Rahul. B. Lanjewar, D. S. Chaudhari, "Speech Emotion Recognition: A Review", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 2, March 2013, pp. 68-71.
8. Ruhi Sarikaya, Brian L. Pellom and John H.L. Hansen, 'Wavelet Packet Transform Features with Application to Speaker Recognition', 1998, pp.912-915.
9. Xuejing Sun, 'Pitch Determination and Voice Quality Analysis using Subharmonic-To-Harmonic Ratio', Department of Communication Sciences and Disorders, Northwestern University, 1999, pp. 561-563.
10. Moataz M. H. El Ayadi, Mohamed S. Kamel and Fakhri Karray, "Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models', ICASSP, 2007, pp.957-960.
11. Tsang-Long Pao, Wen-Yuan Liao and Yu-Te Chen, 'A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition', Speech Recognition Technologies and Applications, 2008, pp. 550-552.