Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm

Aastha Joshi
Student, Department of Computer Science and Engineering
Sri Guru Granth Sahib World University, Punjab, India

Abstract: Speech is an interactive interface medium as it is possible to express emotions and attitude through speech. In this paper, a hybrid of Hidden Markov Models (HMMs) and Support Vector Machines (SVM) has been proposed to classify four emotions viz. happy, angry, sad and aggressive. Combining advantage on capability to dynamic time warping of HMM and pattern recognition of SVM. HMMs, which export likelihood probabilities and optimal state sequences, have been used to model speech feature sequences i.e. our proposed system is trained using HMM algorithm for emotions considered, while SVM has been employed to make a decision i.e. for classification. The recognition result of the hybrid classification has been compared with the isolated SVM and the maximum recognition rates have reached 98.1% and 94.2% respectively.

Keywords: SER System, feature extraction, HMM algorithm, SVM classifier, Performance Parameters.

I. INTRODUCTION

Communication is an important capability, not only based on the linguistic part but also based on the emotional part. In the field of HCI, emotion recognition from computer is still a challenging issue, especially when recognition is based solely on voice, which is the basic mean of human communication. It is an important preparation for automatic classification and recognition of emotions to select a proper feature set as a description to the emotional speech. The efficiency of Speech emotion recognition (SER) system is highly dependent upon naturalness of database used in the system. SER is not an easy task as it requires a set of successive operation such as voice activity detection, feature extraction, training & classification. In this scientific world, everything is going digital. Emotion detection in speech processing is one of the burning arenas in this filed. Many different researchers have tried their approach in this filed but accuracy is the major factor of the processing [1].

Our system has been fully implemented (in MATLAB R2012a) and tested for audio wave files. Result analysis is done using WEKA tool. The emotional speech input to the system is the collection of speech data. After collection of database which is considered as the training samples, necessary features were extracted from the speech signal to train the system using HMM algorithm. A feature set of 14 potentially features is extracted, analyzed and database is prepared in excel spreadsheet. Then the recorded test samples is presented to the SVM classifier which classifies the test sample into one of the emotion considered in our study and gives the recognized emotion as output.

Our basic problem is to detect the kind of emotion from a pitch file. To perform such operations we need to classify audio file on the basis of the following vector spaces.

i. Frequency map per of the audio file.
ii. Length of the audio.
iii. Type of the content of the audio file.

The processing steps will be creating a predefined cluster of audio files for the following criteria:

i. Happy Voice
ii. Angry Voice
iii. Sad Voice s
iv. Aggressive Voice

We will have to find the max and min frequency processed over at least ten files per section so that we can reach to a conclusion. Now our proceeding will then involve finding out the frequency range of the new uploaded file .This would be done with the help of the HMM algorithm which would identify the frequency parameters. Then after finding the exact length of the file, we will have to get into the predefined clusters. Mugging into the predefined clusters would be achieved by the SVM algorithm and each cluster will rollback to a result value. The exact cluster which would give us the maximum problitical analysis of the file would be our target cluster.

II. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some features which contain emotional information from the speaker's voice, and taking appropriate pattern recognition methods to identify emotional

states [2]. SER system has wide range of applications like in Human-Robotic Interfaces, in small call-centers, in intelligent spoken tutoring system etc.

The general architecture of our SER system has following steps:

i. Our speech processing system extracts some appropriate features from signal.
ii. Database is prepared for different emotions in excel spreadsheet.
iii. Using HMM algorithm our system is trained in a supervised manner with example data how to associate the features to the different emotions.
iv. SVM classifier is used to recognize different emotions by matching the features of uploaded audio file with the features of trained system.

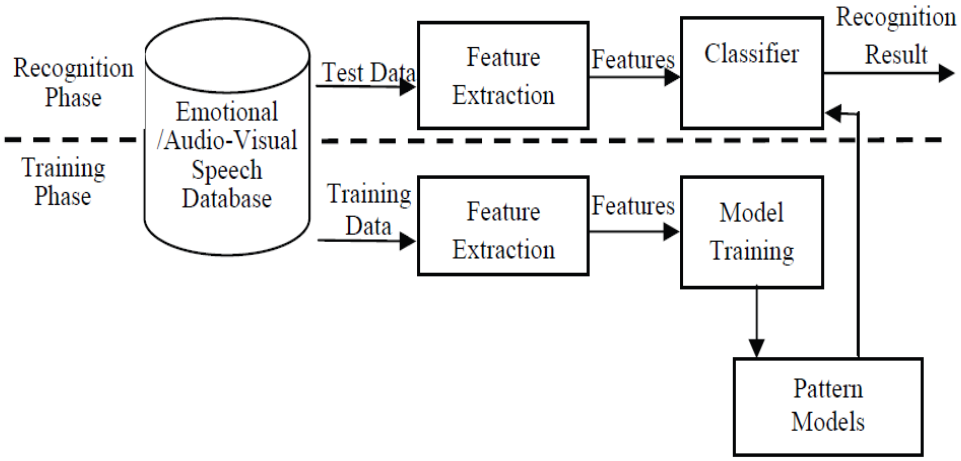

Fig 1: Speech Emotion Recognition System [3].

III. FEATURE EXTRACTION

Any emotion from the speaker's speech contains large number of parameters and the changes in these parameters will result in corresponding changes in emotions. In SER, feature extraction is one of the special forms of dimensionality reduction. Feature resources required to describe a large set of data accurately. Basically feature extraction is based on partitioning speech into small intervals known as frames [4].

A few desirable properties of features are:

i. High discrimination between sub word classes.
ii. Low Speaker variability.
iii. Invariance to degradation in the speech signal due to channel and noise.

The goal is to find the set of properties of an utterance that have an acoustic correlates in the speech signal, that is, the parameters that somehow are computed or estimated through processing of the signal waveform. Such parameters are termed as features [9].

Features can be derived either from frequency domain or time domain. In time domain, amplitude of signal is plotted with respect to time. Using the Fourier transform, a signal in the time domain can be transformed into the frequency domain, also known as the spectrum of signal [10]. Fourteen features have been evaluated for use in the system.

The features extracted to train our system are:

- *Pitch:* It is the main feature of an audio file. Sounds may be generally characterized by pitch, loudness, and quality. The perceived pitch of a sound is just the ear's response to frequency, i.e., for most practical purposes the pitch is just the frequency.
  Pitch = frequency of sound.
- *Standard Deviation:* standard deviation (represented by the symbol sigma) shows how much variation or dispersion exists from the average (mean), or expected value. A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.
- *Energy Intensity:* This feature represents loudness of an audio signal, which is correlated to amplitude of signal.
- *Energy Entropy:* It expresses abrupt changes in the energy level of an audio signal. In order to calculate this feature, frames are further divided into K-sub windows of fixed duration.
- *Shimmer:* A frequent back and forth changes in amplitude (from soft to louder) in the voice. Shimmer Percent provides an evaluation of the variability of the peak-to-peak amplitude within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability of the peak-to-peak amplitude.
- *Jitter:* It is defined as varying pitch in the voice, which causes a rough sound. Compare to shimmer, which describes varying loudness in the voice. Jitter is the undesired deviation from true periodicity of an assumed periodic signal. Jitter Percent provides an evaluation of the variability of the pitch period within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability.
- *Autocorrelation:* It is the cross-correlation of a signal with itself. Informally, it is the similarity between observations as a function of the time lag between them. It is a mathematical tool for finding repeating patterns,

such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.

- *Noise to Harmonic ratio:* Noise is an undesirable component that obscures a wanted signal. NHR is an average ratio of energy of the inharmonic components in the range 1500-4500 Hz to the harmonic components energy in the range 70-4500 Hz. It is a general evaluation of the noise presence in the analyzed signal (such as amplitude and frequency variations, turbulence noise, sub-harmonic components and/or voice breaks).
- *Harmonic to Noise ratio:* HNR represents the degree of acoustic periodicity, also called as Harmonicity object. Harmonicity is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is 10*log10(99/1) = 20 dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise.
- *Short Time Energy:* The amplitude of the speech signal varies appreciably with time. In particular, the amplitude of unvoiced segment is generally much lower than the amplitude of voiced segments. Short Time energy provides a convenient representation that reflects these amplitude variations. The major significance of this is that it provides a basis for distinguishing voiced speech from unvoiced speech.

- *Zero Crossing Rate:* It is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval, being a key feature to classify percussive sounds.
- *Spectral Centroid:* It is the weighted mean frequency. It indicates where the "center of mass" of the spectrum is. Because the spectral centroid is a good predictor of the "brightness" of a sound [5], it is widely used in digital audio and music processing as an automatic measure of music timbre.

$$C_{t=\frac{\sum_{n=1}^{N} M_t}{\sum_{n=0}^{N} M_t[n]}}[n] * n$$

where $M_t(n)$ is magnitude of Fourier transform at frame t and frequency bin n. The centroid is a measure of spectral shape and higher centroid values correspond to “brighter” textures with more high frequencies.

- *Spectral Rolloff :* Spectral Rolloff point is defined as the *Nth* percentile of the power spectral distribution, where *N* is usually 85% or 95% [7]. This measure is useful in distinguishing voiced speech from unvoiced: unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum, where most of the energy for voiced speech and music is contained in lower bands.

$$\sum_{n=1}^{\mathrm{Rt}} M_t(n) = 0.85 * \sum_{n=1}^{N} M_t(n)$$

Where R_t is the frequency below which 85% of the magnitude distribution is concentrated.

- *Spectral Flux:* It is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against power spectrum for the previous frame. More precisely, it is usually calculated as the Euclidean distance between the two normalized spectra.

These features have been extracted for every uploaded wave file and then database of these features is prepared for each emotion in excel spreadsheet. When the spreadsheet is uploaded in MATLAB to train the system it is stored as .mat file to form four different clusters of emotion.

IV. IMPLEMENTATION

System is trained using HMM and tested using SVM classifier i.e. output of HMM is taken as input of SVM for classification. TP Rate, FP Rate, Correctly Classified, Incorrectly Classified, Mean Absolute Error, Root Mean Squared Error, ROC Area are chosen as performance parameters for calculating accuracy and these performance parameters are analyzed with increasing percentage split, which are discussed as under.

A. Hidden Markov Model

Hidden Markov Model (HMM) is having the long history in the field of speech applications. The HMM consist of the first order markov chain whose states are hidden from the observer therefore the internal behavior of the model remains hidden. The hidden states of the model capture the temporal structure of the data [8]. Hidden Markov Models are statistical models that describe the sequences of events. HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix.

In this stage of our study, extracted audio features are used to train HMM for distinguishing between happy, sad, angry and aggressive emotions. The output of this model will be the input for SVM classifier to recognize the uploaded emotion for testing.

B. SVM Classifier

The support vector machine is a learning algorithm which addresses the general problem of learning to discriminate between positive & negative members of given n-dimensional vectors. The SVM is used for classification & regression purpose. The main idea of SVM classification is to a transform the original input set to a high dimensional feature space by using kernel function.

In Classification, training examples are used to learn a model that can classify the data samples into known classes. The Classification process involves following steps:

a. Create training data set.
b. Identify class attribute and classes.
c. Identify useful attributes for classification (Relevance analysis).
d. Learn a model using training examples in Training set.
e. Use the model to classify the unknown data samples.

SVM is a supervised learning process comprising of two steps:

i. Learning (Training): Learn a model using the training data.
ii. Testing: Test the model using unseen test data to assess the model accuracy.

The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of four categories, SVM testing algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM constructs a hyper plane or set of hyper planes in a high- or infinite- dimensional space, which can be used for classification. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class, since the larger the margin the lower the generalization error of the classifier.

C. Performance Parameters

TP Rate: The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall.

FP Rate: The False Positive (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x.

TP Rate and FP Rate are calculated in accordance with mean and variance. For calculating TP Rate & FP Rate, Gain is calculated by processing all the rows of uploaded file. Value of TP is equal to gain calculated and value of FP is equal to value of gain subtracted from the whole database [11].

Receiver Operating Characteristic (ROC): ROC is used as a performance parameter for comparing various classification algorithms. ROC Curve is also called threshold curve. ROC Area of different classification algorithms is compared on the basis of ROC Area values using 10-Fold cross validation.

V. RESULTS

Result Analysis for our study is done in Weka3.7.9. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset (using GUI) or called from your own Java code (using Weka Java library). A database file of 240 tuples and 14 attributes has been made in excel file, then conversion of this file to ARFF file is done. ARFF files are readable in Weka. The generated ARFF file is opened in Weka and then different pre-processing steps are applied for the classification of data. Classification is implemented using inbuilt 'SVM' and new hybrid of HMM & SVM classifier package built during our research and results are recorded and studied the TP Rate, FP Rate, Correctly Classified and Incorrectly Classified of these algorithms. A learning curve is drawn using training rate and performance parameters. To find out the learning rate of these two algorithms, the training is started from 10 % percentage split and keeps on increasing till 90% percentage split. Results of these two algorithms have been recorded and analyzed and interpretation has been done according to the analyses.

Table I

TP Rate and FP Rate of two classifiers using percentage split method

PERCENTAGE SPLIT (training set and rest testing set)	TP RATE		FP RATE	
	SVM CLASSIFIER	HYBRID CLASSIFIER	SVM CLASSIFIER	HYBRID CLASSIFIER
10 (training set)	0.752	0.479	0.082	0.168
20	0.832	0.678	0.056	0.102
30	0.868	0.780	0.043	0.064
40	0.859	0.859	0.047	0.042
50	0.877	0.915	0.041	0.024
60	0.933	0.962	0.022	0.010
70	0.936	0.949	0.024	0.010
80	0.942	0.981	0.022	0.005
90	0.923	0.962	0.036	0.007

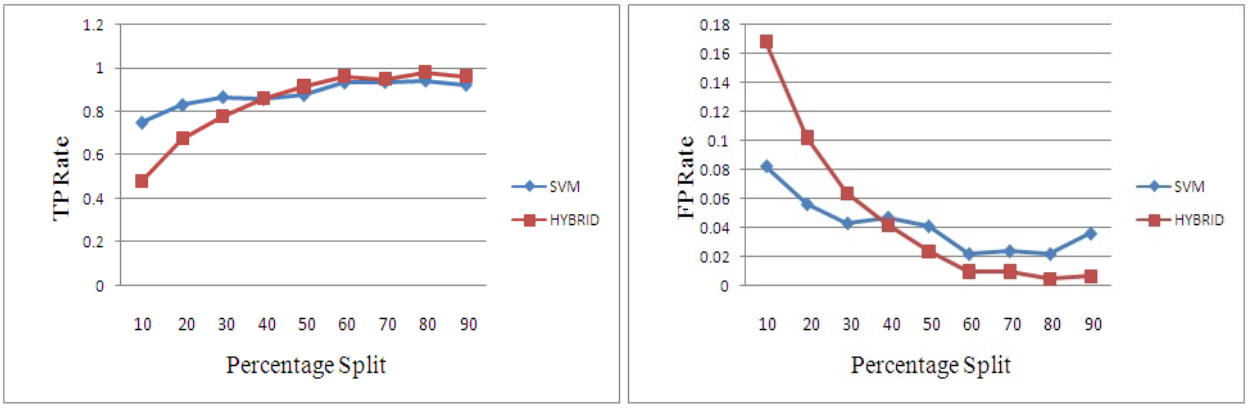

Fig. 1 TP Rate and FP Rate Vs Percentage Split

The graph in figure: 2 shows increasing TP Rate and decreasing FP Rate of both the algorithms with increasing training set. It is observed that our proposed hybrid algorithm shows better results compared to inbuilt SVM classification algorithm and at 80 percentage split TP Rate value comes out to be maximum and FP Rate value comes out to be minimum.

Table II

Correctly Classification and Incorrectly Classification of two classifiers using percentage split method

PERCENTAGE SPLIT (training set and rest testing set)	CORRECTLY CLASSIFIED		INCORRECTLY CLASSIFIED	
	SVM CLASSIFIER	HYBRID CLASSIFIER	SVM CLASSIFIER	HYBRID CLASSIFIER
10 (training set)	75.2137%	47.8632%	24.7863%	52.1368%
20	83.1731%	67.7885%	16.8269%	32.2115%
30	86.8132%	78.022%	13.1868%	21.978%
40	85.8974%	85.8974%	14.1026%	14.1026%
50	87.6923%	91.5385%	12.3077%	8.4615%
60	93.2692%	96.1538%	6.7308%	3.8462%
70	93.5897%	94.8718%	6.4103%	5.1282%
80	94.2308%	98.0769%	5.7692%	1.9231%
90	92.3077%	96.1538%	7.6923%	3.8462%

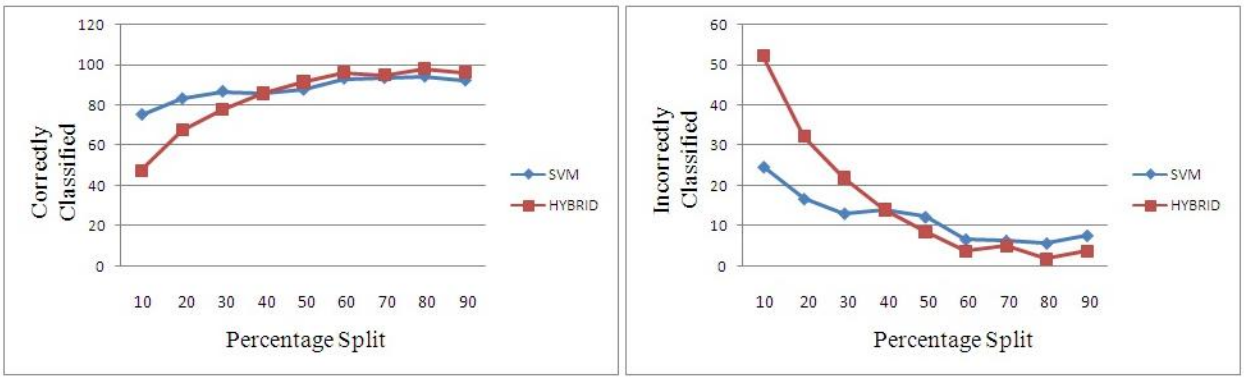

Fig. 3 Correctly Classified and Incorrectly Classified Vs Percentage Split

The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. The graph in figure: 3 shows that with increasing training set, data set is classified more correctly and vice versa. At 80 percentage split our proposed hybrid classifier correctly classified 98.0769% of instances whereas SVM correctly classifies 94.2308%. This shows our proposed hybrid classifier is better and gives more accuracy as compared to SVM alone.

Table II

Mean Absolute Error and Root Mean Squared Error of two classifiers using percentage split method

PERCENTAGE SPLIT (training set and rest testing set)	MEAN ABSOLUTE ERROR		ROOT MEAN SQUARED ERROR	
	SVM CLASSIFIER	HYBRID CLASSIFIER	SVM CLASSIFIER	HYBRID CLASSIFIER
10 (training set)	0.281	0.2607	0.3571	0.5106
20	0.2684	0.1611	0.335	0.4013
30	0.2656	0.1099	0.335	0.3315
40	0.2666	0.0705	0.3365	0.2655
50	0.2615	0.0423	0.328	0.2057
60	0.2572	0.0192	0.3232	0.1387
70	0.2572	0.0256	0.3236	0.1601
80	0.2564	0.0096	0.3219	0.0981
90	0.2596	0.0192	0.3269	0.1387

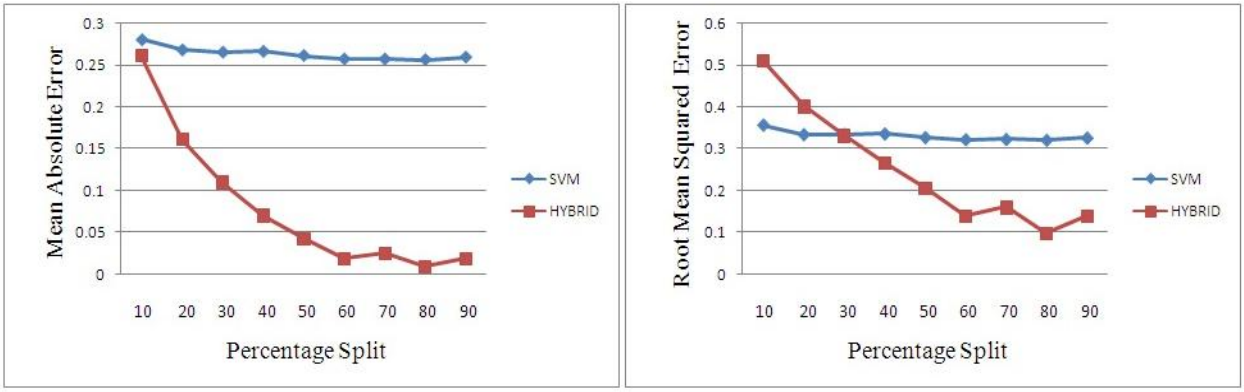

Fig.4 Mean Absolute Error and Root Mean Squared Error Vs Percentage Split

Mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. Whereas Root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. RMSE is a good measure of accuracy.

The graph in figure: 4 clearly shows that MAE decreases to 0.0096 at 80 percentage split for our proposed hybrid classifier whereas for SVM, MAE has minimum value of 0.2564 only which is much larger compared to proposed hybrid classifier. On the other hand RMSE decreases to 0.00981 at 80 percentage split for our proposed hybrid classifier whereas for SVM, RMSE has minimum value of 0.3219 only which is much larger compared to proposed hybrid classifier.

VI. CONCLUSION AND FUTURE SCOPE

We have studied on emotion speech recognition by means of HMMs and SVM, and we believe that HMM makes significant impact on speech emotion recognition as HMM proves to be a better training algorithm and SVM as better classification algorithm. Furthermore, a speech emotion recognizer that combines of HMM & SVM have been proposed. Performance of the hybrid classification and isolated SVMs were collected by experiments using emotional database. Undergoing the estimated procedure, we have expected our conclusion to be a better accurate system for the analysis of the audio files to detect the emotions in the field of clustering. We expect the accuracy to be increased in comparison with the hybrid of HMM & ANN. SVM combined with HMM is expected to work in better manner because the training set created with the help of SVM and HMM puts a strong emphasis in searching into the inner clusters of the files. Our result shows that our proposed hybrid classifier gives accuracy of 98.1% whereas isolated SVM classifier gives an accuracy of 94.2%. In future, work can be done to create more groups into the inner cluster of the files stored so that the searching becomes easy. To perform such task, one can use the CART algorithm which creates a regression tree which is a substitute of the binary decision trees. The future parameters can add the time slots of the frequencies at which the frequencies are consistent. Further our future work will explore the possibility to integrate other channels such as facial expression to increase the recognition rate.

References

[1] Ayadi M. E., Kamel M. S. and Karray F., "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", *Pattern Recognition*, 44(16), 572-587, 2011.

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

[3] [Online]. Available: http://crteknologies.fr/projects/emospeech/

[4] Shen P., Changjun Z. and Chen X., "Automatic Speech Emotion Recognition Using Support Vector Machine", *Proceedings of International Conference On Electronic And Mechanical Engineering And Information Technology*, 621-625, 2011.

[5] Grey, J. M., Gordon, J. W., 1978., "Perceptual effects of spectral modifications on musical timbres", *Journal of the Acoustical Society of America* 63 (5), 1493–1500, doi:10.1121/1.381843

[6] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals", *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 10, no. 5, JULY 2002.

[7] Dalibor,Matthias and Christian, "Features of Content Based Audio", *Advances in Computers*, vol. 78, pp. 71-150,2010.

[8] Ayadi M. E., Kamel M. S. and Karray F., "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", *Pattern Recognition*, 44(16), 572-587, 2011.

[9] 1Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur, "The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 2, Issue 2, February 2012.

[10] Darin Brezeale and Diane J. Cook, Senior Member, IEEE, "Automatic Video Classification: A Survey of the Literature", *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, 2007.

[11] Shruti Aggarwal, Naveen Aggarwal, "Classification of Audio Data using Support Vector Machine",*IJCST*, vol. 2, Issue 3, September 2011.

