SVM Scheme for Speech Emotion Recognition using MFCC Feature

A. Milton
Asst. Professor, Dept. of ECE
SXCCE, Chunkankadai
Kanyakumari District
Tamil Nadu, India

S. Sharmy Roy
Department of ECE
SXCCE, Chunkankadai
Kanyakumari District
Tamil Nadu, India

S. Tamil Selvi, PhD.
Professor, Department of ECE
National Engineering College,
Kovilpatti
Tamil Nadu, India

ABSTRACT

Emotion recognition from speech has developed as a recent research area in Human–Computer Interaction. The objective of this paper is to use a 3-stage Support Vector Machine classifier to classify seven different emotions present in the Berlin Emotional Database. For the purpose of classification, MFCC features from all the 535 files present in the database are extracted. Nine statistical measurements are performed over these features from each frame of a sentence. The linear and RBF kernels are employed in hierarchical SVM with RBF sigma value equal to one. For training and testing of data, 10-fold cross-validation is used. Performance analysis is done by using the confusion matrix and the accuracy obtained is 68%.

General Terms

Speech Processing, Emotion Recognition System

Keywords

Speech Emotion Recognition, MFCC, SVM, RBF, Linear Kernel.

1. INTRODUCTION

Identification of emotions from human interactions serves to be a tough criterion. The aim of emotion recognition system [27] is to enable Human–Computer Interaction (HCI). In the recent past, a great number of researches have been carried out in this area [10]. The MFCC features are extracted from the speech signal for further classification. It is necessary to select the best feature for effective emotion recognition of any system and so MFCC, which is one of the spectral features, is used [31]. MFCC's are obtained as a result of linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Then, the SVM is used for classification. SVM is an effective binary classifier [14] used for both classification and regression purposes. To solve multi-class problems numerous single stage SVM's are used. Added to that, one–against–all approach is adopted in this paper.

2. LITERATURE REVIEW

The speech emotion recognition system employs various processes. Some of the important steps are feature extraction, subset selection and classification. The different features and classifiers used in various papers have been analyzed here. A summary of the different types of available database corpuses is also presented below.

2.1 Databases – Review

For characterization of emotions, either for synthesis or for recognition, suitable emotional speech database is the primary requirement. There are three types [12] of speech corpuses. A brief literature about these corpuses is discussed below.

2.1.1 Actor Based Emotional Speech Database

Simulation or actor based emotional speech corpora are collected from experienced, professional and trained radio artists. These emotions are also known as full blown emotions. Some of the common databases used are, a speaker independent Tamil language database [10], Burmese and Mandarin database [17] and Danish Emotional Speech Database [18], which was recorded using 4 radio artists, 2 male and 2 female. Some others include Emotional Speech Database for Basque [9] which consists of six basic emotions recorded from professional dubbing artists and IITKGP-SESC [13], a corpus which was recorded by professional artists from All India Radio, Vijayawada. The speech samples were recorded using SHURE dynamic cardioid microphone and were sampled at 16 kHz. There are totally 12000 utterances. Hence there are about 1500 utterances for each of the 8 emotions such as anger, disgust, fear, happiness, neutral, surprise, sarcastic and compassion. In [1] a database of emotional speech was collected in an anechoic chamber with the help of condenser cardioid microphone RODE-NT2. Ten speakers were fitted with neck impedance so that the glottal signal could be recorded simultaneously using a larynograph.

2.1.2 Elicited Emotional Speech Database

Elicited [25] otherwise known as induced speech corpora are collected by simulating an artificial emotional situation, without the knowledge of the speaker. Designation of such databases is tedious. Hence very few elicited speech corpuses are available until date.

2.1.3 Natural Emotional Speech Database

Naturally available data may be recorded from call center conversations, cockpit recordings, emotional conversations in public places etc. They are hence real emotions. In [15] a database created from a German TV talk show "Vera am Mittag" was used. Recordings were manually separated at the utterance level. The emotions are described in an emotion space using three emotion primitives. They are valence, activation and dominance.

2.2 Features – Review

Choosing suitable features for developing any of the speech systems is a crucial decision. The following subsections present the literature on three important speech features.

2.2.1 Excitation Source Features

These features are extracted from source signals such as vocal tract characteristics, linear prediction residual, glottal volume velocity etc. [8] at the epoch level. These features contain all information of the speech signal and hence can be used to classify emotions effectively [12]. Literature shows that very less work has been carried out using these features.

2.2.2 Spectral Features

They are obtained from the cepstral domain at the frame level. The cepstrum is the Fourier transform of the log magnitude spectrum and hence gives short time information. Some of the spectral features are MFCC [31], LPCC [21], LFPC [5], [24] etc. Combination of two or more features presents more accuracy [13]. These features are successfully used in speech and speaker recognition systems.

2.2.3 Prosodic Features

These features are extracted at the utterance level [12]. They include energy, pitch, duration and their derivatives. Few attempts have been carried on so far to explore the dynamic nature of these features. Researchers mainly concentrate on their static nature alone.

2.3 Classifiers – Review

In general, the classifiers used for speech emotion recognition can be classified into two broad categories [12] namely,

- Linear classifiers
- Non-linear classifiers.

Linear classifier classifies based on the value of linear combination of the object characteristics whereas non-linear classifiers are based on weighted combination values of the object. A short summary about various classifiers used in different papers are presented below.

2.3.1 Hidden Markov Model

Hidden Markov Model [17], [30] is popular for speech recognition and hence they are adopted for emotion recognition. HMM has long history in the field of speech applications. It consist of first order Markov chain whose states are hidden from the observer and therefore the internal structure of the model remains hidden. The hidden states of the model capture the dynamic nature of the data [2]. The structure of HMM generally adopted is left-to-right structure.

2.3.2 Gaussian Mixture Models

GMM is one of the most prominent statistical methods used for clustering. Expectation Maximization [13] is an iterative method used in GMM. It captures distribution pattern of data points. It is a single-state HMM and is known to have least performance [1].

2.3.3 K-Nearest Neighbor Classifier

The K-Nearest Neighbor classifier is one of the simplest machine learning algorithm that identifies the object by the majority vote of its neighbors based on Euclidean distance. If the value of k is large, big classes will overwhelm the small ones. On the other hand, if value of k is too small, the advantage of KNN algorithm will not be established. When $K \rightarrow \infty$, a less biased classifier is obtained [22].

2.3.4 Optimum-Path Forest Classifier

The Optimum-Path Forest (OPF) classifier [1] was recently proposed as an alternative approach to overcome the problems faced by the previous classifiers. It is in fact a simple, multi-class and parameter independent supervised pattern recognition technique that does not make any assumption about shape and can handle some degree of separability between the data. The overall performance of OPF is faster than SVM.

2.3.5 Auto Associative Neural Network

AANN models [13] are basically Feed-Forward Neural Network (FFNN) models, which maps an input vector onto itself and hence the name auto-association or identity mapping. It consists of one input layer, one output layer and numerous hidden layers. The number of units in the input and output layers is equal to the dimension of the input feature vectors. The number of nodes in one of the hidden layers is less than the number of units in either the input or output layer. This hidden layer is known as dimension compression layer.

2.4 Performance Analysis – Review

The combination of energy, pitch and duration features provided an accuracy of 65% in [13]. Moreover, the combination of global and local prosodic features [13] produced an accuracy of about 65.63%. In [18] the recognition rates obtained by using SVM classifier for linear, polynomial, RBF and sigmoid kernel function are 68%, 60%, 55.4% and 60% respectively. The classification accuracy using LPCC, MFCC and LFPC were 56.1%, 59% and 77.1% respectively in [17]. The accuracy obtained by SVM are 62% and 71.66% for Berlin Database and Hindi database respectively in [16]. The standard HMM model produced an accuracy of about 68.57% and the hierarchical 3 model HMM achieved 71.75% accuracy in [5].

3. PROPOSED SYSTEM

The 3-stage hierarchical SVM is proposed in this paper. In the first stage emotions anger, disgust, fear and happiness are separated from emotions boredom, neutral and sadness. In the second stage 2 SVM’s are used. The first one classifies the emotions anger, happiness and disgust, fear whereas the second one classifies the emotions boredom, neutral and sadness. In the third stage 3 SVM’s are employed. The first one isolates anger and happiness whereas the second one classifies disgust and fear and the last one separates boredom and neutral. Thus 6 binary SVM’s are needed to construct a 3-stage SVM to separate 7 emotions individually. The hierarchical structure of a 3-stage SVM is shown in Figure 2.

3.1 Berlin Emotional Database

The Berlin Emotional Database [7] was recorded at the Technical University of Berlin with 10 professional artists, 5 male and 5 female. It is an internationally known speech corpus. It consists of 535 acted emotions in German language with 7 different emotions and is a multi-speaker database making it easy to perform speaker-independent tests. It also consists five short and five long sentences typically between 1.5s and 4s for database construction. The raw database basically consists of about 800 utterances which were screened by 20 listeners. The database [16] was recorded using the Sennheiser MKH 40 P48 microphone, with sampling frequency of 16 kHz. Samples are stored as 16 bit numbers. This database is normally preferred to other manually created databases because of its availability and easy in testing, using leave-one-out cross-validation [15]. The various emotions and the number of speech files present in the database are listed below in Table 1.

Table 1. Number of emotional speech files in Berlin Database

Emotion	No. of Speech Files
Anger (A)	127
Boredom (B)	81
Happiness (H)	71
Fear (F)	69
Sadness (S)	62
Disgust (D)	46
Neutral (N)	79
Total number of files	535

3.2 MFCC Feature Extraction

MFCC is one of the most widely used spectral features available [20], [23]. It has numerous advantages like simple calculation, better ability of distinction and high robustness to noise. Here MFCC features are extracted from the Praat software [6] with window length 20ms and time step 10ms. Generally, the hamming window is preferred because of its high frequency resolution and good side lobe suppression properties. First, the silence regions present in the database was removed based on the zero–crossing rate and also by thresholding the energy. The silence region does not contain any useful information and is hence removed. Human perception of hearing does not follow a linear scale and hence MFCC follows the Mel scale [14] which is a frequency scaling having linear spacing below 1000Hz and logarithmic spacing above 1000Hz. The formula to compute the Mel frequency for any given frequency f in Hz is given below [19],

$$\text{Mel}(f) = 2595 \times \log\left(1 + \frac{f}{700}\right) \quad (1)$$

The Mel scale filter bank has a triangular series of uniform overlapping filters with constant bandwidth equal to 100 and their centre frequencies at 50. This is what is believed to occur in the human auditory system [28]. This corresponds to the spacing on the Mel frequency scale.

3.3 SVM Training and Classification

The main idea of SVM [26], [29] is to transform the original input set to a high dimensional feature space by using a kernel function, in which input space consisting of input samples is converted into high dimensional feature space and therefore the input samples become linearly separable [14]. It is clearly explained by using an optimal separation hyperplane in Figure 1. The main advantage of SVM is that it has limited training data and hence has very good classification performance. For linearly separable data points, classification is done by using the following formula [3],

$$\langle w \cdot x \rangle + b_0 \geq 1, \forall y = 1 \quad (2)$$

$$\langle w \cdot x \rangle + b_0 \leq -1, \forall y = -1 \quad (3)$$

where, (x, y) is the pair of training set. Here, $x \in R^n$ and $y \in \{+1, -1\}$.

$\langle w \cdot x \rangle$ represents the inner product of w and x whereas b_0 refers to the bias condition.

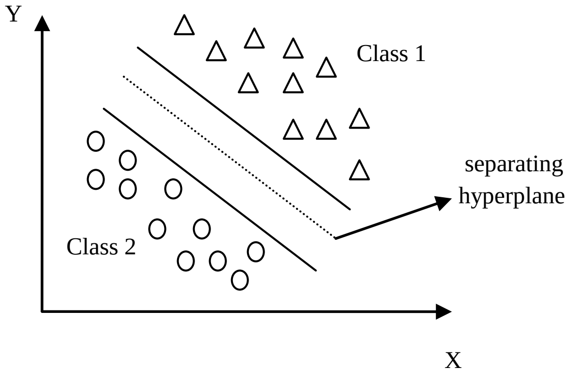

Fig 1: SVM structure

SVM that employs both the linear kernel function and the Radial Basis Kernel (RBF) function [18] is used here. The linear kernel function is given by the formula below,

$$\text{Kernel } (x, y) = (x \cdot y) \quad (4)$$

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

$$\text{Kernel } (x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}} \quad (5)$$

4. EXPERIMENTAL EVALUATION

Initially, 24 MFCC features are extracted from each utterance of the database using the Praat software. About nine statistical measurements like mean, median, maximum value, minimum value, range, inter–quartile range, standard deviation, kurtosis and skewness are performed over these features. It is found that mean is the most prominent one among all other statistics. A 3-stage SVM hierarchy which is shown in Figure 2 has been developed for experimentation.

In the first stage, the 10th MFCC feature is trained using an SVM with RBF kernel. This is chosen because of its high accuracy of 97.76% when compared with the other features. 10-fold cross-validation is used for testing and training the observation. The confusion matrix is as shown in Table 2. Hence, the emotions are classified into two groups in this stage. The first group consists of the emotions, anger, disgust, fear and happiness whereas the second group consists of boredom, sadness and neutral. It is to be noted that all the training in SVM with RBF kernel is done by using the sigma value of one.

Table 2. Confusion matrix of 1st stage RBF SVM classifier

Emotion	Emotion Recognition (%)	
	AHDF	BNS
AHDF	99.04	0.96
BNS	4.05	95.95

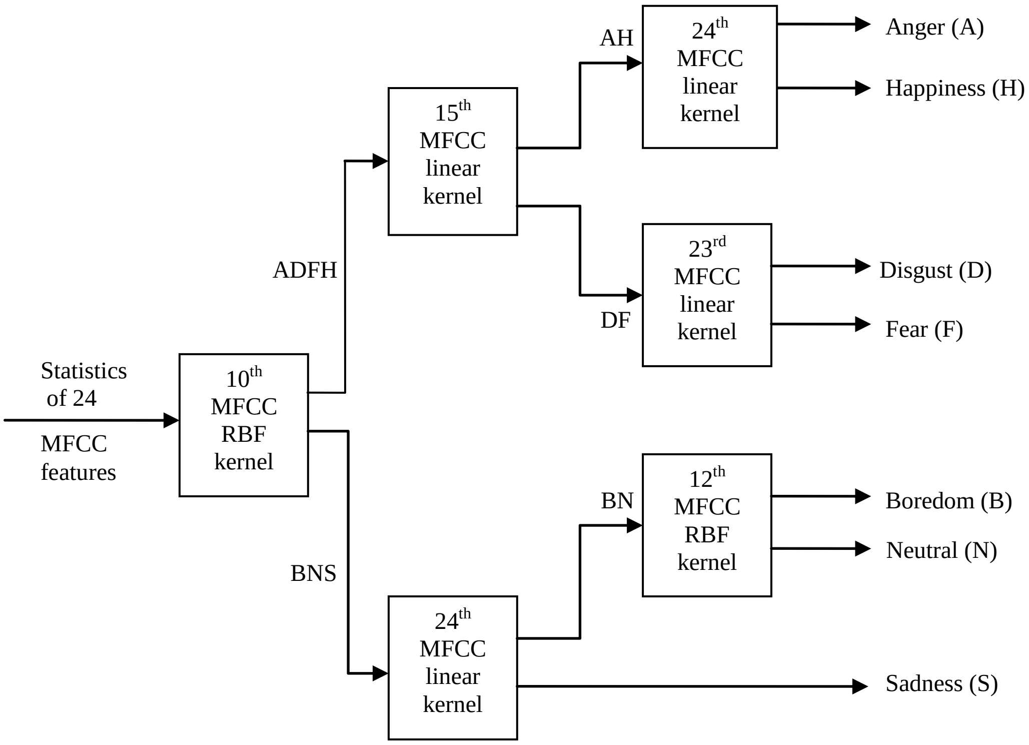

Fig 2: Structure of 3-stage hierarchical SVM

In the second stage, the 15th MFCC feature is trained using an SVM with linear kernel function to produce an accuracy of 91.96%. This is shown in Table 3. The emotions anger and happiness are generally confused the most and hence they are grouped into one and all the remaining emotions are grouped into class 2.

Table 3. Confusion matrix of 2nd stage linear kernel SVM classifier

Emotion	Emotion Recognition (%)	
	AH	DFBNS
AH	89.4	10.6
DFBNS	6.53	93.47

Next, the 24th MFCC feature is trained using an SVM with linear kernel function to obtain an accuracy of 95.89% and is shown in Table 4. Here, one-against-all approach is used. Hence the emotion sadness is identified and all other emotions are grouped separately. Sadness gets confused mostly with boredom and neutral.

Table 4. Confusion matrix of 2nd stage linear kernel SVM classifier

Emotion	Emotion Recognition (%)	
	S	ADFHBN
S	80.65	19.35
ADFHBN	2.16	97.89

In the third stage, the 24th MFCC feature outperforms all other features and is hence used to be trained with an SVM using linear kernel function. It gives an accuracy of 93.83% and is shown in Table 5. Here anger is grouped separately whereas all others in class 2.

Table 5. Confusion matrix of 3rd stage linear kernel SVM classifier

Emotion	Emotion Recognition (%)	
	A	HDFBNS
A	87.4	12.59
HDFBNS	4.17	95.83

Here the emotions anger and happiness are isolated and hence identified. Next, the 23rd MFCC with SVM using linear kernel function is trained. An accuracy of 95.33% has been obtained. It is shown in Table 6.

Table 6. Confusion matrix of 3rd stage linear kernel SVM classifier

Emotion	Emotion Recognition (%)	
	D	AHFBNS
D	60.87	39.13
AHFBNS	1.43	98.57

Emotions disgust and fear are identified now. Normally, the identification of the emotion disgust is very tedious. It is confused with the emotion fear the most. Then 12th MFCC using RBF kernel in SVM is employed. It produces an accuracy of 90.09% and is shown in Table 7. Here class 2 is more biased than class 1.

Table 7. Confusion matrix of 3rd stage RBF SVM classifier

Emotion	Emotion Recognition (%)	
	N	ADFHBS
N	39.2	60.76
ADFHBS	1.1	98.9

After this step the emotions boredom and neutral are classified separately. At the end of the third stage all emotions are identified individually and the overall confusion matrix is obtained as shown in Table 8.

Table 8. Confusion matrix of overall RBF SVM classifier

Emotion	Emotion Recognition (%)						
	A	D	F	H	B	N	S
A	83.5	0	0	3.15	13.4	0	0
D	0	82.7	0	6.17	0	6.17	4.93
F	2.2	2.17	47.8	30.43	17.4	0	0
H	8.7	1.45	2.9	73.9	10.1	2.9	0
B	16.9	1.4	1.4	22.5	57.7	0	0
N	0	22.6	0	3.22	0	74.2	0
S	0	56.9	0	2.53	0	1.26	39.2

Finally, it yields an overall accuracy of 68%. More concentration has to be made in order to use the derivatives of MFCC features delta and delta-delta for a higher accuracy. The proposed method is compared with many other existing methods to prove its efficiency. The performance comparison is tabulated and shown in Table 9. The combination of various features and classifiers used are discussed. When prosodic features and SVM is used the accuracy is just 48%. It is hence well clear that MFCC features perform better in all cases. Moreover SVM that uses linear kernel also produces a higher accuracy of 65% when compared to polynomial and radial basis function kernel that produce only 60% and 55.4% respectively.

5. CONCLUSION AND FUTURE WORK

Due to very less knowledge about this field there are very few researches going on in the area of speech processing. But a large amount of work can be done by processing the spectral features effectively to recognize the emotions. Here, using 24 MFCC features an accuracy of 68% has been achieved in the Berlin emotional database. Moreover, higher accuracy can be obtained using the combination of more features. To sum up, future work is to extract the delta features from each utterance and then use the SVM hierarchical structure for classification. Also while increasing the sigma value from the default value one, significant results may be obtained.

Table 9. Tabulation for performance comparison

Features & Classifiers Used	Accuracy (%)
MFCC & HMM [17]	59
MFCC & K-NN [31]	67
MFCC & FFNN [25]	55
Energy + pitch + duration & SVM [13]	48
MFCC & SVM using linear kernel [5]	65
MFCC & SVM using polynomial kernel [18]	60
MFCC & SVM using radial basis function kernel [2]	55.4
MFCC & 3 – Stage SVM [proposed method]	68

6. REFERENCES

[1] Alexander I. Iliev, Michael S. Scordilis, Joao P. Papa and Alexandre X. Falcao, 2010,"Spoken emotion recognition through optimum-path forest classification using glottal features", Computer Speech and Language 24, pp. 445 - 460.

[2] 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.

[3] Bhoomika Panda, Debananda Padhi, Kshamamayee Dash and Prof. Sanghamitra Mohanty, 2012, "Use of SVM Classifier & MFCC in Speech Emotion Recognition System", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 3.

[4] Bottou L. and Chih-Jen Lin, 2007, "Support Vector Machine Solvers", Dorling Kindersley Publication.

[5] Enrique M. Albornoz, Diego H. Milone and Hugo L. Rufiner, 2011,"Spoken emotion recognition using hierarchical classifiers", Computer Speech and Language 25, pp. 556 –570.

[6] http://www.fon.hum.uva.nl/praat/, Last accessed on 12.11.2012.

[7] http://www.expressive-speech.net/, Berlin emotional speech database, Last accessed on 25.10.2012.

[8] Iker Luengo, Eva Navas and Inmaculada Hernaez, 2010, "Feature Analysis an Evaluation for Automatic Emotion Identification in Speech", IEEE Transactions on Multimedia, Vol. 12, No. 6, pp. 490 - 501..

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

[10] Jeong-Sik Park, Ji-Hwan Kim and Yung-Hwan Oh, 2009, "Feature Vector Classification based Speech Emotion Recognition for Service Robots", IEEE Transactions on Consumer Electronics, Vol. 55, No. 3, pp. 1590 – 1596.

[11] Lawrence R. Rabiner and Ronald W. Schafer, 1978, "Digital Processing of Speech Signals", Prentice Hall.

[12] Shashidhar G. Koolagudi and K. Sreenivasa Rao, 2011, "Emotion recognition from speech: a review", Int J Speech Tech, pp. 119 – 128.

[13] Shashidhar G. Koolagudi and K. Sreenivasa Rao, 2012, “Emotion recognition from speech using source, system and prosodic features”, Int J Speech Tech, pp. 265 – 289.

[14] Simon Haykin, 1999, “Neural networks: A Comprehensive Foundation”, Pearson Education.

[15] Siqing Wu, Tiago H. Falk, Wai-Yip Chan, 2011,“Automatic Speech Emotion Recognition Using Modulation Spectral Features”, Speech Communication 53, pp. 768 - 785.

[16] Sujata B.Wankhade, Pritish Tijare and Yashpalsing Chavhan, 2011, “Speech Emotion Recognition System Using SVM AND LIBSVM”, International Journal Of Computer Science And Applications, Vol.4, No. 2.

[17] Tin Lay Nwe, Say Wei Foo, Liyanage C. De Silva, 2003, “Speech Emotion Recognition Using Hidden Markov Models”, Speech Communication 41,pp. 603 - 623.

[18] 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.

[19] Vibha Tiwari, 2010, “MFCC and its applications in speaker recognition”, International Journal on Emerging Technologies, ISSN : 0975-8364.

[20] Vimala.C, Dr.V.Radha, 2011, “Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM”, International Conference on Communication Technology and System Design, Speech Communication 46.

[21] Yixiong Pan, Peipei Shen and Liping Shen, 2012, “Speech Emotion Recognition Using Support Vector Machine”, International Journal of Smart Home, Vol. 6, No. 2.

[22] Yongjin Wang and Ling Guan, 2008, “Recognizing Human Emotional State from Audiovisual Signals”, IEEE Transactions on Multimedia, Vol. 10, No. 5, pp. 936 – 946.

[23] Emily Mower, Maja J Mataric, and Shrikanth Narayanan, 2011, “A Framework for Automatic Human Emotion Classification Using Emotion Profiles”, IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, No. 5, pp. 1057 - 1070.

[24] Chung-Hsien Wu and Wei-Bin Liang, 2011, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels”, IEEE Transactions on Affective Computing, Vol. 2, No. 1, pp. 10 - 21.

[25] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, 2011,“Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases”, Pattern Recognition 44, pp. 572 – 587.

[26] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, 2004,“Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture”, IEEE, ICASSP, pp. I – 577 - I – 580.

[27] Cowie.R, 2011,“Emotion Recognition in Human-Computer Interaction”, IEEE Signal Processing Magazine, Vol. 18, No.1, pp.22-80.

[28] Rabiner, L.R., & Juang, B.H., 1993,“Fundamentals of Speech Recognition”, Englewood Cliffs, Prentice-Hall.

[29] Hsu C.W, Chang.C, Lin C.J,“A Practical Guide to Support Vector Classification”, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

[30] Lin.Y, Wei.G, 2005, “Speech Emotion Recognition Based on HMM and SVM”, International Conference on Machine Learning and Cybernetics, Vol.8, pp. 4898-4901.

[31] 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).