Emotion Extraction and Recognition from Music

Fan Zhang[1] and Hongying Meng[1]

[1]Dept. of Electronic and Computer Engineering
Brunel University London
Uxbridge, UB8 3PH, UK

Maozhen Li[1,2]

[2]The Key Laboratory of Embedded Systems and Service
Tongji University
Shanghai 200092, China

Abstract— **Music makes our life lovely because it can affect our mental states significantly with its emotional information inside. Different people might be affected differently from the same music when they listen the music in different situation and mental states. However, the common emotion information the music can be agreed even from peoples with quite different background and cultures. In this paper, we propose an automatic emotion recognition system for the music by extracting different features from the music and machine learning method learning from common knowledge on emotional state of the trained data. Firstly, two-channel audio signals are processed, and typical audio features are extracted. Then some other features used for EEG signal analysis are also extracted. Finally, these features are combined and the random forest classifier is used for the classification. The proposed method has been tested on a public music dataset and the experimental results demonstrate its efficiency in comparison with the state-of-the-art performance in the same dataset.**

Keywords-component; Musical Emotion Recognitio (MER); EEG; Random Forest; Music

I. INTRODUCTION

Music as one of the most classic expression being humans invention, it appeared in many artworks, such as songs, movies and theater. It can be seen as another language, used to express the author's thoughts and feelings. In many cases, music can extract meaning and emotion emerged where it works to express, which is the author's hope and the audience felt. At the same time, music as a sound, we want to express through sound different meanings not only by changing the sound characteristics, such as frequency, amplitude, so that the audience have different feelings. In the famous TV show 'I'm a singer', a singer who can only by the rhythm of the song adapted, with the original tone to express different emotions. Most of the audience will be affected with such sentiments, but they are not as accurate expression of this change as professional judges. So a musical emotion recognition (MER) system will be helpful in the society. A piece of music with clear emotion information inside will be beneficial for the entire entertainment industry and even the arts because the artist can use this information for their design and performance.

A typical MER system is to extract audio features from the music, and then these features are used for classification based on machine learning methods. The system can determine what kind of emotion this music belongs to.

In this paper, we carefully select the typical audio features extracted from both channels of the music recordings and then add other features used for EEG signal analysis. In addition, random forest method is used to combine the features in an efficient way to boost the overall performance.

The rest of the paper is organized as the following. In the section 2, related works are reviewed. In section 3, the proposed MER system is introduced. In section 4, the detailed experimental results are given and section 5 summarize the work with conclusions.

II. RELATED WORK

In recent years, there is a strong trend on emotion recognition from music due to the requirement of entertaining and digital media industry. Good datasets have been produced and lots of experiments have been done by many researchers all over the world. .

In 2008, Yang et al [1] used the Daubechies wavelets coefficient histogram (DWCH) algorithm [10], spectral contrast algorithm with PsySoud [5] and Marsyas [6] extracting the features as pervious work. Then he trained the data with features by multiple linear regression (MLR) [7], support vector regression (SVR) [8], and daBoost.RT (BoostR) [9]. Later, he used the MEVD (music emotion variation detection) [2]-[4] to do the same ground true data and features for fear comparison. At last he got his result that the regression approach has more promising prediction operation than normal arousal-valence (AV) [2]-[4], [11] computation algorithms in doing MER. And also the regression approach can be applied in MEVD.

In 2009, Han et al [12] used Juslin's theory[13] along with Thayer's emotion model[14] to analysis the 11 kind of emotions (angry, bored, calm, excited, happy, nervous, peaceful, pleased, relaxed, sad and sleepy) and their cause: 7 music characteristicals (pitch, tempo, loudness, tonality, key, rhythm and harmonics) an first. Then he compared his result, by using support vector regression (SVR) [15] as a classifier to train two regression functions for predicting arousal and valence values based on the low-level features, with the GMM (Gaussian Mixture Model[16]) and SVM (Support Vector Machine[17]). The final results showed that, the SVR can increase the accuracy from 63.03% to 94.55%, but the GMM can only grow 1.2% (91.52% to 92.73%).

In 2010, Kim et al [18] has done a comprehensive work in music emotion recognition review. He started with the psychology research on emotion with the Valence-Arousal space and the perceptual considerations. In feature part he described the lyrics feature selection and the acoustic features.

Lyrics features selection was based on the pleasure (valence), arousal and dominance (PAD) [19], Affective Norms for English Words (ANEW) [20] and Affective Norms for Chinese Words (ANCW) [21] to select the affective features from the signals. For the acoustic features, he gave 5 types (Dynamics Timbre, Harmony, Register, Rhythm and Articulation) 17 features (RMS energy, Mel-Frequency cepstral coefficients [22], spectral shape, spectral contrast，Roughness, harmonic change, key clarity, majorness，Chromagram, chroma centroid and deviation，Rhythm strength, regularity, tempo, beat，histograms，Event density, attack slope, attack time). But he focused the MFCCs (Mel-Frequency cepstral coefficients) to comment. And in machine learning part, he mentioned the SVM, Logistic Regression, Random Forest, GMM, K-NN, Decision Trees and Naive Bayes Multinomial classifiers. In the end of his paper, he gave example of some combinations of the emotion types. For example, Yang and Lee [23] combined the both audio and lyrics for emotion classification with 12 low-level MPEG-7 descriptors and increase 2.1% (82.8% vs 80.7% with only audio) of the results. And other review was for Bischoff [24] combining audio and tags (use Music-IR tasks as classification to make the tags like happy and sad [25]) by Bischoff with 240 dimensional features and 178 mood tags. Then Bischoff prove that the combination is better than the single one.

In 2013, 'The MediaEval 2013 Brave New Task: Emotion in Music' [26] attracted a large amount of challenger to take part in the task. In this task, they split the 1000 songs as 700 for development and 300 for test. And the first task was for the continuously time to determine the emotional dimensions, arousal and valence, and the automatically detection for the arousal and valence of the songs were the second task in static emotion characterization. Later on, this challenge was held in 2014 and 2015.

In MediaEval 2013, Konstantin Markov [27] extracted standard features tailored for music processing such as MFCC, Statistical Spectrum Descriptors (SSD), Chroma, Spectral Crest Factor (SCF), and Spectral Flatness Measure (SFM). Then he used the Gaussian Processes regression to get the results as the GPR got the better results for static emotion estimation when it compared with the SVR. Anna Aljanaki [28], has done the data filtering before which delete some useless information, such as containing speech, noise and environmental sounds. Then he used 3 toolboxes (PSYsound [29], VAMP [30] and MIRToolbox [31]) to extract the 44 features as loudness, mode, rms and mfcc 1-13 etc. Next step, he finished the feature selection in both arousal and valence with RReliefF feature selection algorithm in Weka in order to top 10 significant features [1]. At last he used the classifier multiple regression, Support Vector Regression, M5Rules, Multilayer Perceptron and other regressive techniques available in Weka to do the machine learning. Then got the result shows that the multiple regression performs as good as more sophisticated models – M5Rules. Because the high degree of correlation between valence and arousal, the prediction accuracy of valence is lower than that for arousal

MediaEval 2014 was similar with MediaEval 2013. One of the task was "dynamic emotion characterization". But there is a new task for second task: feature design, which asks the researchers find a new feature never being found or apply to MER. They prepared 744 songs for training and 1000 songs for evaluation. All the music come from large numbers of tags, and belongs to different genres.

Naveen Kumar [44] has done a nice work in the feature design part of this task. His work designed 3 new features for the task: Compressibility feature (comp), Median spectral band energy (MSBE) and Spectral centre of mass (SCOM). Next step was frame level prediction, what was that he used openSMILE [35] to extract the features at an interval of 0.5s for directly predicting continuous arousal and valence ratings for each song. Then he used the separate linear regression models to train for the A-V over frame. In another prediction way, dynamic emotion ratings, he use features extracted over the entire 45 second clip of songs to predict the dynamic ratings. The he predicted the features by Partial Least Squares Regression (PLSR) with 64 Haar coefficient (which compute by Haar transform). In his conclusion, he mentioned that local characteristics and global compressibility feature can both decided the dynamic emotion ratings in a song. And the features would improve the prediction by take in to account from adjacent frames.

The MediaEval task in 2015 [43] only had one task for people -- dynamic emotion characterization. But for the dataset, they do some changes. They connected come royalty-free music from several sources. Then set the development part with 431 clips of 45 second and 50 complete music pieces around 300 second for test.

Thomas Pellegrini [45] chose the recurrent neural networks (RNN) [47] to predict the A-V for the system for their sequence modeling capabilities. He competed the 260 baseline features which are given by [43] with 29 acoustic feature types which extract by ESSENTIA toolbox [46]. After 10-fold cross-validation (CV) setup, he got the result that valence and arousal values is correlated as r=0.626 (r: Pearson's correlation coeffcients).

In 2014, Gao et al. [33] used the 6552 standard large emotion feature which extract from OpenSMILE to compare with his multi-feature (GMM super vector – GSV [37], positive constranin matching pursuit—PCMP [38], multiple candidates and optimal path) with the APM dataset. Then he do the machine learning by SVM classifiers with RBF kernel and got the confusion matrix of the results. For the results, his working proved that his multiple features are more effective than only use the openSMILE one by 5.25% (74.9% to 80.15%)

In our work, we tried to build another MER system based on Gao's system. The main differences were that: 1).We extracted the feature from two channels; 2). We extracted extra features based on EEG analysis methods; 3). We used Random Forest to combine these features together in the classification efficiently.

III. MUSIC EMOTION RECONGNITION SYSTEM

Based on preliminary analysis of the data, we focus on two channels of data for feature selection.. We will try to build a better system for the musical emotion recognition with better features and better classifiers.

A. System Overview

The overall system is shown in Figure 1. Both channels of the music signals are processed separately and typical audio features are extracted from them. In addition, some statistic features used for EEG analysis are also extracted. Then these features are combined together as the overall features for classification. Finally random forest classifier is chosen to ensemble all these features together and predict the emotion categories of the music.

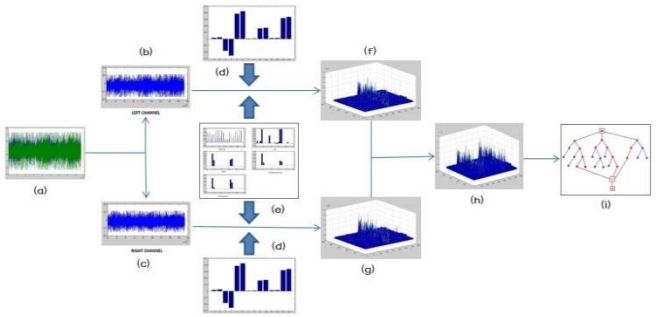

Figure 1 The overview of the music emotion recognition system. (a). original signal, (b). left channel, (c). right channel, (d). EEG statistic features, (e). typical audio feature, (f). left channel combination feature, (g). right channel combination feature, (h). two channel combination feature, (i). random forest classifier)

B. Typical Audio Feature Extraction

There are many features existed for audio signal analysis. Here some of the typical emotion related features are selected based on 'The INTERSPEECH 2009 Emotion Challenge' [34], in which these features can be very accurately reflect the most emotional characteristics of the audio signal. In Table I, the features and the functionals are shown. These features include 16 low-level-descriptors (LLD) and 12 functionals. Totally, 16*2*12=384 features are selected.

TABLE I. THE FAN'S RERULTS WITH APM DATABASE

LLD	Functionals
PCM_RMSenergy	Max, min
MFCC(1-12)	Range, max position, min position
PCM_ZCR	Arithmetic mean
Voiceprob	The slope (m) of a linear approximation, the offset (t) of a linear approximation
F0	The quadratic error computed as the difference of the linear approximation, the standard deviation of the values
	Skewness, kurtosis

1) RMSenergy (Root-mean-square signal frame energy)

RMS ENERGY [32] is based on the amplitude of the peak value of the audio signal to calculate the power of a signal, which shows the energy carried by the signal. It is a common audio feature representation.

2) MFCCs (Mel-Frequency cepstral coefficients)

Mel frequency is a major concern for the human hearing characteristic frequency. Because humans often unconsciously for a certain period of frequency is very sensitive, if the analysis of such frequency will greatly improve accuracy. MFCC is a linear mapping of the audio spectrum to Mel spectrum, then the conversion factor cepstral frequency domain when available.

3) ZCR (Zero-crossing rate of time signal)

ZCR means the signal from one end of the axis 0 of the cross to the other side of the ratio. It is usually expressed by a signal of a predetermined value of 0, or as a filter are searching for a rapidly changing signals.

4) F0(The fundamental frequency computed from the Cepstrum)

The fundamental frequency, in a complex wave is a periodic waveform the lowest sum of frequency of. In music, it is usually the lowest pitch notes.

5) Voiceprob(The voicing probability computed from the ACF)

Figure 2 in the following show the examples of the five typical audio features based on one music signal.

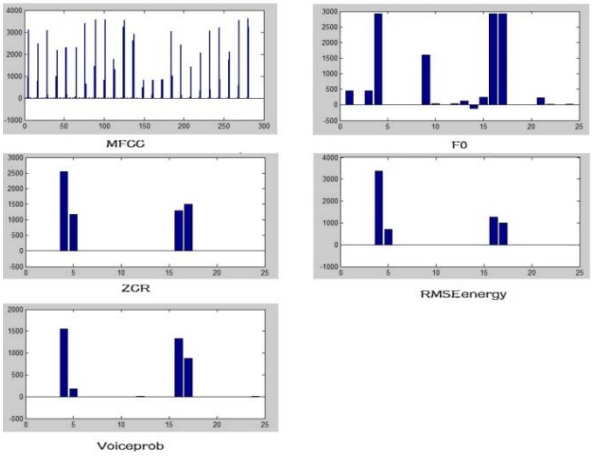

Figure 2. Five typical audio features

C. Feature Extraction based on EEG Analysis

In the emotion recognition feature extraction on the audio, the paper according to the EEG feature [36][40][41] extraction method to extract the brain wave data and audio features closest to several groups. After testing found that this set of features greatly enhance the musical emotion recognition accuracy.

Since the EEG signal and an audio signal differences, these characteristics as the basic statistical characteristics of the signal in the time domain.

- Power

$$P_\varepsilon = \frac{1}{T}\sum_{-\infty}^{\infty}|\varepsilon(t)|^2 \quad (1)$$

- Mean

$$\mu_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}\varepsilon(t) \quad (2)$$

- Standard deviation

$$\sigma_\varepsilon = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\varepsilon(t)-\mu_\varepsilon)^2} \quad (3)$$

- First difference

$$\delta_\varepsilon = \frac{1}{T-1}\sum_{t=1}^{T-1}|\varepsilon(t+1) - \varepsilon(t)| \quad (4)$$

- Normalized 1st difference

$$\bar{\delta} = \frac{\delta_\varepsilon}{\sigma_\varepsilon} \quad (5)$$

- Second difference

$$\gamma_\varepsilon = \frac{1}{T-2}\sum_{t=1}^{T-2}|\varepsilon(t+2) - \varepsilon(t)| \quad (6)$$

- Normalized 2nd difference

$$\bar{\gamma} = \frac{\gamma_\varepsilon}{\sigma_\varepsilon} \quad (7)$$

Figure 3 in the following show the examples of the features. In the figure, each bar means one feature in EEG stat features.

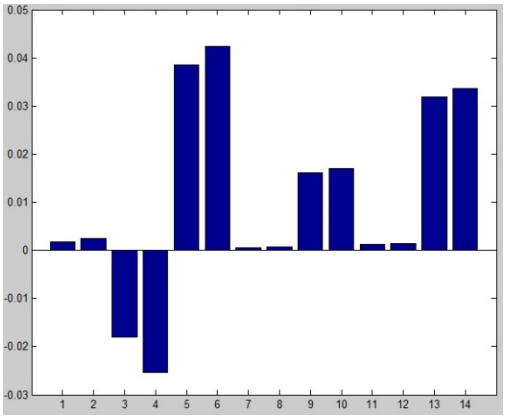

Figure 3 EEG analysis feature

D. Random Forest Classifacation

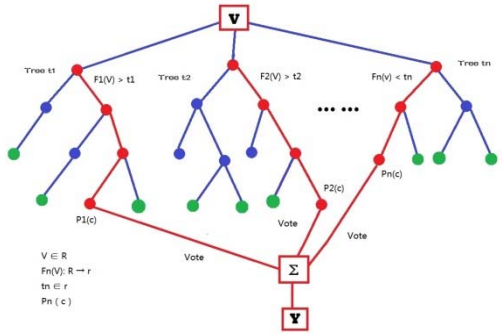

Figure 4 Random Forest classifier in which several decision trees are selected and combined together for the classification automatically

Random forests as a mainstream machine learning, it can handle high-dimensional data, and can be used to assess the importance of each category according to the final results of forecast accuracy. Most importantly, he can estimate the missing part of the sample, and provides high accuracy results. This approach combines machine learning 'baging' ideas and features. It is possible to sample the original sample and the formation of a new training set. Then they were randomly training for different tree. After a few times of packet classification training, get the final result by voting the decision tree.

The figure 4 in front is showing the basic concept of random forest. R are the features, V are the parts of features, Fn (V) are the class rule. The next step of the branch depend on if the feature is satisfy the requirement of the rule. After many of the trees go to final branch by the classifier, they will voting for which one is the most, which is the final result of the classification.

IV. Experemental Risults

The following experiments are to evaluate our proposed system for emotion recognition from music recordings. APM database[39] was used this experiment. In this study, firstly Matlab extracted one-dimensional audio signal, the signal and two channels separately. Use OpenSMILE extracted five characteristics of the signal, RMSennergy, MFCC, ZCR, F0 and Voice Prob. Then, using EEG feature extraction, feature extracted Stat. It features a total of 12 set last two channels together. Random forest classifier using machine learning, and get the final accuracy of the confusion matrix correct rate.

The experimental results will be compared with the results in Gao's work because the same APM dataset was used.

A. APM Database

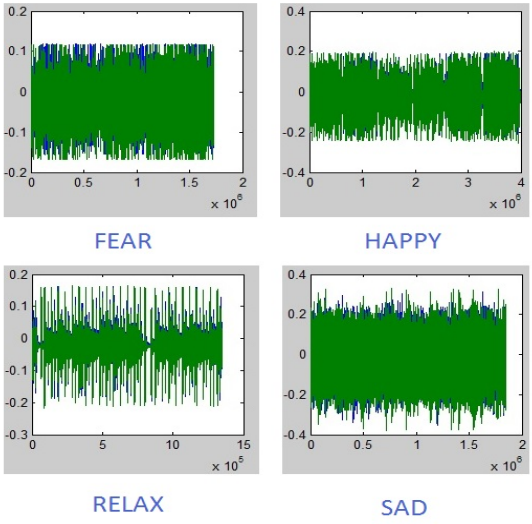

Figure 5 Four audio signals for 4 different emotions

In order to make sure the effect of the features during the real situation, an experiment based on APM music emotion recognition data set be started. APM music data set is a music data in wav format. The data set including 4 kinds of emotion: happy, sad, relax and fear. Each of the emotion has 100 audio samples with around 35 seconds long. I need to set the testing and training set by myself or I prefer to use cross validation to test the features and the arithmetic more accuracy and more effective.

APM is the largest production music library in the industry. It contains more than 40 libraries, more than 475,000 tracks and CDs. And he almost contains almost every type and style of music. Therefore, this article chose him as this experiment database.

Figure 5 shows some examples of the music data with different emotions.

B. Experimental Setting

In this experiment, I used Matlab software to separate the audio signal input in two channels. The system uses software to extract music emotion OpenSMILE part features. The two parts of the audio signals are introduced into OpenSMILE software, and then select the IS09 as the configuration, and set the export .CSV file. The second step, the exported file split into two parts, and calculated using the Weka machine learning and accurate rates.

In Weka, we selected RandomForest as a machine learning classifier. Decision trees are set to 800 and the numFeature(the number of attributes to be used in random selection) is changed from default 1 to 200. Then, we set 5 times 2-fold cross-validation same as Gao's work. Finally, the prediction results obtained with the actual results were combined with confusion matrix accuracy of analysis to get the overall accuracy.

C. Performance measurement

The performance is measured based on the classification rate of the prediction on the testing dataset. The accuracy is defined by the following equation (true positive (TP), true negative (TN), accuracy (A)):

$$A = \frac{TP+TN}{P+N} \quad (8)$$

D. Results and Comparison

In Gao's work on APM database, OpenSMILE was used to extract 6552 emotional features with an accurate rate of 74.9%. The detailed results is shown in Table II in the following. Further, PCMP feature [38] was used in combining with OpenSMILE feature and the accuracy is improved to 77%. In addition, he used multiple techniques (GSV+PCMP+multiple candidates+optimal path) to get the best accuracy rate of 80.15%.

TABLE II. COMPARISON BETWEEN THE GAO'S RERULTS WITH APM DATABASE AND THEFAN'S RERULTS WITH APM DATABASE

Method	Feature	Accuracy (%)
SVM classifiers with RBF kernel [33]	OpenSMILE	74.9
	OpenSMILE+PCMP	77.4
	GSV+PCMP+multiple candidates+optimal path	80.15
Random Forest (proposed method)	OpenSMILE (left channel)	76.06
	OpenSMILE (right channel)	78.3
	OpenSMILE (2 channels)	77.81
	EEG Stat	69.08
	OpenSMILE (2 channels)+EEG	**83.29**

Our experimental results are shown in the Table II. In our work, only a few of these audio features were selected and the accuracy was 76.06%. Two channels have got the same performance although the confusion matrix are different. The combined features from two channels makes the accuracy of 77.81%. Although the EEG features did not perform well with 66.8% only, the combined performance is 83.29% that is better than all the results on this dataset. The associated confusion matrix is shown in Table III.

In comparison with Gao's results in Table II, Our selected feature achieved better performance. Although features from EEG analysis are not very good, its compensation in feature space to the audio feature has made contributions on the performance improvement. The result of the accuracy increases by 3.14%. The overall performance of the proposed emotion recognition system achieved the best results.

From the confusion matrix in Table III, it can be seen that "Happy" is the easiest emotion to be recognized. "Fear" and "Sad" are difficult because they are very similar.

TABLE III. THE COMFUTION MATRIX OF PROPOSED SYSTEM(ACTUAL CLASSES IN COLUMNS,PREDICTED CLASSES IN ROWS)

	Fear	Happy	Relax	Sad
Fear	82	0	8	11
Happy	1	95	4	0
Relax	3	2	84	11
Sad	13	1	13	73

V. CONCLUSION AND DISCUSSION

In this paper, an automatically musical emotion recognition system was proposed. As it can be seen from Table 1, the same use OpenSMILE feature extraction, the feature selection one exceed non-selection feature by 2%. After the combination with EEG based features, and the Random Forest produce the best results over Gao performance ratio higher 0.4%. This shows that this way of feature selection and machine learning can indeed improve the accuracy of musical emotion recognition.

In the future work, more dataset will be used for the testing of the proposed system. More other features should also be considered.

ACKNOWLEDGEMENT

This research is partially supported by the 973 project on Network Big Data Analytics funded by the Ministry of Science and Technology, China. No. 2014CB340404.

REFERENCES

[1] Y.H. Yang, et al.: "A regression approach to music emotion recognition," IEEE Trans. on ASLP, Vol. 16 (2), pp. 448–457, 2008.

[2] Y.-H. Yang, C.-C Liu, and H. H. Chen, "Music emotion classification: A fuzzy approach," Proc. ACM Multimedia, Santa Barbara, USA, pp. 81–84, 2006.

[3] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," IEEE Trans. Multimedia, vol. 7, no. 1, pp. 143–154, 2005.

[4] E. Schubert, "Measurement and time series analysis of emotion in music," Ph.D. dissertation, School of Music & Music Education, Univ. New South Wales, Sydney, Australia, 1999..

[5] D. Cabrera, "PSYSOUND: A computer program for psychoacoustical analysis," Proc. Australian Acoustic Society Conf., pp. 47–54, 1999.

[6] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Trans. Speech and Audio Processing, vol. 10, no.5, pp. 293–302, 2002.

[7] A. Sen and M. Srivastava, Regression Analysis: Theory, Methods, and Applications, New York, Springer, 1990..

[8] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” Statistics and Computing, 2004.

[9] D.P. Solomatine and D.L. Shrestha, “AdaBoost.RT: A boosting algorithm for regression problems,” Proc. IEEE Int. Joint Conf. Neural Networks, pp. 1163–1168, 2004.

[10] T. Li and M. Ogihara, “Content-based music similarity search and emotion detection,” Proc. Int. Conf. Acoustic, Speech, and Signal Processing, Toulouse, France, pp. 17–21, 2006.

[11] M. D. Korhonen, D. A.Clausi, and M. E. Jernigan, “Modeling emotional content of music using system identification,” IEEE Trans. Systems, Man., and Cybernetics, vol. 36, no. 3, pp. 588–599, 2006.

[12] Han, B.J., Dannenberg, R.B., Hwang, “SMERS: music emotion recognition using support vector regression”. In: Proc. ISMIR. pp. 651–656 (2009).

[13] P.N. Juslin and J.A. Sloboda: “Music and Emotion: Theory and research,” Oxford Univ. Press, 2001.

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

[15] Smola, Alex J., et al.: “A tutorial on support vector regression,” Statistics and Computing, Vol.14, pp.199-222, 2004.

[16] Ramesh Sridharan, “Gaussian mixture models and the EM algorithm”, Avilable in: http://people.csail.mit.edu/rameshvs/content/gmm-em.pdf

[17] CC Chang, CJ Lin,” LIBSVM: a library for support vector machines”, ACM Transactions on Intelligent Systems and Technology, Volume 2, issue 3, 2011.

[18] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, “Music emotion recognition: A state of the art review,” in Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), pp. 255–266, Utrecht, Netherlands, 2010.

[19] A. Mehrabian and J. A. Russell, An Approach to Environmental Psychology. MIT Press, 1974.

[20] M. M. Bradley and P. J. Lang, “Affective norms for English words (ANEW),” The NIMH Center for the Study of Emotion and Attention,University of Florida, Tech. Rep., 1999.

[21] Y. Hu, X. Chen, and D. Yang, “Lyric-based song emotion detection with affective lexicon and fuzzy clustering method,” in Proc. of the Intl. Society for Music Information Conf., Kobe, Japan, 2009.

[22] L. Mion and G. D. Poli, “Score-independent audio features for description of music expression,” IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 2, pp. 458–466, 2008.

[23] D. Yang and W. Lee, “Disambiguating music emotion using software agents,” in Proc. of the Intl. Conf. on Music Information Retrieval. Barcelona, Spain: Universitat Pompeu Fabra, October 2004.

[24] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, “Music mood and theme classification-a hybrid approach,” in Proc. of the Intl. Society for Music Information Retrieval Conf., Kobe, Japan, 2009.

[25] P. Knees, E. Pampalk, and G. Widmer, “Artist classification with web-based data,” in Proc. of the Intl. Symposium on Music Information Retrieval, Barcelona, Spain, 2004, pp. 517– 524.

[26] M. Soleymani.“The MediaEval 2013 Brave New Task: Emotion in Music“, epartment of Computing. Imperial College London, UK, 2013

[27] K.Markov, M.Iwata and T.Matsui. :”Music emotion recognition using gaussian preocesses”. In: Processdings of the ACM Multimedia 2013 Worksop on Crowdsourching for Multimedia, CrowdMM, ACM, Barcelona, Spain 2013.

[28] A Aljanaki, F.Wiering, and R.C. Veltkamp,“MIRUtrecht participation in MediaEval 2013: emotion in music task,” in MediaEval Workshop, Vol. 1043，Barcelona，2013.

[29] Lartillot, O., Toiviainen, P., 2007. A Matlab Toolbox for Musical Feature Extraction FromAudio, International Conference on Digital Audio Effects, Bordeaux, 2007.

[30] Cabrera, D., 1999. PSYSOUND: A computer program for psychoacoustical analysis, in Proc. Australian Acoust. Soc. Conf., 1999, pp. 47–54.

[31] Sonic Annotator. http://www.omras2.org/SonicAnnotator

[32] K Sakhnov, E Verteletskaya, B Simak，“Approach for energy-based voice detector with adaptive scaling factor”，IAENG International Journal of Computer Science 36 (4), 394，2009

[33] B. Gao,"Contributions to music semantic analysis and its acceleration techniques",PhD dissertation, Ecole Centrale de Lyon, 2014

[34] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” Speech Communication, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[35] F. Eyben, M. Woellmer, and B. Schuller,"OpenSMILE, the Munich open Speech and Music Interpretation by Large Space Extraction toolkit", Institute for Human-Machine Communication Technische Universitaet Muenchen (TUM) D-80333 Munich, Germany ,2010

[36] Jenke, R., Peer, A. and Buss, M. Feature extraction and selection for emotion recognition from EEG. IEEE: Transactions on Affective Computing, 5 (3). pp. 327-339. ISSN 1949-3045, 2014

[37] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. Signal Processing Letters, IEEE, 2006, 13(5): 308-311.

[38] B. Gao, E. Dellandréa, and L. Chen, “Music sparse decomposition onto a midi dictionary of musical words and its application to music mood classification,” in Proceedings of 10th IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6, 2012.

[39] APM Music, http://www.apmmusic.com/about-apm-music-and-vision.

[40] K. Takahashi, “Remarks on emotion recognition from multimodal bio-potential signals,” in Proc. Int. Conf. Ind. Technol., 2004, pp. 1138–1143.

[41] X. Wang, D. Nie, and B. Lu, “EEG-based emotion recognition using frequency domain features and support vector machines,” in Proc. Int. Conf. Neural Inf. Process., 2011, pp. 734–743.

[42] A. Aljanaki, Y. Yang, and M. Soleymani. “Emotion in music task at mediaeval 2014”. In Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17, 2014.

[43] A. Aljanaki, Y.-H. Yang, and M. Soleymani. “Emotion in music task at mediaeval 2015”. In Working Notes Proceedings of the MediaEval 2015 Workshop, September 2015.

[44] N. Kumar, R. Gupta, T. Guha, C. Vaz, M. Van Segbroeck, J. Kim, and S. Narayanan. “Affective feature design and predicting continuous affective dimensions from music". In MediaEval Workshop, Barcelona, 2014.

[45] T Pellegrini, V.Barrière. “Time-continuous Estimation of Emotion in Music with Recurrent Neural Networks”. In MediaEval Workshop, Barcelona, 2014.

[46] D. Bogdanov, N. Wack, E. Gomez, S. Gulati, P. Herrera, O. Mayor, and et al. “ESSENTIA: an Audio Analysis Library for Music Information Retrieval”. In Proc. International Society for Music Information Retrieval Conference (ISMIR'13), pages 493{498, Curitiba, 2013.

[47] E. Coutinho, F. Weninger, B. Schuller, and K. Scherer. “The Munich LSTM-RNN”, Approach to the MediaEval 2014, aAIJEmotion in MusicaAI Task. In Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, 2014.