Speech Emotion Recognition by Using Combinations of Support Vector Machine (SVM), and C5.0

Mohammad Masoud Javidi[1] and Ebrahim Fazlizadeh Roshan[2]

[1,2]Department of Computer Science, Shahid Bahonar University of Kerman, Kerman, Iran.

ABSTRACT:

Speech emotion recognition enables a computer system to records sounds and realizes the emotion of the speaker. we are still far from having a natural interaction between the human and machine because machines cannot distinguishes the emotion of the speaker. For this reason it has been established a new investigation field, namely "the speech emotion recognition systems". The accuracy of these systems depend on the various factors such as the type and the number of the emotion states and also the classifier type. In this paper, the classification methods of C5.0, Support Vector Machine (SVM), and the combination of C5.0 and SVM (SVM-C5.0) are verified, and their efficiencies in speech emotion recognition are compared. The utilized features in this research include energy, Zero Crossing Rate (ZCR), pitch, and Mel-scale Frequency Cepstral Coefficients (MFCC). The results of paper demonstrate that the effectiveness proposed SVM-C5.0 classification method is more efficient in recognizing the emotion of the between -5.5 % and 8.9 % depending on the number of emotion states than SVM, C5.0.

KEY WORDS:

Emotion recognition, Feature extraction, Mel-scale Frequency Cepstral Coefficients, C5.0, Support Vector Machines

1. INTRODUCTION

Speech emotion recognition aims are to design of the operator systems which receive the speech signals and extract the emotional states from them. This technology enables a sound recording computer (e.g. a computer that has a microphone) to realize the emotion of the speaker. However, there are a lot of challenges before any speech emotion recognition system. The most important challenges are emotional databases, feature extraction and classification models. A machine learning framework for emotion recognition from speech are displayed in figure 1.

In the last three decades, several attempts have been done in order to recognize the speech emotion which the most important ones include [1-6]. The voice and the prosodic features, the speaking style, the speaker's characteristics, and the linguistic features can affect the emotion [5]. Several models such as hidden Markov models (HMM) in [9], Gaussian mixture model (GMM) in [5, 14, 15], NN in [13, 8], and SVM in [5, 16] have been utilized in order to recognize the speech emotion. It is well investigated in [7] that the SVM and the HMM lead to the most and the least recognitions, respectively. Another model was suggested in 2013 in which the neural network, SVM, C5.0 and their combination were analyzed [27].

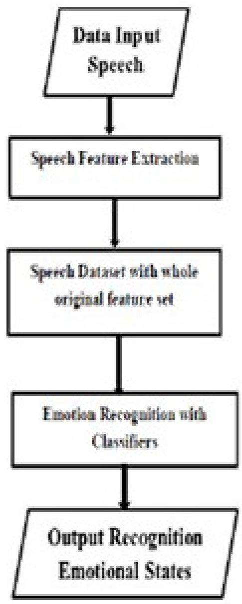

Figure 1. A machine learning framework for emotion recognition from speech.

We propose in this paper the SVM-C5.0 method in order to model the speech emotionrecognition systems more precisely. Regarding the existing features in the speech processing such as energy, ZCR, pitch, MFCC, and etc.

Emotion states of anger, happiness, fear, sadness, disgust, boredom, and neutral have been considered. We will apply these recognizing emotion states to different classification models of SVM, C5.0 and SVM-C5.0 by using the SPSS IBM MODELER software and verify the results. The results of paper demonstrate that the effectiveness proposed SVM-C5.0 classification method is more efficient in recognizing the emotion of the between -5.5 % and 8.9 % depending on the number of emotion states than SVM, C5.0.

The rest of this paper is organized as follows: in section 2, we review the most recently proposed speech emotion recognition systems, briefly. In section 3, we verify the database of Berlin. Section 4 and 5 explain the extracted features and utilized models, respectively. In section 6, we present the experimental results and also compare different methods. In the last section, the conclusion of this paper is explained.

2. RELATED WORK

In 1990's, the most recognizing models were proposed based on Linear Recognize Classification (LDC)

[17] and Maximum Likelihood Bays (MLB) [10, 26]. In the recent years, on the other hand, GMM [5,12, 14], NN [13, 8], Multi-Layer Perceptron (MLP) [5], K-Nearest Neighbor (KNN) [11, 14], HMM [9], and SVM [5, 16] used to recognize speech emotion.

Table 1 shows the utilized models in the recent years along with the number of emotion stats and their recognition rates. In [26], Haq et al. used the MLB classification model and the following seven emotion states: the anger, disgust, fear, happiness, neutral, sadness, and surprise. They could attain the recognition rate of 53%. In [10], Ververidis and Kotropoulos used the MLB classification model and the anger, happiness, neutral, and sadness emotion states and could achieve the recognition rate of 53.7%. In [8], Yu et al. used the SVM and ANN classification models and the following four emotions: anger, happiness, neutral, and sadness, and could reach the recognition rates of 71% and 42% respectively for the SVM and ANN models, which demonstrate that the SVM is more high performance than the ANN.

Table1. The most recent utilized models and their associated results

Paper	Emotion number	Classifier	Recognition rate
Petrushin VA (2000) [11]	5	KNN	70%
Yu F, Chang e, Xu y, Shum h (2001) [8]	4	SVM , ANN	71%,42%
Ververidis D, Kotropoulos c (2006) [10]	5	MLB	53.7%
Ayadi M, Kamel S, Karray F (2007) [9]	6	ANN, HMM	55%,71%
Haq S, Jackson PJB, Edge J (2008) [26]	7	MLB	53%
Gharavian D, Sheikhan M, Pezhmanpour M(2011)[12]	4	GMM	65.1%

He L, Lech M, Maddage NC, Allen NB (2011)[14]	5	GMM, KNN	77%,77%
Sheikhan M, Bejamin M, Gharavian D (2012)[5]	3	GMM, C5.0, MLP, MODULAR NEURAL-SVM	65.9%,56.3%,68.3%,76%
Hamidi M, Mansorizadeh M (2012)[13]	5	NN	78%
Fersini E, Messina E, Archetti F (2012)[16]	6	SVM	70.83%
Javidi MM, Roshan EF (2013)[27]	2	NN-C5.0	91.625
Javidi MM, Roshan EF (2013)[27]	3	NN-C5.0	82.122
Javidi MM, Roshan EF (2013)[27]	4	NN-C5.0	82.214
Javidi MM, Roshan EF (2013)[27]	5	NN-C5.0	75.807
Javidi MM, Roshan EF (2013)[27]	6	NN-C5.0	74.946
Javidi MM, Roshan EF (2013)[27]	7	NN-C5.0	72.621

In [9], Ayadi et al. could gain the recognition rates of 71% and 55% for the HMM and ANN models, respectively, by using the their classification models and seven emotion states, which reveals that the HMM model performs more better than the one of ANN. In [11], Petrushin used the KNN model and the anger, happiness, sadness, fear, and neutral emotion states and reached the recognition rate of 70%.

In [12], Gharavian et al. used the GMM model for four emotion states and could reach the recognition rate of 65.1%. In one of the most recent researches in [5], Sheikhan et al. used three emotion states of happiness, anger, and neutral and two classification models, namely modular neural-SVM and C5.0. They demonstrated the recognition rates of 76.3% and 56.3% for the former and later models, respectively. And finally, a model

which is a combination of the NN and C5.0 (NN-C5.0) classification methods was proposed in 2013 [27]. This model was applied to various emotion states and its results for 2, 3, 4, 5, 6, and 7 emotion states were 91.625, 82.122, 82.214, 75.807, 74.946, and 72.621, correspondingly.

3. BERLIN DATABASE OF EMOTIONAL SPEECH

The Berlin's database of the emotional speech is used to classify discrete emotion states. This database is one of the most applicable ones used to recognize the speech emotion state [7] based on which several work have been presented (for example, for two of the most recent researches, we refer to [16] and [18] proposed in 2011 and [27] proposed in 2013).

This database has been implemented in the Technical University of Berlin. There are seven emotion states utilized in this database: the anger, happiness, boredom, sadness, fear, disgust, and neutral. Ten artists, including 5 men and 5 women, implemented the database in German language by saying 10 sentences, including 5 short and 5 long ones with time duration between 1.5 to 4 seconds. Samplings were done as the single channel by 16 kHz frequency. From these obtained speeches, it is possible to recognize seven real emotions of human. Audio files associated with 7 emotion states were classified as follows: anger (127), boredom (81), disgust (46), fear (69), happiness (71), sadness (62) and neutral (79).

4. FEATURE EXTRACTION

The speech feature extraction which is also called speech coding is a very important and is basic part in many of the automatic processing systems of speech. Features of the speech are generally obtained from the digital speech. To do this, various methods are utilized that the aim of it to extract the features of the speech which are useful for the desired aim of the speech automatic process. Features that we have extracted to do this research are: energy, ZCR, pitch, MFCC which are described as follows [21, 19, 27].

4.1. Mel-Frequency Cepstral Coefficients (MFCCs)

As it was stated other type of Cepstrum coefficients are Mel-frequency cepstrum coefficients (MFCCs) [20]. The basic idea of using MFCC inspirited from the properties of the human ear in understanding speech. The human ear function is in a way that its cognitive frequency varies from actually true frequency (physically) of the voice. A Mel is a step unit of measurement a step or the frequency of a sound which is heard. In 1940, Stevens and Volkmann experimentally estimated Mel scale [28].Therefore; in an experiment they called the frequency of 1000Hz by 1000Mel. Then, it was asked the listeners to change linearly the physical frequency to make cognitive frequency which double is of 1000Mel. So, they named it 2000Mel and it was repeated for 10times, 0.5times, 0.1times and etc and in next stage they labeled them 1000Mel, 500Mel and 100Mel and etc. This mapping is linear almost at frequencies below 1000 Hz, while in frequencies higher than 1000Hz, it is logarithmic [29]. In 1959, the following equation was recommended by Faunt [31]in order to express the relationship between the frequency based on Mel criteria and frequency based on Hertz criteria.

$$\text{Fmel} = 1000 \log\left(1 + \frac{F_{Hz}}{1000}\right), \qquad (1)$$

Also, in 1993, Yang presented (4-2) equation [30]

$$\text{Fmel} = 2592 \log\left(1 + \frac{F_{Hz}}{700}\right) \quad , \qquad (2)$$

Research has shown that in human auditory system, getting a certain frequency such as Ω0 which is affected by energies around Ω0 frequency is placed in a critical band where the band width can be changed by frequency. The width for frequencies below 1000 Hz is approximately 100 Hz and for frequencies above 1000 Hz, it is measured logarithmically. Thus, to find the Cepstrum coefficients, the logarithm of the total energy in the critical band around Mel frequencies has been used as input to inverse Fourier transform. For this purpose, first a filter bank is installed on the signal spectrum, so that the center frequency is distributed according to the Mel criteria.

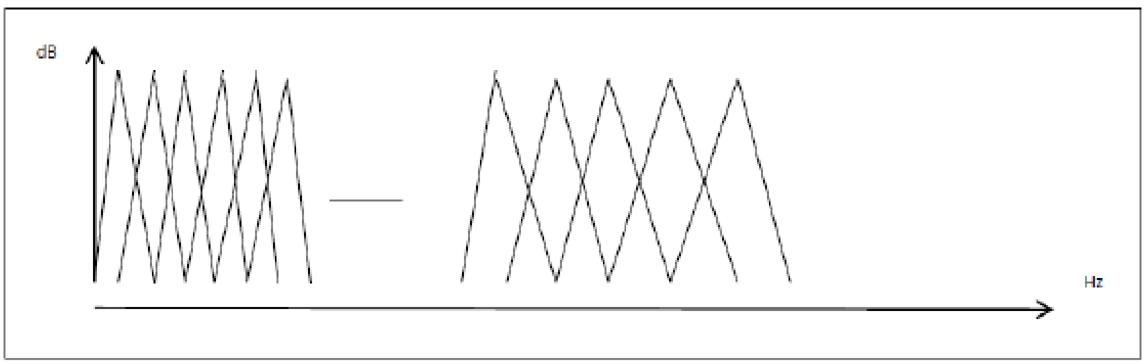

Figure 2. Installation of distributed filter bank based on the Mel criteria on the logarithm spectrum.

Therefore, to extract MFCC features, we do as follow:

- -Selecting the desired frame from the speech signal.
- Measuring the frame Fourier spectrum taking the logarithm of the amplitude.
- Installing the filter bank on the spectrum in a way that distribution of the filters has been done based on Mel criteria.
- Calculating the output of each filter in the filter bank. Measuring MFCC by using (4-3) equation:

Where P is the number of filters, N the number of MFCC coefficients, Xj the output of jth filter and MFCC coefficients are Ci. The number of chosen frame has to be a power of 2, and if not, adding 0, the frame length is given a power of 2.

4.2. Pitch Estimation

Fundamental frequency, F0, of the speech signal which is called a step and the periodicity of it, is of great importance in speech signal automatic processing. The following prosodic information and speech pitch of speech are largely determined by these parameters.

Pitch estimation algorithms can be classified into the following types:

- Algorithms that make use of time domain properties of speech signal.
- Time domain algorithm acts directly on the speech signal and by using the measured waves peaks and valleys of the wave and measuring the zero crossing rate as well as measuring autocorrelation, the step frequency period can be measured . It is assumed that if a quasi-periodic signal be processed properly, a simple estimation of time domain gives a proper approximation of the pitch.

4.2.1. Evaluation criteria of methods of step estimation

There are several criteria to evaluate a step estimation algorithm and some of them were mentioned in below:

- Estimation accuracy in step of consistency and stability of the measurement.
- Speed of operation.
- Algorithm complexity.
- Suitability for hardware implementation.
- Hardware implementation cost.

An introduction to pitch estimation method

If s(n) is nth sample of a frame in speech signal, since its aim is to estimate the step frequency and the frame has N samples, therefore, in this method a diagram of variation in autocorrelation

function amount r (η) based on η (sample frame) has been drawn. The distance of the first peak,is the amount of step period [20].

$$r(\eta) = \frac{1}{N}\sum_{n=0}^{N-1} s(n) * s(n-\eta) \qquad 0 \leq \eta \leq N-1 , \tag{4}$$

4.3. Energy

Also known as power or energy, the intensity of a voice can be physically detected through the pressure of sounds or a subjective level of noisiness. Normally, the simple intensity
is the sum of the absolute values for each data frame. The energy (E) of a signal frame of length is obtained by:

$$E = \sum_{n=0}^{N-1} s(n) * s(n) \quad , \tag{5}$$

4.5. ZCR

zcr is considered as one of the duration-related feature to be used in our experiment, which represents the number of times that the speech signals crossing the zero point. It can be easily calculated by counting the times that the wave touches the level zero reference. Instead of speech rate mentioned in the psychology, ZCR is more appropriate for language-independent speech recognition. The ZCR of a signal frame of length is obtained by:

$$ZCR = \frac{1}{2}\sum_{n=0}^{N-1} |Sign(s(n)) - Sign(s(n-1)), \quad (7)$$

In order to extract the feature, each voice is divided into windows of length 320 sections, and the overlapping size is chosen to be 20. Features have been extracted by using the MATLAB software. The calculation methods of these features are presented in the following (in all equations, and ( ) stand for the window length and the values of samples in the time domain, respectively) [20]. All the extracted features are shown in Table 2.

Table 2. Extracted features of each statement.

Feature	Normal	DFT
Pitch	Pitch_max Pitch_mean Pitch_min Pitch_median Pitch_range Pitch_std Pitch_var	fft_Pitch_max fft_Pitch_mean fft_Pitch_min fft_Pitch_median fft_Pitch_range fft_Pitch-std fft_Pitch_var
Energy	energy_max	fft_energy_max

	energy_mean energy_min energy_median energy_range energy_std energy_var	fft_energy_mean fft_enrgy_min fft_energy_median fft_energy_range fft_energy_std fft_energy_var
ZCR	zcr_max zcr_mean zcr_min zcr_median zcr_range zcr_std zcr_var	fft_zcr_max fft_zcr_mean fft_zcr_min fft_zcr_median fft_zcr_range fft_zcr_std fft_zcr_var
MFCC	mfcc1 mfcc2 mfcc3 mfcc4 mfcc5 mfcc6	mfcc7 mfcc8 mfcc9 mfcc10 mfcc11 mfcc12

5. CLASSIFIERS

Classification is the another component of a speech emotion recognition system. In this research, we used three classification methods: SVM, and C.5 as well as their ombinations. In the following of this section, we briefly explain the three major ones:

5.1. SVM Classifier

The SVM is one of the learning with observation methods which has been used in the classification and regression. In the recent years, this comparatively new method is shown to outperform the other older classification ones such as NN. This method is used to recognize speech emotion state which had resulted in a very good performance [5, 7, 17]. The base of the SVM method is to linearly classify data; and in the linear division of data, that line which has the highest safety margin is selected. By using the quadratic programming (QP) methods which are well-known in solving problems with constraints the optimum line for data is obtained. Before the linear division, in order that the machine can classify data with high complexity, data are transformed to a space with very higher dimensions by using the “phi” function. In order to solve the problem with very high dimensions by using these methods, the Lagrangian dual theorem is utilized to transform the intended minimization problem to its dual form. In the dual form, instead of the complicated high-dimension “phi” function, a simpler one called the core function is appeared. It is possible to use such various core functions as exponential, polynomial, and sigmoid [23].

5.2. C5.0 Classifier

In the recent years, C5.0 has been used as a popular method in data mining. However, it has not been used to recognize emotion till 2012. In this year, Sheikhan used the method, but the results were not satisfying enough [5]. This algorithm is the developed version of the C4.5 and ID3 [24], and it performs by either constructing the decision tree or utilizing a set of rules C5.0. C5.0 algorithm is used to make a decision tree or as rule set. A C5.0 model works by decomposing the sample based on the feature which it benefits from maximum obtained information. Every sub - sample which is defined by the first decomposition will be decomposed again and it takes place based on different fields. Decomposition process is repeated until the obtained subsamples cannot be decomposed any more. Finally, decomposition of the lowest levels is retested and the ones that do not add much value to the model can be removed or pruned. C5.0 can produce two types of models, while a decision tree is regarded as a direct description founded decomposition by algorithms. Every leaf describes a certain subset of training data, while each training data item belongs to a final node in a decision tree .In other words, for each record of proposed data, just a prediction can be done to the decision tree. In contrast, a set of rules is defined as what is attempted to do predictions for individual records. Meanwhile, a set of rules has been extracted form decision trees and so, a simplified and detailed form of the founded data in the decision tree can be provided. However, set of rules usually is not able to maintain their jobs which are decision tree features. The main difference is that in a set of rules, more than one rule may be applied in a record or maybe no rule can be applied in. If several rules are applicable in a record, each rule achieves one weight (vote) based on the amount of certainty and a final prediction is carried out by combining all weighted votes. If a rule does not apply, an arbitrary prediction can be assigned to a record.

C5.0 tree models are totally stable against problems such as missing data and plurality of input features. They usually do not require much time for estimation. Besides, understanding the C5.0 tree model is easier than the other models due to simple interpretation of the rules extracted from the tree. Moreover, C5.0 model proposes powerful Boosting methods to increase the accuracy of classification [25].

6. IMPLEMENTATION AND RESULTS EVALUATION

In this paper, we used the SVM and C5.0 classifiers as well as their combination (SVM-C5.0) in the SPSS IBM MODELER environment.Tthen we stored the data in the EXCEL environment and mined them by the SPSS IBM MODELER software. There were 54 features as the inputs which were obtained by programming in the MATLAB environment from statements of the Berlin's database. Our output was a set of emotion states (anger, happiness, boredom, sadness, fear, disgust, and neutral).

Our statements were 535, 20% of which were randomly chosen for the test while other 80% were used for the learning process. We implemented each set of data 10 times by using the SVM and C5.0 classifiers and their combination (SVM-C5.0), averaged over the obtained values of recognitions. In our first experiment, we tried to recognize two emotion states of neutral and anger.

Our total statements were 206, 127 of which were associated with the anger state and the other 79 statements were related to the neutral state. Results of each test were obtained by 10 times implementations for the following classification methods: SVM, C5.0 and SVM-C5.0. Values of the average recognitions associated with these experiments are shown in Table 3. this table shows, the recognition rate of the SVM-C5.0 , 3.6% is better than that of SVM. Also, 4.2% better is than that of C5.0.

Table3. The average recognition rates for the anger and neutral states.

COUNTER EXECUTE	1	2	3	4	5	6	7	8	9	10	Accuracy Mean
SVM	89	85	88	93	94	86	94	92	96	91	90.8
C5.0	89	89	88	88	88	94	92	90	93	91	90.2
SVM-C5.0	94	93	97	93	94	94	96	92	96	95	**94.4**

In our second experiment, we tried to recognize three emotion states of anger, happiness, and sadness. Like the former case, we used 54 features. Our total statements were 272: 128, 78, and 66 statements were accordingly associated with the anger, sadness, and happiness states. Results of each lassification test are shown in Table 4. As this table shows, the recognition rate of the SVM-C5.0 3.5% is better than that of SVM. Also, C5.0 is 5.5% better than that of SVM-C5.0.

Table 4. The average recognition rates for the anger, happiness, and sadness states.

COUNTER EXECUTE	1	2	3	4	5	6	7	8	9	10	Accuracy Mean
SVM	84	71	82	73	84	91	80	75	84	77	80.1
C5.0	90	92	83	82	90	91	88	90	94	91	**89.1**
SVM-C5.0	86	88	88	75	86	91	78	87	83	84	83.6

In the following, we will examine 4, 5, 6, and 7 emotion states whose results are presented in the table 5. This table demonstrates that for four emotion states, SVM-C5.0 classification method C5.0 and SVM methods to the extent of approximately 0.4%, and 3.9%, correspondingly. Also, for five emotion states, it outperforms C5.0 and SVM methods as much as about 7.4% and 5.3%, respectively. Furthermore, for six emotion states, it outperforms C5.0, and SVM methods to the extent of 8.9%, and 5.6%, correspondingly. Finally, for seven emotion states, it outperforms C5.0, and SVM methods as much as 4.5% and 3.5%, respectively.

Table 5. Average recognition rates for 4, 5, 6, and 7 emotion states.

MODEL	4 Emotions Anger, sadness, happiness, neutral	5 Emotions Anger, happiness, disgust, sadness, fear	6 Emotions Anger, happiness, fear, sadness, disgust, boredom	7 Emotions Anger, happiness, fear, sadness, disgust, boredom, neutral
SVM	84.9	86.5	85.9	85.1
C5.0	88.4	84.4	82.6	84.1
SVM-C5.0	**88.8**	**91.8**	**91.5**	**88.6**

7. CONCLUSION AND DISCUSSION

In this paper, we used pitch, energy, ZCR, and MFCC to recognize speech emotion states. We analyzed 535 emotional speeches from Berlin's database. A set with 54 features were calculated for each statement. Three classification methods: SVM, C5.0, and their combination (SVM-C5.0) were applied.

According to our findings, For anger, and natural emotion states, the recognition rate of SVM-C5.0 was and 3.6% better than that of SVM model and 4.2% better than that of C5.0 model. for happiness, anger, sadness states the recognition rate of SVM-C5.0 was 83.6% and 3.5% better than that of SVM model, approximately and -5.5% better than that of C5.0 model, approximately.

For more emotion states, we obtained the following results: For four emotional states the recognition rate of SVM-C5.0 was 88.8% and 0.4% better than that of C5.0 model and 3.9% better than that of SVM model, approximately. For five emotional states the recognition rate of SVM-C5.0 was 91.8% and 7.4% better than that of C5.0 model and 5.3% better than that of SVM model, approximately. For six emotional states the recognition rate of SVM-C5.0 was 91.5% and 8.9% better than that of C5.0 model and 5.6% better than that of SVM model, approximately. For seven emotional states the recognition rate of SVM-C5.0 was 88.6% and 4.5% better than that of C5.0 model and 3.5% better than that of SVM model, approximately. It is evident that the proposed SVM-C5.0 classification method is more accurate than the other ones used in this paper for two, four, five, six, and seven emotional states. Consequently, it should be discussed , according to the table, that why the accuracy of 5 and 6 emotional state is higher than 3 and 4 emotional state . Happiness and angriness diagnosis is confusing in model prediction. Thus, the accuracy rate decreases. However, this state is too low for the six senses and the following tables approved the results. For other states, accuracy rate increases and decreases.

Table 6. Confusing Matrix for the anger, happiness, and sadness states.

'Partition' = 2_Testing	A	H	S	$null$
A	24	5	0	3
H	2	7	1	0
S	1	0	15	0

Table 7. Confusing Matrix for the anger, happiness, fear, sadness, disgust, boredom.

'Partition' = 2_Testing	A	B	D	F	H	S
A	14	0	0	0	0	0
B	0	22	0	0	0	0
D	0	0	8	0	0	0
F	0	0	3	12	0	1
H	1	0	0	1	8	0
S	0	0	1	0	0	11

REFERENCES

[1] Pao T, Chen Y, Yeh J, Chang Y. Emotion recognition and evaluation of mandarin speech using weighted D-KNN classification. IntInnovComput Info Control 2008; 4: 1695- 1709.

[2] Altun H, Pollat G. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection.Expert Syst. Appl 2009; 36: 1897-8203.

[3] Yang ML. Emotion recognition from speech single using new harmony feature. Single Process 2010; 90: 1415-1423.

[4] He L, Lech M, Maddage NC, Allen NB. Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech.Biomed Signal Process Control 2011; 6, 139-146 .

[5] Sheikhan M, Bejamin M, Gharavian D. Modular neural-SVM scheme for speech emotion recognition using ANOVA feature for method. Neural Comput&Applic 2012.

[6] Schuller B, Rigoll G, Lang M. Speech emotion recognizing combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in proceeding of the ICASSP 2004; 1: 397-401.

[7] Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition 2011; 44: 572–587.

[8] Yu F, Chang E, Xu Y, Shum H. Emotion detection from speech to enrich multimedia content. In proceedings of the IEEE Pacific Rim conference on multimedia. Advances in multimedia information processing,pp. 550-557, 2001

[9] Ayadi M, Kamel S, Karray F. Speech emotion recognition using Gaussian mixture vector autoregressive models. In proceeding of the international conference on acoustics, speech, and signal processing, 5,pp. 957-960, 2007.