Speech Emotion Recognition Using Support Vector Machines

Thapanee Seehapoch[1] Sartra Wongthanavasu[2]
[12]Cellular Automata and Knowledge Engineering (CAKE) Laboratory
[2]Machine Learning and Intelligent Systems (MLIS) Laboratory
Department of Computer Science, Faculty of Science, Khon Kaen University,
Khon Kaen 40002, Thailand
[1]s.thapanee@kkumail.com, [2]wongsar@kku.ac.th

Abstract—**Automatic recognition of emotional states from human speech is a current research topic with a wide range. In this paper an attempt has been made to recognize and classify the speech emotion from three language databases, namely, Berlin, Japan and Thai emotion databases. Speech features consisting of Fundamental Frequency (F0), Energy, Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) from short-time wavelet signals are comprehensively investigated. In this regard, Support Vector Machines (SVM) is utilized as the classification model. Empirical experimentation shows that the combined features of F0, Energy and MFCC provide the highest accuracy on all databases provided using the linear kernel. It gives 89.80%, 93.57% and 98.00% classification accuracy for Berlin, Japan and Thai emotions databases, respectively.**

Keywords-component; Speech Emotion Recognitions; Support Vector Machines

I. INTRODUCTION

A way of speaking is very important in human communication. It is the most natural way to express in different story; expressing the emotion and feeling of the speaking. However, tone of voices are also the way to express the status in emotion. Sometimes, a person speaks the sentence while stay in some emotion which makes the tone of speech change the meaning of the sentence completely.

Up to date, Automatic Speech Emotion Recognition (ASER) is a very active research area in Human Computer Interaction (HCI) field and has a wide range of applications. For example, in e-learning system, computer is capable to analyze the emotions of the subject or person and adjust the substance of learning of the subject or student. In automatic remote call center, it is used to timely detect customers' dissatisfaction. In the robot's technology, by teaching robots to respond to the human and receive emotions of human, could be able to verify the human's stresses. Or even in medication, the patience's emotion is examined to diagnose the mental illness.

In recent years, many speech databases were built for speech emotion research, such as Danish Emotional Speech corpus (DES) [1], Berlin Emotional Database (EMO-DB) [2], Spanish Emotional Speech Database (SES) [3], Chinese Emotion Speech Database [4], Japanese Emotional Speech Database [5] etc. Process of the speech emotion recognition system have 2 major factors; first, feature extraction and the other, Emotion Classification. Feature extraction is the important part in finding the substitute of the tone of speech which expresses the emotion. Prosodic features and Spectral features can be used for speech emotion recognition because both of these features contain the emotional information. Many researchers tried to intercept the important features of speech such as Pitch, Energy, Formant frequency [6], Jitter, Shimmer [7], Zero Crossing Rate (ZCR) [8], Linear Predictive Coding (LPC), Linear Prediction-based Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC) [9], Postfiltered Cepstral Coefficient (PFC), Greenwood Function Cepstral Coefficient (GFCC) [10], Perceptual Linear Prediction (PLP) Cepstral Coefficients [11], and RASTA-PLP [12], etc. In work [13], lpc, jitter and energy were used as the features and it reported classification accuracy at 62.35%. In work [14] lpcc, mfcc were used and reported the classification accuracy of 83.9%. In some other works like [15], energy, zero crossing rate and fundamental frequency were used and reached classification accuracy of 97.8%. In emotion classification, many researchers explored several promising classification methods, such as K-nearest neighbor (k-NN) [16], support vector machines (SVM) [17], Neural network (NN) [18], Hidden Markov model (HMM) [19], etc.

As was investigated, a number of stated features play a vital role in accuracy performance. In addition, the promising classification models are capable of enhancing the peak accuracy rate. This paper is to investigate and integrate the integrated features to arrive at the highest accuracy performance by using Support Vector Machines.

The paper is organized as follows. Following this, section II provides the details in ASER. Section III discussed the experimentation. Results and performance comparison are given in section IV. Section V gives conclusions and discussion.

II. SPEECH EMOTION RECOGNITION SYSTEM

The structure of the speech emotion recognition system studied in this paper is depicted in figure 1. The speech signal is the first pre-process by pre-emphasis, framing and windowing. In this paper, five short time features are extracted, which are Fundamental Frequency (F0), Energy, Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC)

978-1-4673-4853-9/13/$31.00 ©2013 IEEE

and Mel Frequency Cepstral Coefficient (MFCC). Feature normalization means that the statistical features are calculated for every window of a specified number of frames by statistical method. Feature fusion is to combine different features to build different training models. Finally, Support Vector Machines (SVM) is used as emotion classifier.

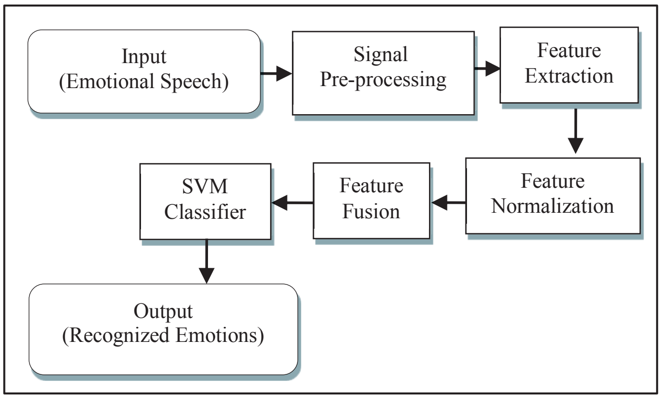

Figure 1. Speech Emotion Recognition System.

A. Signal Pre-processing

The basic operations used in the speech pre-processing [20] include the following: pre-emphasis, framing and windowing.

1) Pre-emphasis: In order to flatten speech spectrum, a pre-emphasis process is carried out on the speech signal using a high-pass Finite Impulse Response (FIR) filter given in (1).

$$H_{pre}(z) = 1 - a_{pre} z^{-1} \tag{1}$$

Adjusts the amplitude signal into normal form by taking the maximum value of the signal as nominator divided by the signal. Amplitude will be between -1 to 1.

2) Framing: The speech signal is divided into a sequence of frames where each frame can be analyzed independently and represented by a single feature vector. Frame shift is the time difference between the start points of successive frames, and the frame length is the time duration of each frame. The frame block is of length 10 msec to 40 msec from the filtered signal at every interval of 1/2 or 1/3 of frame length.

3) Windowing: In order to reduce the discontinuities of the speech signal at the edges of each frame [21], a tapered window is applied to each one. The most common used window is Hamming window, shown in (2).

$$w = 0.54 - 0.46\, cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

B. Feature Extraction

The Speech signal contains a large number of parameters that reflect the emotional characteristics, and the different parameters result in changes in emotion. Thus, the most important step in speech emotion recognition is how to extract the feature parameters, which can express mostly the emotion of speech. In our research, we calculate the statistics of five short time features, namely, Pitch, Energy, Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC). These are extracted as acoustic features for speech emotion recognition

1) Fundamental Frequency (f0): Fundamental frequency is often processed on a logarithmic scale, rather than a linear scale, to match the resolution of the human auditory system. Normally, The autocorrelation function [22] for the center clipped section is computed over a range of frequency from 50 to 500 Hz for voiced speech (the normal range of human pitch frequency).

2) Energy: It can simply be computed from the speech samples s_n [6] within the time window by (3).

$$E_v = \sum_{n=1}^{N} {s_n}^2 \tag{3}$$

3) Zero Crossing Rate: The short-time zero crossing rate is defined as the weighted average of the number of times the speech signal [8] changes sign within the time window, shown in (4).

$$zcr = \sum_{n=1}^{N} \tfrac{1}{2}[sgn(s_n) - sgn(s_{n-1})] \tag{4}$$

where $$sgn(s) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{5}$$

4) Linear Predictive Coding (LPC): Linear Prediction (LP) analysis is based on the source-filter model [20], where the vocal tract transfer function is modeled by an all-pole filter with a transfer function shown in (6).

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{6}$$

It is based on the source-filter model, where a_i is the filter coefficients. The speech signal s_n assumed to be stationary over the analysis frame which is approximated as a linear combination of the past p samples, shown in (7).

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-i} \tag{7}$$

In (7) a_i can be found by minimizing the mean square filter prediction error between $\hat{s}_n$ and s_n.

5) Mel Frequency Cepstral Coefficient (MFCC): MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Fast Fourier Transform (FFT) algorithm is ideally used for converts each frame of samples from the time domain into the frequency domain, as being described in (8).

$$S[k] = \sum_{n=0}^{N-1} s[n].e^{-\frac{j2\pi nk}{N}} \ , 0 \leq k \leq N-1 \tag{8}$$

The mel filter bank consists of overlapping triangular filters with the cutoff frequencies determined by the center frequencies of the two adjacent filters [23]. The filters have linearly spaced center frequencies and fixed bandwidth on the mel scale. The logarithm has the effect of changing multiplication into addition, described in (9).

$$F[m] = log\left(\sum_{k=0}^{N-1} |\tilde{x}[k]|^2 H_m[k]\right), 0 \leq m \leq M \tag{9}$$

Finally, The Discrete Cosine Transform (DCT) of the log filter bank energies is calculated to find the MFCC, described in (10).

$$c[n] = \sum_{m=1}^{M} F[m] \cos\left(\frac{\pi n(m-1)}{2M}\right), 0 \leq n \leq M \tag{10}$$

C. Feature Normalization

The state-segments have different lengths. In order to obtain isometric state segments and reduce redundancy of data, this paper adopts statistical method [14] to normalize the states. For each coefficient, mean, variance, median, maximum and minimum across all frames are calculated.

D. Support Vector Machines (SVM) for Emotion Classification

SVM is a non-linear classifier by transforming the input feature vectors into a generally higher dimensional feature space using a kernel mapping function. Maximum discrimination is obtained with an optimal placement of the separation plane between the borders of two classes. The plane is spanned by the support vectors leading to a reduction of references. By given a set P of points $x_i \in R^M$ with $i = 1, \dots, N$. Each point x_i belongs to either of two classes labeled $y_i \in \{-1, +1\}$. The goal is to establish the equation of a hyperplane that divides P . This purpose needs some preliminary definitions. If the set P is linearly separable there exists $w \in R^d$ and $b \in R$ to satisfy

$$y_i[(\mathrm{w} \cdot \mathrm{x}_i) + b] \geq 1, \qquad \forall i = 1, 2, \dots, N \tag{11}$$

The pair (w, b) defines a hyperplane

$$(\mathrm{w} \cdot \mathrm{x}_i) + b = 0 \tag{12}$$

This plane is called the separating hyperplane. The problem of finding the optimal separating hyperplane is converted to an optimal problem as follows:

$$\text{Minimize } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{13}$$

$$\text{subject to: } \sum_{i=1}^{N} \alpha_i y_i = 0, \;\; \alpha_i \geq 0, \;\; \forall i = 1, 2, \dots, N$$

Choosing suitable non-linear kernels, therefore, classifiers that are non-linear in the original space can become linear in the feature space. Some common kernel functions [20] are shown below:

- Linear kernel:

$$K\left(x_i, x_j\right) = x_i \cdot x_j \tag{14}$$

- Polynomial kernel:

$$K\left(x_i, x_j\right) = (x_i \cdot x_j + \beta)^d \tag{15}$$

- Radial Basis Function (RBF) kernel:

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{16}$$

A single SVM itself is a classification method for two category data. In speech emotion recognition, there are usually multiple emotion categories. Therefore we must generalize SVM to adapt to solve multi-classes. If a classification problem can be classified as N classes, any two classes among the N classes can be classified. On the contrary, in an N classification problem, if any two classes can be classified then the problem can be classified as N classes through certain combined rules of Decision Directed Acyclic Graphs (DDAG).

E. Speech Emotion Database

1) Berlin Database of Emotional Speech: It contains approximately 500 utterances spoken by actors in a happy, angry, sad, fearful, bored, disgusted and neutral version. Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all applied emotions, speech lengths in 1 to 5 seconds [2].

2) Japanese Emotional Speech Database: This database currently contains both a set of human speech with vocal emotion spoken by a Japanese male speaker and a set of artificial speech that were synthesized. They call the former 'the recorded speech' and the latter the 'synthesized speech', number of syllables in each word is 2 to 6 syllables [5].

3) Audiovisual Thai Emotion Database: This database consists of six basic emotions; happiness, sadness, surprise, anger, fear and disgust. All six students were recorded separately, so that their styles of reading do not influence on each other, and were asked to read 972 most commonly used Thai words from LINKS database (number of syllables in each word is 1 to 7 syllables). However, not all of them were useful, meaning that only words that were correctly classified by human ear were used. If an emotion from curtain file was correctly classified by at least 4 out of 5 people, the file was kept, otherwise, it was deleted [15].

III. EXPERIMENTATION

Berlin emotion database has seven emotions, namely, anger, boredom, disgust, fear, happiness, sadness and neutral version. Japan emotion database has seven emotions from ‘the recorded speech, namely, anger, boredom, disgust, fear, happiness, sadness and neutral version. Thai emotion database has six emotions, namely, anger, disgust, fear, happiness, sadness, surprise and neutral version. In the experiment design, 584, 140 and 1200 speech file from Berlin, Japan and Thai emotion databases were selected, respectively.

In step of signal pre-processing, pre-emphasis set the coefficient $a_{pre} = 0.9375$, resulting in the optimal result of filtering [25].

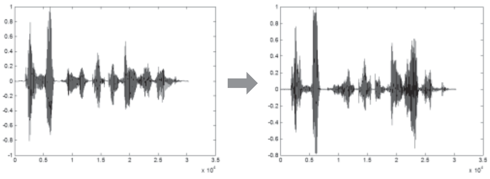

Figure 2. Pre-emphasis.

In framing, the frame blocking was made of length 30 msec or 480 sample from the filtered signal at every interval of 15 msec or 240 sample. In windowing, hamming window is then applied to each signal frame to reduce signal discontinuity in order to avoid spectral leakage.

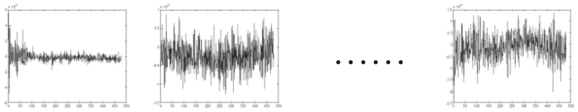

Figure 3. Frame blocking.

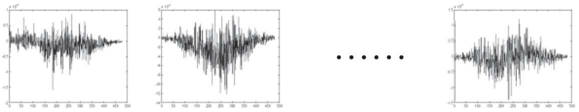

Figure 4. Hamming window.

This paper uses SVM with a number of kernels function, such as linear, quadratic, polynomial and radial basis functions with k-fold cross-validation in experimentation. Cross-validation is a common practice used in performance analysis that randomly partitions the data into N complementary subsets, with N−1 of them used for training in each validation and the remaining one used for testing.

The objective of this paper is to identify the feature set which gives the highest recognition performance using SVMs. The feature combination among F0, Energy, ZCR, LPC and MFCC are inputs for the SVM. Table I depicts a feature combination list in experimentation.

TABLE I. FEATURE COMBINATION LIST.

No.	Features	Quantity[a]
1	Fundamental Frequency (F0)	5
2	Energy	5
3	Zero Crossing Rate	5
4	LPC	105
5	MFCC	105
6	1 + 2	10
7	1 + 3	10
8	1 + 4	110
9	1 + 5	110
10	2 + 3	10
11	2 + 4	110
12	2 + 5	110
13	3 + 4	110
14	3 + 5	110
15	4 + 5	210
16	1 + 2 + 3	15
17	1 + 2 + 4	115
18	1 + 2 + 5	115
19	1 + 3 + 4	115
20	1 + 3 + 5	115
21	1 + 4 + 5	215
22	2 + 3 + 4	115
23	2 + 3 + 5	115
24	2 + 4 + 5	215
25	3 + 4 + 5	215
26	1 + 2 + 3 + 4	120
27	1 + 2 + 3 + 5	120
28	1 + 2 + 4 + 5	220
29	1 + 3 + 4 + 5	220
30	1 + 2 + 3 + 4 + 5	225

[a] Number of features.

IV. RESULTS

The SVMs is trained and tested on five feature vectors using each kernel function. The experimentation is carried out by varying cost values. Moreover, among the entire kernel functions, notice that linear kernel (linear), polynomial kernel at degree 3 (poly-3) and RBF kernel at sigma 7 (rbf-7) are the first three to give the best results.

From experimental result shown in Table II – IV, MFCC features gives the best recognition rate. It provides classification accuracy at 78.04%, 89.29% and 92.42% for Berlin, Japan, and Thai emotion databases, respectively.

TABLE II. RECOGNITION RATE OF BERLIN EMOTION DATABASE ON FIVE FEATURES AND THREE KERNELS.

No.	Features	Accuracy (%)		
		linear	*poly-3*	*rbf-7*
1.	F0	53.47	62.43	52.07
2.	Energy	62.26	67.83	58.19
3.	ZCR	51.10	57.68	46.50
4.	LPC	58.45	55.03	53.57
5.	MFCC	**78.04**	73.75	71.44

TABLE III. RECOGNITION RATE OF JAPAN EMOTION DATABASE ON FIVE FEATURES AND THREE KERNELS.

No.	Features	Accuracy (%)		
		linear	*poly-3*	*rbf-7*
1.	F0	70.71	80.71	67.86
2.	Energy	71.43	64.29	67.14
3.	ZCR	39.29	44.29	32.14
4.	LPC	59.29	47.14	65.00
5.	MFCC	**89.29**	64.29	87.14

TABLE IV. RECOGNITION RATE OF THAI EMOTION DATABASE ON FIVE FEATURES AND THREE KERNELS.

No.	Features	Accuracy (%)		
		linear	*poly-3*	*rbf-7*
1.	F0	62.08	62.75	59.25
2.	Energy	61.83	72.08	56.42
3.	ZCR	75.00	76.25	66.33
4.	LPC	81.92	76.17	78.17
5.	MFCC	**92.42**	87.67	91.00

For comprehensive performance evaluation, the empirical experiments in entire variant features combination and all possible different kernel functions are implemented. In this regard, it reports that the feature set of F0 + Energy + MFCC with linear kernel provides the highest accuracy of recognition. Table V - VII depicted the results of classification from Berlin, Japan and Thai Emotion databases at 89.80%, 93.57% and 98.00%, respectively.

TABLE V. CONFUSION MATRIX FOR THE FEATURE SET F0 + ENERGY + MFCC OF BERLIN EMOTION DATABASE.

Emotions	Recognized Emotions (%)						
	Angry	*Bored*	*Disgust*	*Fear*	*Happy*	*Natural*	*Sad*
Angry	**92.86**	0	0	0	7.14	0	0
Bored	0	**88.75**	1.25	0	0	6.25	3.75
Disgust	2.27	2.27	**86.36**	6.82	0	2.27	0
Fear	2.99	1.49	1.49	**89.55**	4.48	0	0
Happy	7.04	0.00	4.23	11.27	**77.46**	0	0
Natural	0	2.56	1.28	3.85	0	**92.31**	0
Sad	0	3.23	0	0	0	0	**96.77**

TABLE VI. CONFUSION MATRIX FOR THE FEATURE SET F0 + ENERGY + MFCC OF JAPAN EMOTION DATABASE.

Emotions	Recognized Emotions (%)						
	Angry	*Bored*	*Disgust*	*Fear*	*Happy*	*Natural*	*Sad*
Angry	**100**	0	0	0	0	0	0
Bored	0	**95**	0	0	0	5	0
Disgust	0	0	**85**	0	10	5	0
Fear	0	0	0	**95**	5	0	0
Happy	0	0	5	5	**90**	0	0
Natural	0	5	5	0	0	**90**	0
Sad	0	0	0	0	0	0	**100**

TABLE VII. CONFUSION MATRIX FOR THE FEATURE SET F0 + ENERGY + MFCC OF THAI EMOTION DATABASE.

Emotions	Recognized Emotions (%)					
	Angry	*Disgust*	*Fear*	*Happy*	*Sad*	*Surprise*
Angry	**96.5**	3.5	0	0	0	0
Disgust	3.5	**95.5**	0	0	1	0
Fear	0.5	0	**98.5**	0	0.5	0.5
Happy	0.5	0	0	**99.5**	0	0
Sad	0	1.5	0	0	**98**	0.5
Surprise	0	0	0	0	0	**100**

V. CONCLUSION

This paper aims at determining the optimal basic feature integration in speech emotion recognition using SVM-based classification. It implements in three emotion recognition databases, namely, Berlin, Japanese and Thai. In this regard, the feature combination of F0, energy and MFCC implementing in SVM with linear kernel provides the best results for emotional classification. It reports the classification accuracy at 89.80%, 93.57% and 98.00% for Berlin, Japan and Thai emotion databases, respectively. From this research, we can conclude the result in several ways. Firstly, Speech Emotion Recognition System is high recognition rate that uses both prosodic and spectral features. Secondly, Japan and Thai language aren't variant tone language. So, when speakers express their emotions its will be more obviously and better communication than German language which is an over/low tone language. Finally, Japan and Thai databases record 1-7 syllables while German database records by sentence, Therefore, if the speech is too long, the emotion is also difficult to recognize.

REFERENCES

[1] I. S. Engberg, and A. V. Hansen, "Documentation of the Danish Emotional Speech Database (DES)", Internal AAU report, Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark, September 1996.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A database of German emotional speech", Proc. Interspeech, 2005.

[3] J. M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, and J. M. Pardo, “Analysis and modelling of emotional speech in Spanish”, in Proc. ICPhS’99, pp. 957-960, San Francisco 1999.

[4] F. Yu, E. Chang, Y.Q. Xu, and H.Y. Shum, “Emotion detection from speech to enrich multimedia content”, in Proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp.550-557, Beijing, China, October 2001.

[5] Tsuyoshi Moriyama, Shinya Mori, and Shinji Ozawa. “A synthesis method of emotional speech using subspace constraints in prosody”. Journal of Information Processing Society of Japan, 50(3):1181–1191, 2009. (in Japanese).

[6] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification”, in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004.

[7] J. Kreiman and B. R. Gerrat, “Perception of aperiodicity in pathological voice”, Acoustical Society of America, vol.117, pp. 2201-2211, 2005.

[8] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” Bell System Technical Journal, vol. 54, no. 2, pp. 297-315, February 1975.

[9] Specht, D. F., “Probabilistic neural networks for classification, mapping or associative memory”, Proceedings of IEEE International Conference on Neural Network, Vol. 1, pp.525-532, Jun. 1988.

[10] Johnson, M.T., Clemins, P.J., Trawicki, M.B., “Generalized Perceptual Features for Vocalization Analysis Across Multiple Species”, ICASSP.2006 Proceedings.,2006.

[11] Hermansky, H., “Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, Vol. 87, No. 4, pp. 1738-1752, 1990.

[12] Hynek Hermancky, N.M., Aruna Bayya and Phil Kohn, “RASTA-PLP Speech Analysis”. 1991.

[13] Aishah Abdul Razak,Ryoichi Komiya and Mohamad Izani Zainal Abidin, “Comparison Between Fuzzy and NN Method for Speech Emotion Recognition”, Proc. of the Third International Conference on Information Technology and Applications, pp.297 – 302, July. 2005.

[14] Xia Mao, Lijiang Chen and Bing Zhang, “Mandarin speech emotion recognition based on a hybrid of HMM/ANN”, INTERNATIONAL JOURNAL of COMPUTERS Volume 1, pp.321-324, 2007.

[15] Stankovic, I., Karnjanadecha, M., and Delic, V., “Improvement of Thai speech emotion recognition by using face feature analysis”, Proceedings of the Nineteenth IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2011), Chiang Mai, Thailand, December 7-9, pp. 87, 2011.

[16] Dellaert, F., Polzin, T. & Waibel, A., “Recognizing emotion in speech”, Fourth International Conference on Spoken Language Processing, Vol. 3, pp. 1970-1973, Oct. 1996.

[17] Vapnik, V., “The nature of statistical learning theory”, Springer-Verlag, 1995, ISBN 0-387-98780-0, 1995.

[18] Nicholson, J., Takahashi, K. & Nakatsu, R., “Emotion recognition in speech using neural networks”, 6th International Conference on Neural Information Processing, Vol. 2, pp. 495–501, 1999.

[19] Rabiner, L.R., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proceedings of IEEE, 77(22): pp. 257-286, 1989.

[20] Ling Cen, Minghui Dong, Haizhou Li, Zhu Liang Yu and Paul Chan., “Machine Learning Methods in the Application of Speech Emotion Recognition”. Intech Publications, ISBN 978-953-307-035-3, Feb. 2010.

[21] Joseph W. Picone, “Signal modeling techniques in speech recognition”, Proc. of the IEEE, Vol. 81, No. 9, pp.1215-1245, Sep. 1993.

[22] Sondhi, M., “New Methods of pitch extraction”, IEEE Trans. ASSP, 16(2): pp.262-266, 1968.

[23] Peipei Shen, Zhou Changjun and Xiong Chen. “Automatic Speech Emotion Recognition Using Support Vector Machine”. Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference, pp.859-862, Aug. 2011.

[24] Richard O. Duda, Peter E. Hart and David G. Stork. “PATTERN CLASSIFICATION”. 2nd ed. New York : Wiley-Interscience, pp.128-138, Oct. 2000.

[25] Milan Sigmund, “Voice Recognition By Computer”, Tectum Verlag publication, pp.20-22.