Emotion Recognition from Audio Signals using Support Vector Machine

M.S. Sinith, Aswathi E., Deepa T. M., Shameema C. P. and Shiny Rajan
Electronics and Communication Engineering
Govt. Engineering College Thrissur, 680009
Email: sinith@gectcr.ac.in, aswathie1894@gmail.com, deepatm003@gmail.com,
shameemacp111@gmail.com, shinyrajan.92@gmail.com

***Abstract*—The purpose of speech emotion recognition system is to differentiate the speaker's utterances into four emotional states namely happy, sad, anger and neutral. Automatic speech emotion recognition is an active research area in the field of human computer interaction (HCI) with wide range of applications. Extracted features of our project work are mainly related to statistics of pitch and energy as well as spectral features. Selected features are fed as input to Support Vector Machine (SVM) classifier. Two kernels linear and Gaussian radial basis function are tested with binary tree, one against one and one versus the rest classification strategies. The proposed speaker-independent experimental protocol is tested on the Berlin emotional speech database for each gender separately and combining both, SAVEE database as well as with a self made database in Malayalam language containing samples of female only. Finally results for different combination of the features and on different databases are compared and explained. The highest accuracy is obtained with the feature combination of MFCC +Pitch+ Energy on both Malayalam emotional database (95.83%) and Berlin emotional database (75%) tested with binary tree using linear kernel.**

***Keywords*—Speech, Emotion, Automatic Emotion Recognition, Pitch, Energy, MFCC, SVM**

I. INTRODUCTION

The relevance of emotion recognition from speech has increased in recent years to upgrade the efficiency of human-machine interactions [1]. 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.

Emotion recognition is a very complex and a difficult job to achieve with very high accuracy. Researchers have conducted a number of studies on different features [1] that influence the emotion deeply. Still there does not exist a perfect feature set that characterize the emotion correctly. This is because of the factors like different speakers, sentences, speaking styles, speaking rates and different languages .The same speech sample may show different emotions in its different portions and it is very difficult to differentiate these portions .Also speaking style of the person is greatly influenced by his age, culture and environment. Emotions occurring in spontaneous speech seem to be more difficult to recognise compared to acted speech. Widely used features [2] include both spectral features such as LPC, LPCC, MFCC [3] and its derivatives and prosodic features such as pitch and energy [4],[5],[6].

Many databases were built for emotion recognition research such as Emo-DB (Berlin Emotional Database) [7] DES (Danish Emotional Speech) that is Danish Corpus, SES (Spanish Emotional Speech) that is Spanish Corpus etc. Various types of classifiers have been used for the task of speech emotion recognition like HMM, GMM, SVM, artificial neural networks (ANN), k-NN and many others [8],[9],[10]. In fact, there has been no accordance on which classifier is the most suitable one for emotion classification. Each classifier has its own advantages and limitations.

In this system, three databases are being used - Berlin Emotional Database [7], SAVEE, an English database and a self made database in Malayalam to train and test the system. The extracted features include MFCC and its derivatives, energy and its derivatives and pitch. The system is trained and tested using SVM classifier [11],[12] built in Binary tree,one against one and one versus the rest methods using both linear and RBF kernel. A comparison is done on the recognition rate of different combination of speech features in each of the classifiers for three databases.

The paper is organized as follows: Section II describes the databases used in the experiments. Section III introduces implementation of the system explaining the feature extraction and the Support Vector Machine algorithm in detail. Experiments and the results obtained are mentioned in the section IV. Section V concludes this paper.

II. EMOTIONAL SPEECH DATABASE

To date, several studies [12],[5],[19] on this topic employed on Emo-DB,DES,SES and [5],[19] used databases which were self built. In particular, for the paper[19], overall accuracy obtained with Emo-DB is 63.5% (five classes) and 67.6% on DES(seven classes) dataset using ensemble of SVM with 10 fold cross-validation. Moreover, the paper presented a self built database based on an animation movie which obtained

978-1-4673-6670-0/15/$31.00 ©2015 IEEE

an accuracy of 77.5%(4 classes) and 66.8%(5 classes). The work is based on three databases-Emo-DB which is a German database, SAVEE which is an English database and our self made database in Malayalam.

The Berlin Database consists of 535 speech samples in total of 10 different sentences, which contain German speech samples related to emotions such as anger, disgust, fear, joy, sadness, surprise and neutral, spoken by native professional actors (five males and five females). The Berlin database was chosen to be used because of the high quality of its recording and its popular use in emotion recognition based research work. The database is used for each gender separately and combining both.

Surrey Audio-Visual Expressed Emotion (SAVEE) database has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, anger, disgust, fear, happiness, sadness, surprise and neutral, 480 British English utterances in total. There are 15 sentences per emotion, chosen from the standard TIMIT corpus.

The self made database in Malayalam consists of a total of 120 samples, by taking 5 samples (5 different sentences) for each emotion from 6 female students, out of which 96 are training samples and remaining are taken as testing samples.

III. SYSTEM IMPLEMENTATION

The emotion recognition system consists of four basic blocks: Emotional speech input, Feature extraction, feature labelling and SVM classification [13] as shown in the figure 1.

Input to the system is wav file of the speech samples in

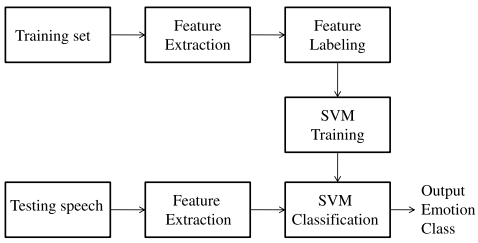

Figure 1: Speech Emotion Recognition

the emotional database. Samples of each database is divided into two groups-training set and test set. Extracted features include pitch, energy and its dynamic coefficients, MFCC and its dynamic coefficients. In the training section, a vector of mean, standard deviation, maximum, minimum and range of the extracted features of the training set are given as input to SVM after labelling each sample by the corresponding emotion class, to generate an output model. In the testing section, feature vector extracted from the wav file of test set is given to SVM classifier and it predicts the emotion class based on the output model.

A. Feature Extraction

Feature extraction is the very intricate and the most important stage in the emotion recognition system as there does not exist an optimum feature set which maximizes the recognition accuracy.

Emotion imparted in speech is represented by large number of features comprised in it and they vary with changes in emotion. Mostly, these features also carry some other information like speaker's identity, gender of the person etc. which may result in erroneous recognition of emotion. Language, speaking rate and style also results in variations in these feature values even if same emotion is produced. Hence, extraction of a general set of features cannot contribute the same accuracy with the different database. Several researches have been conducted on this. Prosodic features form the most widely applied feature type for SER, and Spectral features, on the other hand, also play a significant role in SER as they are related to the frequency content of the speech signal.Some common features are speech rate, energy, pitch, formant, and some spectrum features, such as Linear Prediction Coefficients (LPC), Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative and so on. The paper [5] compares the recognition rate based on the different combinations of the features. The best result(95.1%) is obtained with MFCC+MEDC+Energy on Emo-DB.

The current paper introduces a system which uses the statics of the features pitch, energy and its dynamic coefficients and MFCC and its first and second derivatives for classification. Statistical parameters include mean, standard deviation, minimum, maximum and range. General trend of these features [1] on the four emotional states is as follows: Anger is characterized by the highest value of mean, variance and much wider pitch range and its mean energy is the highest. Happy has much higher mean value of pitch and much wider pitch range. Its energy is also higher. While sadness has its mean pitch value below normal and has narrower pitch range and energy is very low. All these are in comparative to that of neutral.

1) Mel Frequency Cepstral Coefficients: MFCC are widely used in speech recognition and speech emotion recognition studies [3]. A human auditory system is assumed to process a speech signal not in a linear manner. Since lower frequency components of a speech signal contain more phoneme specific information, a non linear Mel scale filter bank has been used to emphasize lower frequency components over higher ones. In speech processing, the Mel frequency cepstrum is a representation of the short term power spectrum of a speech frame using a linear cosine transform of the log power spectrum on Mel frequency scale [14]. Conversion from normal frequency f to Mel frequency m is given by the equation

$$m = 2595 log_{10}(\frac{f}{700} + 1)$$

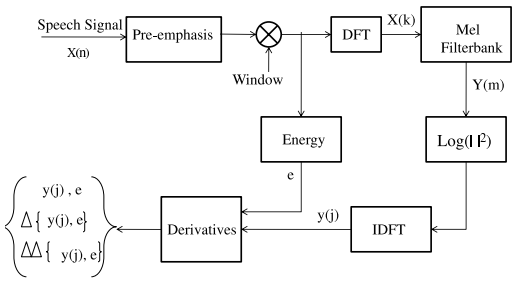

Figure 2: MFCC Feature Extraction

The steps used for computing mel frequency cepstral coefficients (MFCCs) from a speech signal are as follows:

- Pre-emphasize the speech signal.
- Divide the speech signal into a sequence of frames with a frame size of 20 ms and a shift of 10 ms. Apply the hamming window over each of the frames.
- Compute the magnitude spectrum for each frame by applying DFT.
- Compute the mel spectrum by passing the DFT signal through a Mel filter bank.
- DCT is applied to the log mel spectrum to derive the desired MFCCs.

Thirteen filter banks are used to compute 12 MFCC features from a speech frame of 20 ms, with 10 ms overlap each time. The purpose of using MFCCs is to take the non-linear auditory perceptual system into account, while performing automatic emotion recognition.

The MFCC feature vector describes only the power spectral envelope of a single frame, but speech would also have information in the dynamics i.e. what are the trajectories of the MFCC coefficients over time. These are called dynamic coefficients calculated using the formula

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

where d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+N} to c_{t-N} . A typical value for N is 2. Delta-Delta (Acceleration) coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients. Hence a total of 36 coefficients are calculated. For each coefficients, computed the mean, variance, maximum, minimum and range in entire speech.

2) Pitch: Pitch is the perceived fundamental frequency of speech. The pitch signal has information about emotion as it depends on the tension and vibration of the vocal folds [6],[15]. Two features namely the pitch frequency (F_0) and the glottal velocity volume are widely used. The latter denotes the air velocity through the glottis during the vocal fold vibration. F_0, also known as fundamental frequency contains information of the vibration rate of the vocal folds. Many algorithms for estimating the pitch signal exist [16]. Here, a simple auto-correlation analysis [17] is performed at every frame. Statistics [4] such as mean, standard deviation, minimum, maximum and range are calculated in the whole speech sample. As the first, step framing is done with 20 ms frame size and 10 ms overlap. Then autocorrelation sequence of each frame is found and the time lag corresponds to the second largest peak from the central peak gives pitch period and its reciprocal gives pitch frequency.

3) Energy: Energy is another prosodic feature that expresses the variation in emotion. Energy of a speech signal is a representation that reflects the amplitude variations. It contain information of the arousal level [6] of emotions. Short term energy, first and second derivatives of the logarithm of the mean energy are calculated in each frame. Then the mean, standard deviation, minimum, maximum and range [4] of energy in the whole speech is calculated. For extracting energy, framing of speech signal is done with 20 ms duration and 10 ms overlap. After that windowing is done by multiplying each frame with a hamming window in order to keep the continuity of the first and the last points in the frame. Then energy is calculated using the expression,

$$e(n) = \sum_{m=-\infty}^{\infty} (s(m)w(n-m))^2$$

B. Feature Labelling

After the feature vector is extracted, they are labelled according to the emotion class of the corresponding speech sample and stored as a database. This is further loaded for training the SVM.

C. SVM Training And Classification

SVM is the most widely used classifier for pattern recognition applications because of its simplicity in use and the good recognition accuracy it give with limited training data. Idea behind the SVM classifier is that it classifies the data by finding the best hyper plane that separates all data points of one class from those of other class [10]. It thus maximizes the margin between the two classes. There can be two types of data, linearly separable and non-linearly separable.
For linearly separable data, minimize the $||w||$ in order to maximize the margin where 'w' is the normal vector to the hyper plane. So, the optimization problem in this case is:

$$a = \min(\tfrac{||w||^2}{2}) \text{ subject to } \forall k,\ y_k(< w.x_k > +b) \geq 1$$

Where $< w.x_k >$ denote the inner product of w and x_k for the training set of instance-label pairs (x_k, y_k) , $y_k \epsilon \{+1, -1\}$

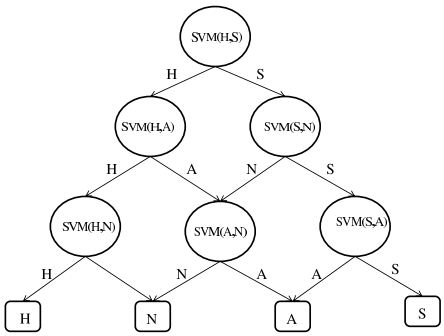

Figure 3: Binary tree classification

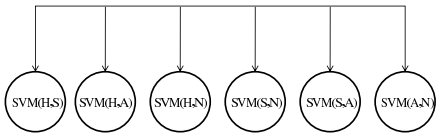

Figure 4: One against one classification

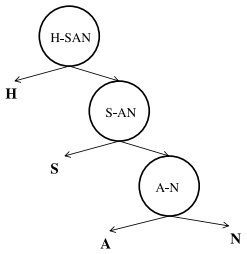

Figure 5: One versus the rest classification

SVM classifies the data so by transforming the original input feature space into a high dimensional feature space using a kernel function where the data can be linearly separated [15] SVM that employs both the linear kernel function and the Radial Basis Kernel (RBF) function is used here [10]. The linear kernel function is given by the formula below,

$$kernel(x, y) = (x.y)$$

The radial basis kernel function is given by the following formula,

$$kernel(x, y) = e^{\frac{-||x-y||^2}{(2\sigma^2)}}$$

The paper [12] describes the performance of SVM classifier for different kernels. Out of which, RBF kernel contributed 79.55% accuracy when implemented in multiclass SVM. Three methods are employed for classification-Binary tree, One against one and One versus the rest and compared their performance.

IV. EXPERIMENTATION AND RESULTS

In the current paper, SVM using linear and Radial Basis Function is implemented in the three methods-Binary tree, One against one and One versus the rest method. Three databases namely Emo-DB, SAVEE and our self built Malayalam database are tested. Database is divided into training set and testing set. During training [8],[10], features are extracted from training set and SVM is trained. In testing, feature vector extracted from an unclassified sample is given to SVM classifier. Based on the model generated by training, SVM predict the output emotion class label. Performance different combination of features is compared. Also performance of different classification methods is also compared. Among all, maximum accuracy (95.83%) is obtained with the self made database tested using binary tree implemented in linear kernel while considering the entire feature set which include pitch, energy and MFCC and their dynamic coefficients.

In Emo-DB, system is implemented for male only, female only and combining both. For each emotion, divide these speech utterances into two subsets as training subset and testing subset. For male 80 training samples were selected ie, 20 samples from each emotion and 40 testing samples ie, 10 samples from each emotion. In the case of female there are 120 training samples, 30 samples from each emotion and 40 testing samples, 10 samples from each emotion. Apart from this ,there is a general system also, taking 160 training and 80 testing samples. Equal number of samples are chosen for each emotion in this case too.

Among the three methods using both the kernels, binary tree using linear kernel has got the maximum recognition rate. The recognition rate in binary tree using both kernels is given in the confusion matrices below.

While using the Malayalam database ,it is divided into training set consisting of 96 samples and test set consisting of 24 samples. Result of the experiment is that the best recognition rate (95.83%) is achieved with this database tested on binary tree using linear kernel,considering the entire feature set.

Another database used is SAVEE. For classification purpose, 160 training samples ie, 40 samples from each emotion and 80 testing samples ie, 20 samples from

TABLE I: Confusion matrix of the linear SVM binary tree classifier Emo-DB(General)

Emotion	**Emotion Recognition(%)**			
	Anger	**Happy**	**Neutral**	**Sad**
Anger	75	25	0	0
Happy	25	75	0	0
Neutral	15	30	50	5
Sad	0	0	0	100
Total Accuracy				75

TABLE II: Confusion matrix of the linear SVM binary tree classifier Emo-DB(Female)

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	90	10	0	0
Happy	70	30	0	0
Neutral	0	10	60	30
Sad	0	0	0	100
Total Accuracy				70

TABLE III: Confusion matrix of the linear SVM binary tree classifier Emo-DB(Male)

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	70	30	0	0
Happy	20	80	0	0
Neutral	30	40	20	10
Sad	0	0	0	100
Total Accuracy				67.5

TABLE IV: Confusion matrix of the RBF SVM binary tree classifier Emo-DB(General)

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	40	60	0	0
Happy	15	65	20	0
Neutral	10	5	85	0
Sad	10	0	55	35
Total Accuracy				67.5

TABLE V: Confusion matrix of the RBF SVM binary tree classifier Emo-DB(Female)

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	80	20	0	0
Happy	40	60	0	0
Neutral	40	0	60	0
Sad	0	0	20	80
Total Accuracy				70

each emotion were selected. SVM is trained in the three classification strategies in linear as well as RBF kernel. Accuracy obtained for this database is poor and is noted in table XI. Emotion recognition performance can boost significantly if appropriate and reliable features are extracted. Computed the performance of various feature set in this database using binary tree and one against one classification strategy in linear kernel and conclusion drawn is that Pitch and

TABLE VI: Confusion matrix of the RBF SVM binary tree classifier Emo-DB(Male)

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	80	10	10	0
Happy	30	50	20	0
Neutral	0	10	90	0
Sad	0	0	100	0
Total Accuracy				55

TABLE VII: Confusion matrix of the Linear SVM binary tree classifier Malayalam Database

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	100	0	0	0
Happy	16.67	83.33	0	0
Neutral	0	0	100	0
Sad	0	0	0	100
Total Accuracy				95.83

TABLE VIII: Confusion matrix of the RBF SVM binary tree classifier Malayalam Database

Emotion	Emotion Recognition(%)			
	Anger	Happy	Neutral	Sad
Anger	66.66	33.34	0	0
Happy	0	100	0	0
Neutral	0	0	100	0
Sad	0	0	0	100
Total Accuracy				91.66

TABLE IX: Percentage accuracy of different feature set for Malayalam database in binary tree with linear kernel

Feature set	Accuracy(%)
mfcc	83.33
mfcc+Delta	83.83
mfcc+Delta+Delta-Delta	87.5
mfcc+Delta+Delta-Delta+Energy+Pitch	95.83

energy when considered alone didn't give good performance and the best performance is obtained when pitch, energy and MFCC are considered together.

Table XI shows the comparison between the classification strategies using linear and RBF kernel in various databases.

TABLE X: Percentage accuracy of different feature set for Malayalam database in one against one with linear kernel

Feature set	Accuracy(%)
Energy only	62.5
Pitch only	62.5
Energy+Pitch	75
mfcc+Delta+Delta-Delta+Energy	91.66

TABLE XI: Comparison between the classification strategies using both linear and RBF kernel in the different databases

Database	Kernel	Classification Strategy	Percentage Accuracy (%)
Our database	Linear	Binary tree	95.83
Emo-DB (General)	Linear	Binary tree	75
Emo-DB (Female)	Linear	Binary tree	70
Emo-DB (Male)	Linear	Binary tree	67.5
Savee	Linear	Binary tree	61.25
Our database	Linear	One against one	95.83
Emo-DB (General)	Linear	One against one	73.5
Emo-DB (Female)	Linear	One against one	70
Emo-DB (Male)	Linear	One against one	67.5
Savee	Linear	One against one	60
Our database	Linear	One versus the rest	87.5
Emo-DB (General)	Linear	One versus the rest	73.75
Emo-DB (Female)	Linear	One versus the rest	82.5
Emo-DB (Male)	Linear	One versus the rest	92.5
Savee	Linear	One versus the rest	57.5
Our database	RBF	Binary tree	91.66
Emo-DB (General)	RBF	Binary tree	56.25
Emo-DB (Female)	RBF	Binary tree	70
Emo-DB (Male)	RBF	Binary tree	55
Our database	RBF	One against one	91.66
Our database	RBF	One versus the rest	87.5

V. Conclusion

The paper describes that, distinct combinations of features can obtain different emotion recognition rate, and the sensitivity of different emotional features in different languages are also different. So the adjustment of features to different various corpuses is essential.

From the experiment, it can be concluded that the combination of various features provide different results. Emotion recognition performance can be boost up significantly if appropriate and reliable features are extracted. We computed the performance of various feature set and came into a conclusion that pitch and energy when considered alone didn't give good performance and the best performance is obtained when pitch, energy and MFCC are considered together.

While considering three classification strategies, maximum accuracy is obtained with the system implemented in binary tree. Out of the three databases, maximum accuracy of 95.83% was obtained for the database created by us tested in linear kernel.

ACKNOWLEDGEMENT

This work was carried out as part of "Main project" during eighth semester B Tech Electronics and Communication Engineering in partial fulfillment of the requirement for the award of Bachelor of Technology Degree in Electronics and Communication Engineering under the University of Calicut.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor," Emotion recognition in humancomputer interaction", IEEE Signal Process. Mag., 18 (2001), pp. 3280

[2] Ashish B. Ingale, D. S. Chaudhari,"Speech Emotion", International Journal of Soft Computing and Engineering (IJSCE)2012, Vol.2, Issue-1.

[3] Bhoomika Panda, Debananda Padhi, Kshamamayee Dash, Prof. Sanghamitra Mohanty,"Use of SVM Classifier & MFCC in Speech Emotion Recognition System", ijarcsse,Vol.2,Issue-3.

[4] B. Schuller, G. Rigoll, M. Lang, "Hidden Markov model-based speech emotion recognition", Proceedings of the IEEE ICASSP Conference on Acoustics,Speech and Signal Processing, vol.2, pp. 1-4, April 2003.

[5] Yixiong Pan, Peipei Shen and Liping Shen, 2012, "Speech Emotion Recognition Using Support Vector Machine", International Conference on Electronic and Mechanical Engineering and Information Technology,2011

[6] Iker Luengo, Eva Navas, Inmaculada Hernaez and Jon Sanchez, 2005,"Emotion Recognition using Prosodic Parameters", Interspeech, pp. 433 442.

[7] http://www.expressive-speech.net/. Berlin emotional speech database

[8] Ashish B. Ingale and Dr.D.S.Chaudhari, 2012, "Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine", International Journal of Advanced Engineering Research and Studies, Vol. 1,Issue 3.

[9] 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, June July 2011.

[10] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification", Technical Report,Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

[11] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Computer Applications, vol. 1, pp.6-9,February 2010

[12] Vaishali M. Chavan, V.V. Gohokar, 2012, "Speech Emotion Recognition by using SVM-Classifier",International Journal of Engineering and Advanced Technology, IJEAT, Vol. 1, Issue 5.

[13] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition 44,PP.572-587, 2011.

[14] Han Y, Wang G, Yang Y, 2008, "Speech Emotion Recognition Based on MFCC", Journal of Chong Qing University of Posts and Telecommunication, Natural Science Edition 20(5).

[15] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", (ISBN: 0-7803-8484-9), pp. I- 57780, IEEE International Conference on Acoustics,Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04), May 17-21 2004.

[16] "Pitch Extractionand Fundamental Frequency:History and Current Techniques", David Gerhard Technical Report ,November, 2003

[17] "Pitch Detection",Naotoshi Seo,Project

[18] Simina Emerich, Eugen Lupu, Anca Apatean,"Emotions recognition by speech and facial expressions analysis", 17th European Signal Processing Conference (EUSIPCO 2009).

[19] "Emotion Classification of Audio Signals Using Ensemble of Support Vector Machines" Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, June 16-18, 2008, Proceedings.