A Music Recommendation System with Consideration of Personal Emotion

Chuan-Yu Chang[1*], Chun-Yen Lo[1], Chi-Jane Wang[2] and Pau-Choo Chung[3]
[1]Department of Computer Science and Information Engineering,
National Yunlin University of Science and Technology
[2]Department of Nursing, National Cheng Kung University
[3]Department of Electrical Engineering, National Cheng Kung University
[*]chuanyu@yuntech.edu.tw

***Abstract*—Emotions evoked by were analyzed. An estimation of the correlation coefficient was applied to determine features of music that evoke an emotion. These features were then used to train two support vector machines (SVMs) for an individual subject to classify music that evokes happiness, anger, sadness, and peacefulness. The proposed approach can be used to classify music that evokes an emotion and to build a personal emotion-cognitive music recommendation system for an individual subject. Experiment results show the effectiveness of the proposed approach.**

Keywords: emotion recognition; music recommendation

I. INTRODUCTION

Listening to music can be relaxing. However, not all music is effective in changing emotions. An automatic emotion-cognitive music recommendation system can be used to select music that is appropriate for a given mood [1-2]. A recommendation system analyzes music and the emotions of the subject. Analyzing emotions evoked by music and classifying them are difficult tasks [3]. Emotion-cognition is subjective. The emotion evoked by music greatly depends on the person. This subjectivity makes evaluation of an emotion-cognitive music classification system difficult to implement. In order to reduce subjectivity, a subject-dependent classification approach, where a respective classification system is applied to an individual subject, is proposed.

In order to discover the relationship between music and the emotion that it may evoke, emotions have been categorized into many classes and pattern recognition procedures have been applied to classify the music [3-4]. Emotions such as happiness, anger, and sadness have been classified using various emotion models, such as Thayer's model [5], the arousal-valence model [6], and Russell's model [7]. Yang *et al*. proposed a combination model, which utilized Thayer's arousal-valence emotion plane (see Fig. 1), that integrates the arousal-valence model [6] and Russell's model [7]. Based on Thayer's arousal-valence emotion plane, four types of emotion, namely happiness, nervousness, sadness, and peacefulness, can be classified. Arousal is the intensity of the emotion response, which ranges from low to high, and valence is the degree of positive or negative emotion. Thayer's arousal-valence emotion plane is adopted to classify emotions in the present study.

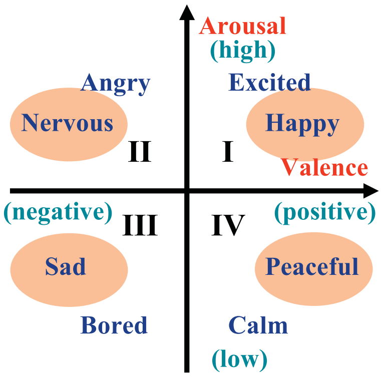

Figure 1. Thayer's arousal-valence emotion plane.

Many features have been proposed to describe music, such as linear prediction coefficients (LPC), linear prediction cepstrum coefficients (LPCC) [8-9], Mel-frequency cepstrum coefficients (MFCC) [2, 8-10], entropy and dynamism [11], timbre [1-2, 4, 8, 11], intensity [4], rhythm [4, 12], pitch [12], amplitude envelop [9], and Daubechies wavelet coefficient histograms [2]. Although these features can be directly applied, they characterize music over a short time period. We believe that these short-term features may neglect potential properties in long-term information that can evoke emotions in subjects. A long-term approach, where features of music are organized in series as sequences of features, is thus proposed in this study.

To classify sequences for respective types of emotion, a classifier is required. In previous studies, a number of classifiers have been used, including support vector machines (SVMs) [8], support vector regression (SVR) [3, 9], fuzzy C-mean [13], gaussian mixture models (GMM) [4], multi-layer perceptrons (MLPs), hidden markov model (HMM) [11], K-nearest neighbor (KNN) [12, 14], AdaBoost [1], and radial basis function neural networks (RBFNNs) [8]. Among these classifiers, SVM achieves the highest classification accuracy [15]. Therefore, it was adopted as the classifier in the present study.

The rest of this paper is organized as follows. Section II describes the proposed approach including feature extraction, feature selection, correlation coefficients, the support vector machine, and classification. Section III presents experimental results. The conclusion is given in Section IV.

978-1-4244-7640-4/10/$26.00 ©2010 IEEE

II. Proposed Method

Fig. 2 illustrates main processes of the proposed approach. Features are extracted from non-overlapping frames. A feature selection approach is applied to obtain the optimal feature set. Then, highly correlated sequences are selected to train the classifier. The respective processes of the proposed approach are described in the following sub-sections.

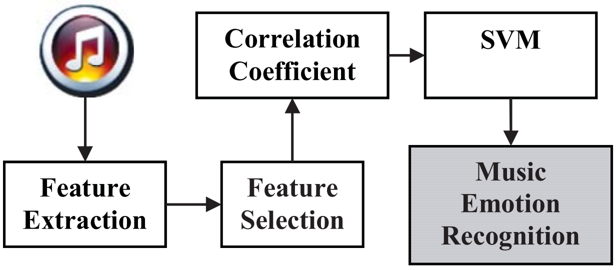

Figure 2. Processes of the proposed analysis approach.

A. Feature Extraction

Lie *et al*. and Huron *et al*. pointed out that the intensity feature is necessary for emotion detection [4, 16]. The intensity feature is simple and easy to compute. Timbre refers to audio signal information. Timbre features are usually described by centroid, bandwidth, roll-off, flux, zero crossing rate, and octave-based spectral contrast [1-2, 4, 9, 12] and computed based on a short-time Fourier transform.

Before features are acquired in the frequency domain, each signal frame is multiplied by a pre-emphasized Hamming window. The operation is defined as follows:

$$x_i'(n) = x_i(n) \times h(n) \tag{1}$$

where $x_i(n)$ is the input signal of the i -th frame and $h(n)$ is a pre-emphasized Hamming window defined as follows [8, 10] :

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

where $0 \le n \le N-1$ and N is the number of points in each frame.

The fast Fourier transform (FFT) is then applied on each frame to obtain the corresponding spectrum. Since the FFT transforms signals in the time domain to the frequency domain, we preserve some features in the time domain.

The intensity features and the timbre features are defined as follows:

Intensity:

$$I_i = \sum_{j=0}^{N-1} \left| x_i'(j) \right| \tag{3}$$

where $x_i'(j)$ is the j -th sample point in the i -th frame.

Average:

$$A_i = \frac{1}{N} \sum_{j=0}^{N-1} x_i'(j) \tag{4}$$

Variance:

$$V_i = \sum_{j=0}^{N-1} \left(x_i'(j) - A_i \right)^2 \tag{5}$$

Maximum:

$$M_i = \max\{ x_i'(j) \mid 0 \le j < N \} \tag{6}$$

Centroid:

$$C_i = \sum_{j=0}^{N-1} x_i'(j) \times j \Big/ \sum_{j=0}^{N-1} x_i'(j) \tag{7}$$

Bandwidth:

$$B_i = \sqrt{ \sum_{j=0}^{N-1} \left(\left| x_i'(j) \right|^2 \times (j - C_i) \right) \Big/ \sum_{j=0}^{N-1} \left| x_i'(j) \right|^2 } \tag{8}$$

Rolloff:

$$\sum_{j=0}^{R_i} \left| x_i'(j) \right|^2 = 0.95 \times \sum_{j=0}^{N-1} \left| x_i'(j) \right|^2 \tag{9}$$

where R_i is below 95% of the power distribution.

Flux:

$$F_i = \sum_{j=0}^{N-1} \left(\left| x_i'(j) \right| - \left| x_{i-1}'(j) \right| \right)^2 \tag{10}$$

Zero Crossing Rate:

$$ZCR_i = \sum_{j=0}^{N-1} \left| \operatorname{sgn}\left(x_i'(j) \right) - \operatorname{sgn}\left(x_{i-1}'(j) \right) \right| \tag{11}$$

where the sgn function is 1 for positive arguments and 0 for negative arguments.

LPCC and MFCC are general features used in audio signal recognition and classification [2, 8-10, 17-19]. They are robust and reliable.

Twenty-one features are extracted in each frame. They are average (TA), variance (TV), intensity (TI), centroid (TC), bandwidth (TB), rolloff (TR), and flux (TF) in the time domain, and LPCC and MFCC with thirteen coefficients; peak (FSP), valley (FSV), and contrast (FSC) with seven sub-band; average (FA), variance (FV), intensity (FI), maximum (FM), centroid (FC), bandwidth (FB), rolloff (FR), flux (FF), and zero crossing rate (FZCR) in the frequency domain. The features extracted from each frame are summarized in Table I.

TABLE I. EXTRACTED FEATURES

Time Domain		**Frequency Domain**	
Intensity	Average Variance Intensity	***Intensity***	Average Variance Maximum Intensity
Timbre	Centroid Bandwidth Rolloff Flux	***Timbre***	Centroid Bandwidth Rolloff Flux Peak Valley Contrast Zero Crossing Rate
		LPCC	
		MFCC	

B. Feature Selection

As mentioned above, 21 features are extracted in each frame. However, extracting these features and using them to directly train a classifier is very time-consuming. To reduce the time required and to improve accuracy, it is necessary to select significant features. Thus, a feature selection algorithm called sequential floating forward selection (SFFS) [20] is utilized to find discriminative features.

Sequential forward selection (SFS) and sequential backward selection (SBS) are step-optimal only since the best (worst) feature is always added (discarded) in the algorithms. Both methods suffer from the so-called nesting effect [20]. In the case of the top-down search, discarded features cannot be reselected, whereas in the case of the bottom-up search, a selected feature cannot be discarded later. The SFFS algorithm is a revised algorithm based on sub-optimal feature subset selection. SFFS is superior to SFS and SBS in the ability of dynamically determining the number of forward/backward steps during the search process.

Seven features (TV, TC, FA, FV, FM, FR, and FF) were selected for the arousal SVM, and nine features (TA, TV, TB, TR, TF, FA, FV, FC, and FR) were selected for the valence SVM. The arousal SVM and valence SVM are described in sub-section D.

C. Correlation Coefficient

In order to determine the relationship between music and emotions, the correlation coefficient is used for sequences of music that evoke emotion. The Pearson product-moment correlation coefficient [21] is applied. The correlation coefficient of the Pearson product-moment is a statistics method, which describes the degree of linear correlation between two groups. The coefficient is defined as follows:

$$r^f = \frac{S_{ab}^f}{\sqrt{S_{aa}^f S_{bb}^f}}, -1 \le r^f \le 1 \tag{12}$$

where

$$S_{ab}^f = \sum_{i=1}^{n} \left(a_i^f - \bar{a}^f\right)\left(b_i^f - \bar{b}^f\right) \tag{13}$$

$$S_{aa}^f = \sum_{i=1}^{n} \left(a_i^f - \bar{a}^f\right)^2 \tag{14}$$

$$S_{bb}^f = \sum_{i=1}^{n} \left(b_i^f - \bar{b}^f\right)^2 \tag{15}$$

where n is the length of the sequence; a_i^f and b_i^f are the f -th feature in the i -th frame in group *A* and group *B*, respectively. $\bar{a}^f$ and $\bar{b}^f$ are mean values of the f -th feature in the sequence.

For each sequence, correlations between the sequence and all other sequences are computed. Since the number of sequences with positive correlation is generally large, a threshold (*T*) is set. The parameter *k* is used to determine the number of training samples.

D. Support Vector Machine

The support vector machine (SVM) is a supervised learning machine that attempts to find an optimal hyperplane to separate two classes and to produce a classifier that will work well on unknown patterns.

From Fig. 1, emotions happiness and peacefulness have a positive valence, and emotions nervousness and sadness have a negative valence. Similarly, emotions nervousness and happiness are classified as high-arousal, and emotions sadness and peacefulness are classified as low-arousal. In this study, two SVMs (arousal SVM and valence SVM) were trained to determine low or high arousal, and positive or negative valence, respectively. The SVMs were implemented using the LIBSVM package [22]. The kernel function of the SVMs was the radial-basis kernel function.

E. Music Emotion Recognition

In the testing phase, specific features are extracted in each frame according to the feature selection results. After a comparison with the trained emotional-music sequences, emotional-music sequences with positive correlation are extracted from the tested melodies. These extracted emotional-music sequences are then inputted to the arousal SVM and valence SVM. Results of the arousal SVM may be high (+) or low (-), whereas results of the valence SVM may be positive (+) or negative (-), as shown in Fig. 3. If the arousal SVM classifies a music sequence as high (+) and the valence SVM classifies it as negative (-), the final result is "nervousness". Similarly, if the arousal SVM classifies a music sequence into low (-) and the valence SVM classifies it as negative (-), the final result is “sadness”.

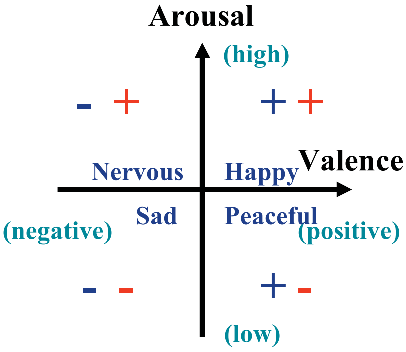

Figure 3. Determination of emotion. Blue signs are results of the valence SVM. Red signs are results of the arousal SVM.

III. Experimental Results

293 music files were tested, including piano performances, symphonic music, pan flute performances, harp performances, and hip hop music. The music (melodies) contained only instruments (no vocals). All tracks were encoded in 44.1 KHz standard stereo WAV format.

The subjects were asked to relax themselves with their eyes closed for two minutes. The system played a two-minute music file for the subjects to listen to. Once the music stopped, the subjects were asked to identify the type of emotion that the music had evoked. Between tracks, the subjects were asked to relax themselves for two minutes again. The experiment continued until at least six tracks for each emotion were identified. A total of 24 emotion-cognitive tracks were thus obtained for each subject.

To evaluate the performance of the proposed method, three experiments were performed. (1) Determination of the three parameters in emotional music recognition; these parameters are the length (*L*) of the emotional music sequence, the threshold of correlation (*T*), and the number of training samples *k*. (2) A comparison between the proposed method and another emotional music recognition approach. (3) Performance of the proposed method in a music recommendation system.

A. Determination of three parameters

In the proposed method, the length (*L*) of the emotional music sequence, the threshold (*T*) of the correlation coefficient, and the number of training samples *k* should be set appropriately to obtain high classification accuracy.

The length of the sequence determines how the music is sampled. The threshold of the correlation coefficient determines how the correlated segments are preserved. The parameter *k* determines the number of sequences that should be used to train the SVMs.

To determine the three parameters, we randomly selected three tracks from each emotion type to train the SVMs. A total of 12 tracks were used for each subject. The recognition rate is defined as:

$$\mathit{recognition\ rate} = \frac{A_P}{A_P + A_N} \tag{16}$$

where A_P denotes the number of tracks that the system properly identified, and A_N denotes the number of tracks that the system missed.

TABLE II. Music emotion recognition rate (%) for *K*=1

Threshold / Length(s)	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
8	N/A	33.85	43.08	43.08	43.08	43.08	43.08	43.08	43.08	43.08
9	N/A	N/A	46.15	40.77	41.54	42.31	42.31	42.31	42.31	42.31
10	N/A	N/A	30.77	38.46	40.77	40.77	40.77	40.77	40.77	40.77
11	N/A	N/A	N/A	39.23	36.92	37.69	37.69	37.69	37.69	37.69
12	N/A	N/A	N/A	33.85	42.31	39.23	39.23	40.00	39.23	39.23
13	N/A	N/A	N/A	N/A	43.08	38.46	40.00	40.77	40.77	40.77
14	N/A	N/A	N/A	N/A	35.38	40.77	40.77	41.54	41.54	41.54
15	N/A	N/A	N/A	N/A	N/A	39.23	40.00	38.46	37.69	37.69
16	N/A	N/A	N/A	N/A	N/A	44.62	36.92	39.23	39.23	39.23

TABLE III. Music emotion recognition rate (%) for *K*=4/5

Threshold / Length(s)	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
8	N/A	45.38	36.92	33.08	34.62	36.15	35.38	37.69	32.31	33.85
9	N/A	N/A	34.62	36.92	35.38	36.15	33.85	36.92	36.92	37.69
10	N/A	N/A	46.15	29.23	31.54	35.38	35.38	36.92	34.62	34.62
11	N/A	N/A	29.23	40.00	30.77	32.31	33.85	28.46	33.85	31.54
12	N/A	N/A	N/A	73.08	41.54	35.38	34.62	37.69	36.15	30.77
13	N/A	N/A	N/A	36.92	34.62	37.69	35.38	36.92	33.85	34.62
14	N/A	N/A	N/A	29.23	41.54	35.38	33.85	32.31	34.62	36.15
15	N/A	N/A	N/A	N/A	49.23	34.62	28.46	29.23	34.62	35.38
16	N/A	N/A	N/A	N/A	34.62	36.92	36.15	31.54	31.54	35.38

The average music emotion recognition rates under various settings are listed in Tables II, III, IV, and V. In these tables, N/A means that the system could not acquire a sequence, and values represent the recognition rate of the system. The highest recognition rate was obtained when the threshold was set to 0.8, the sequence length was set to 12, and *k* was set to 4/5. Hence, subsequent experiments were carried out with *T*=0.8, *L*=12, and *k*=4/5. The results in Tables II-IV show that when the threshold is too small, the correlation of the acquired sequences will be low, resulting in low accuracy. Although a larger threshold can produce more correlated sequences, the number of obtained sequences will be small.

TABLE IV. MUSIC EMOTION RECOGNITION RATE (%) FOR K=3/5

Length(s) \ Threshold	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
8	N/A	29.23	33.85	36.15	34.62	33.85	35.38	34.62	31.54	36.92
9	N/A	38.46	37.69	33.85	34.62	32.31	36.92	37.69	33.08	38.46
10	N/A	N/A	34.62	35.38	33.85	33.08	36.15	35.38	34.62	35.38
11	N/A	N/A	45.38	31.54	27.69	28.46	35.38	32.31	33.85	35.38
12	N/A	N/A	36.92	36.92	31.54	31.54	36.15	37.69	34.62	38.46
13	N/A	N/A	N/A	53.08	31.54	32.31	33.85	35.38	33.08	36.92
14	N/A	N/A	N/A	52.31	43.85	37.69	32.31	35.38	33.08	37.69
15	N/A	N/A	N/A	33.85	38.46	33.85	31.54	29.23	29.23	36.15
16	N/A	N/A	N/A	N/A	42.31	31.54	39.23	30.77	32.31	36.15

TABLE V. MUSIC EMOTION RECOGNITION RATE (%) FOR K=2/5

Length(s) \ Threshold	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
8	30.00	20.00	20.77	23.08	21.54	20.00	23.08	20.77	19.23	23.08
9	N/A	27.69	24.62	23.85	22.31	20.00	23.08	23.85	19.23	26.15
10	N/A	26.15	23.85	20.00	21.54	20.00	23.08	22.31	22.31	23.85
11	N/A	23.08	24.62	14.62	19.23	16.92	22.31	20.77	20.77	23.85
12	N/A	N/A	20.77	24.62	25.38	18.46	23.08	23.08	22.31	24.62
13	N/A	N/A	30.77	24.62	24.62	24.62	22.31	23.08	20.77	23.08
14	N/A	N/A	29.23	28.46	22.31	21.54	20.77	23.08	20.77	23.85
15	N/A	N/A	N/A	30.00	23.85	20.77	19.23	18.46	18.46	24.62
16	N/A	N/A	N/A	33.85	23.08	20.00	22.31	19.23	19.23	23.85

B. Comparison with another emotional music recognition approach

In order to demonstrate the capability of the proposed approach, we compared our method with Yang *et al.*'s approach [3]. They used the SVR as their regressor. In order to reduce subjectivity, subjects were divided into a variety of user groups. A lot of regressors were trained for each group. The groups were defined according to subject information such as generation, sex, occupation, and personality to reduce individual differences for each group. Yang *et al.* used Psysound [23] to extract 15 features, and then used these features to train two regressors for each group. We applied their method on our collection database to train two regressors for each subject as a group.

The comparison results are shown in Table VI. The average recognition rate of Yang's method, the arousal SVM, and the valence SVM are 73.08%, 81.54%, and 85.38%, respectively. The results show that our method outperforms Yang's method. For a group of subjects, Yang *et al.* used the same tracks to train two regressors. However, these tracks may be insufficient to induce emotion in a subject. Hence, using these tracks to train two regressors may lose some information that can evoke emotion.

TABLE VI. COMPARISON WITH YANG'S METHOD

Model \ Methods	Proposed Method	Yang *et al.*[3]
Music emotion recognition	73.08	33.85
Arousal	81.54	48.46
Valence	85.38	60.77

C. Performance of the proposed method in a music recommendation system

TABLE VII. MUSIC RECOMMENDATION RESULTS

Subject \ Emotion	Happy		Nervous		Sad		Peaceful		Hit Rate (%)
	Predict	Hit	Predict	Hit	Predict	Hit	Predict	Hit	
1	9	4	1	1	2	2	3	3	66.67
2	12	8	0	0	1	1	2	2	73.33
3	2	2	3	3	10	7	0	0	80.00
4	2	1	4	4	8	7	1	1	86.67
5	3	1	6	3	5	2	1	1	46.67
6	4	1	2	2	3	2	6	5	66.67
7	0	0	5	5	1	1	9	5	73.33
8	0	0	1	1	14	6	0	0	46.67
9	6	4	1	1	4	4	4	3	80.00
10	1	1	6	5	8	5	0	0	73.33
11	1	1	6	4	3	2	5	2	60.00
12	6	4	4	3	1	1	4	3	73.33
13	2	2	11	5	2	2	0	0	60.00
Accuracy (%)	60.42		74.00		67.74		71.43		68.21

To build a personal emotion-cognitive music database for each subject, the SVMs trained in Experiment A were used. The untrained 281 music files were categorized into four emotions. Then, we randomly selected 15 tracks from personal emotion-cognitive music database for the each subject. 13 subjects participated this experiment.

$$hit\ rate = \frac{Total\ number\ of\ hit}{Total\ number\ of\ hits + Total\ number\ of\ misses} \quad (17)$$

The hit rate is defined as shown in equation 17. The average recommendation results are shown in Table VII, where *predict* denotes the number of tracks that the system recommended, and *hit* denotes the number of tracks for which the subject agreed with the recommendation. For example, the system recommend nine emotion-cognitive tracks for happiness, one for nervousness, two for sadness, and three for peacefulness. However, subject 1 only agreed that four tracks were happy, one was nervous, two music were sad, and three were peaceful. Hence, the average hit rate is 66.67 % for subject 1. The overall average hit rate is 68.21%.

IV. CONCLUSION

In this study, we proposed a correlation-coefficient-based approach to find emotional music sequences which may evoke a specific emotion in subjects. The SFFS method is applied to select significant music features from emotional music sequences. The selected features are used to train SVM classifiers for an individual subject. Hence, we can classify unknown music that evokes an emotion, and can build a personal emotion-cognitive music database for an individual subject. The proposed system can recommend music from a personal emotion-cognitive music database. Experimental results show that the proposed method achieves high classification accuracy, and that the recommended music is close to a subject's emotion perception.

ACKNOWLEDGMENT

This work was supported by the National Science Council, Taiwan, under grant NSC 97-3114-E-006-001.

REFERENCES

[1] X. Zhu, Y. Y. Shi, H. G. Kim, and K. W. Eom, "An integrated music recommendation system," *IEEE Transactions on Consumer Electronics,* vol. 52, pp. 917-925, 2006.

[2] B. Shao, M. Ogihara, D. Wang, and T. Li, "Music Recommendation Based on Acoustic Features and User Access Patterns," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, pp. 1602-1611, 2009.

[3] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, pp. 448-457, 2008.

[4] L. Lie, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, pp. 5-18, 2006.

[5] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York: Oxford University Press, 1989.

[6] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*: University of Illinois Press, 1957.

[7] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology,* pp. 1161-1178, December 1980.

[8] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using SVM and RBFNN," *Expert Systems with Applications,* vol. 36, pp. 6069-6075, 2009.

[9] X. Changsheng, N. C. Maddage, and S. Xi, "Automatic music classification and summarization," *IEEE Transactions on Speech and Audio Processing,* vol. 13, pp. 441-450, 2005.

[10] C. H. Lee, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features," *IEEE Transactions on Multimedia,* vol. 11, pp. 670-682, 2009.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication,* vol. 40, pp. 351-363, 2003.

[12] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing,* vol. 10, pp. 293-302, 2002.

[13] D. C. Park, "Classification of audio signals using Fuzzy c-Means with divergence-based Kernel," *Pattern Recognition Letters,* vol. 30, pp. 794-798, 2009.

[14] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Transactions on Multimedia,* vol. 8, pp. 564-574, 2006.

[15] P. C. Chung and H. J. Kuo, "Using feature selection with support vector machine in gastric histology classification," Master Thesis, National Cheng Kung University, 2004.

[16] D. Huron, "The Ramp Archetype and the Maintenance of Passive Auditory Attention," *Music Perception,* vol. 10, pp. 83-91, 1992.

[17] S. Changwoo, L. K. Yong, and L. Joohun, "GMM based on local PCA for speaker identification," *Electronics Letters,* vol. 37, pp. 1486-1488, 2001.

[18] R. J. Mammone, Z. Xiaoyu, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Processing Magazine,* vol. 13, pp. 58-71, 1996.

[19] G. Velius, "Variants of cepstrum based speaker identity verification," in *International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 583-586 vol.1.

[20] P. Pudil, J. Novovicová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters,* vol. 15, pp. 1119-1125, 1994.

[21] K. Pearson, "Mathematical contributions to the theory of evolution. -III. Regression, heredity and panmixia," *Philos. Trans. R. Soc. London,* vol. 187, pp. 253-381, 1896.

[22] C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines, 2001. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[23] D. Cabrera, "PSYSOUND: A computer program for psychoacoustical analysis," in *Proceedings of the Australian Acoustical Society Conference*, pp. 47-54, 1999.