Music Emotion Recognition Using Deep Gaussian Process

Sih-Huei Chen, Yuan-Shan Lee, Wen-Chi Hsieh, and Jia-Ching Wang
Department of Computer Science and Information Engineering
National Central University, Taiwan, R.O.C.

***Abstract*— Music is a powerful force that evokes human emotions. Several investigations of music emotion recognition (MER) have been conducted in recent years. This paper proposes a system for detecting emotion in music that is based on a deep Gaussian process (GP). The system consists of two parts-feature extraction and classification. In the feature extraction part, five types of features that are associated with emotions are selected for representing the music signal; these are rhythm, dynamics, timbre, pitch and tonality. A music clip is decomposed into frames and these features are extracted from each frame. Next, statistical values, such as mean and standard deviation, of frame-based features are calculated to generate a 38-dimensional feature vector. In the classification part, a deep GP is utilized for emotion recognition. We treat classification problem from the perspective of regression. Finally, 9 classes of emotion are categorized by 9 one-versus-all classifiers. The experimental results demonstrate that the proposed system performs well in emotion recognition.**

***Index Terms*—Music emotion recognition, deep Gaussian process, classification, feature extraction.**

I. INTRODUCTION

Music is usually regarded as a way to release emotions. When people are feeling down, listening to happy music may make them feel full of energy. When a person is studying but bored, exciting music may help to increase efficiency. In these cases, knowing which music can make people feel such emotion may bring more convenience to them. The research for detecting the emotion of songs is called music emotion recognition (MER). According to psychological theory, music emotion recognition system can be roughly divided into parametric models and categorical models.

Parametric approaches represent emotion in music using the valence and arousal (VA) values [1, 3, 8]. Figure 1 displays the VA plane and the four main classes of human emotion- angry, happy, sad and relaxed. The emotion of a music chip can be described as a point in VA plane. However, the emotion tags in the VA plane may be inconvenient for users to quickly search songs. Another way to represent emotions in music is categorical approach that tags songs with emotion labels or adjectives, such as angry, bored, calm, happy, and peaceful. In this work, the categorical approach is utilized to detect emotions in music.

Music emotion recognition involves assigning emotion labels to musical content. Numerous investigations [2, 4, 14] have addressed the task of categorical music emotion recognition. The support vector machine (SVM) [7] is the state-of-the-art classifier for music emotion recognition that finds the relationship between the musical content and emotion labels. Recently, the Gaussian process (GP) has been used as a powerful probabilistic framework for solving regression and classification problems with complex data. A GP can be specified by the mean vector and covariance matrix. The mean vector is commonly assumed to be zero. The covariance matrix, which is obtained from a kernel function, can expresses the relationship among data points. Comparing with SVM, GP provides probability estimates for the predicted data points. Many researchers have utilized GP in regression or classification task on several fields [12, 17]. In the music information recognition (MIR) field, Jensen *et al.* used GP with a specific covariance matrix for a music recommendation system [18]. Markov and Matsui [5] proposed a system based on GP regression to detect the emotion in the VA plane. Recently, the field of deep learning has been developed. Deep hierarchies are constructed by stacking several models. Schmidt and Kim employed the deep belief network (DBN) to learn the sparse feature for music [13]. Considering the advantage of GP and deep learning, Damianou and Lawrence [6] have shown that GP can be used in deep hierarchical structure by stacking them. Deep Gaussian process provides structural learning in Gaussian process model.

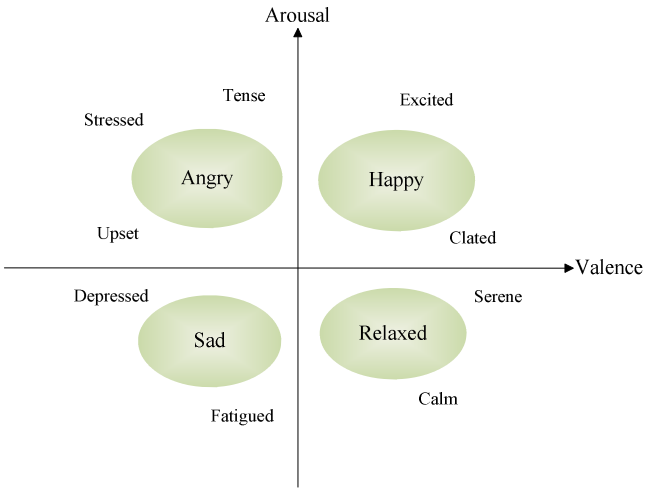

Fig. 1. Thayer's 2-D arousal-valence plane.

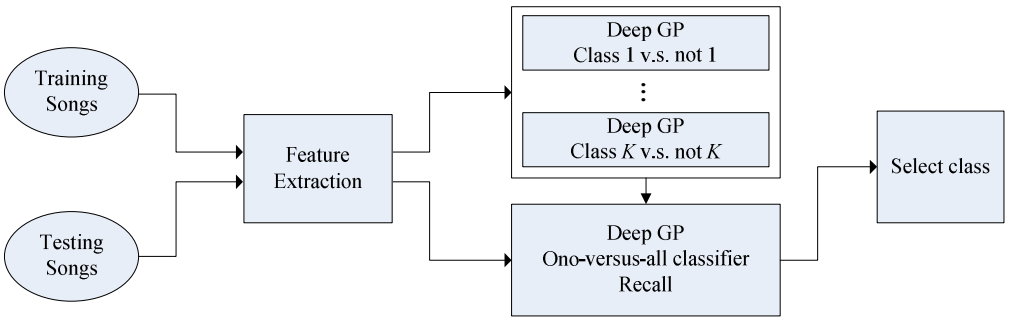

Fig. 2. System architecture.

This paper presents a system for recognizing emotion in music that is based on a deep Gaussian process. The potential and efficiency of learning methodologies with deep architecture have been proved in statistical machine learning. Moreover, the Gaussian process can provide the uncertainty of predictions because of probabilistic properties. Therefore, the deep Gaussian process is a powerful approach to capture the relationship among nonlinear data. In this study, deep GP is utilized for music emotion recognition.

The rest of this paper is composed of the following sections: Section II presents the overview of the proposed system for recognizing emotion in music. Section III presents details of the feature extraction stage. Section IV describes MER based on the standard Gaussian process and the deep Gaussian process. Section V, experimentally elucidates the performance of the proposed system. Finally, the conclusion of this work is written in Section VI.

II. System Overview

This paper constructs a system for recognizing emotion in music that is based on a deep Gaussian process. The procedure of proposed system is shown in Figure 2. The proposed system comprises two parts. In the feature extraction part, 15 acoustical features that are often used in music emotion research are extracted from each music clip. The next section will describe feature extraction in details. In the classification part, a deep Gaussian process is used for recognizing emotion. For $K > 2$ classes, the use of K one-versus-all classifiers, each of which solves a two-class problem, is considered.

TABLE I
EXTRACTED FEATURES

Feature Type	Feature Name (Dimension)
Rhythm	Beatspectrum (1), Event density (1), Tempo (1), Pulse clarity (1)
Dynamics	RMS energy (1), Low energy (1)
Timbre	Zero-crossing rate (1), Roll-off (1), Brightness (1), Roughness (1), MFCC (13)
Pitch	Pitch (1), Harmonicity (1)
Tonality	Chromagram (12), Mode (1)

III. Feature Extraction

The various characteristics of a song (such as rhythm, timbre, tone and others) can induce different emotional responses in humans. A computer does not have a wealth of perception like a human brain. The use of a computer to recognize emotions in a song requires that first, the features of the music are extracted.

Based on relevant theory, this paper extracts 15 acoustical features, which are associated with five types of feature, which are rhythm, dynamics, timbre, pitch and tonality. Table I lists these features. MIR toolbox [9] is used to extract the acoustical features from each music clip.

IV. MER Based on Deep Gaussian Process

A. Gaussian Process

Considering a database of K emotions, each song of interest is divided into several clips. For clip n, the feature obtained from the previous section $\mathbf{x}_n \in R^D$ is regarded as an observation with an emotion label $\mathbf{y}_{n\bullet} \in R^K$ using 1-of-K coding. The classification problem is solved through the perspective of regression. Consider a set of training data comprising N observations $\{\mathbf{x}_n, \mathbf{y}_{n\bullet}\}_{n=1}^N$. In traditional probability nonparametric regression, the model can be assumed by

$$y_{nd} = f_{nd} + \varepsilon_{nd} \tag{1}$$

where $f_{nd} = f_d(\mathbf{x}_n)$, $d = 1,\ldots,K$ and $\varepsilon_{nd} \sim N(0,\sigma_d^2)$ is i.i.d. Gaussian noise. Let $\mathbf{y}_{\bullet d} = [y_{1d}, y_{2d},\ldots, y_{Nd}]^T \in R^N$ be the emotion label that are obtained from the output of function $f_d(\mathbf{X})$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_N]^T$ by adding Gaussian noise with zero mean and covariance matrix $\sigma_d^2\mathbf{I}$. The likelihood of target value $\mathbf{y}_{\bullet d}$ conditioned on function value $f_d(\mathbf{X})$ is given by

$$p(\mathbf{y}_{\bullet d}|\mathbf{f}_d) = N(\mathbf{y}_{\bullet d}|\mathbf{f}_d, \sigma_d^2\mathbf{I}) \tag{2}$$

where $\mathbf{f}_d = f_d(\mathbf{X})$. The latent function f_d is obtained from a Gaussian process $GP(\mathbf{0},\mathbf{K})$ [10], i.e. $p(\mathbf{f}_d|\mathbf{X}) = N(\mathbf{f}_d|\mathbf{0},\mathbf{K})$ where $\mathbf{K}$ is the Gram matrix with elements $\mathbf{K}_{nm} = k(\mathbf{x}_n,\mathbf{x}_m)$ and $k(\mathbf{x}_n,\mathbf{x}_m)$ is a kernel function. Since the GP as prior is

independent, the GP priors and joint likelihood of the target values $\mathbf{Y}=[\mathbf{y}_{\bullet 1},\mathbf{y}_{\bullet 2},\ldots,\mathbf{y}_{\bullet K}]$ conditioned on the value of $\mathbf{F}=[\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_K]$ are given by

$$p(\mathbf{F}|\mathbf{X})=\prod_{d=1}^{K}N(\mathbf{f}_d|\mathbf{0},\mathbf{K}) \tag{3}$$

$$p(\mathbf{Y}|\mathbf{F})=\prod_{d=1}^{K}N(\mathbf{y}_{\bullet d}|\mathbf{f}_d,\sigma_d^2\mathbf{I}) \tag{4}$$

The marginal likelihood of $\mathbf{Y}$ can be computed by integrating over $\mathbf{F}$ as follows:

$$\begin{aligned}p(\mathbf{Y}|\mathbf{X})&=\int p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})d\mathbf{F}\\&=N(\mathbf{Y}|\mathbf{0},\mathbf{K}'+\Sigma)\end{aligned} \tag{5}$$

where $\mathbf{K}'$ is composed of $N\times N$ blocks and $\sum=\mathrm{diag}(\sigma_1^2\mathbf{I},\sigma_2^2\mathbf{I},\ldots,\sigma_K^2\mathbf{I})$. By the variational inference, the model parameters can be obtained [11].

B. Deep Gaussian Process

In this paper, the Deep GP [6] is applied for recognizing emotion in music. Deep Gaussian process is a deep neural network that can be modified by changing the mapping between latent layers. The graphical model of deep GP is shown in Figure 3. The clip-level feature and the emotional label of each data point are regarded as input $\mathbf{x}\in R^D$ and output $\mathbf{y}\in R^K$, respectively. A deep GP with L hidden layers is a composition of vector-valued functions drawn from a Gaussian process:

$$\mathbf{y}=\mathbf{f}^L(\mathbf{f}^{L-1}(\ldots\mathbf{f}^2(\mathbf{f}^1(\mathbf{x})))) \tag{6}$$

where $\mathbf{f}^l$ is drawn from $GP(\mathbf{0},\mathbf{K}^l)$.

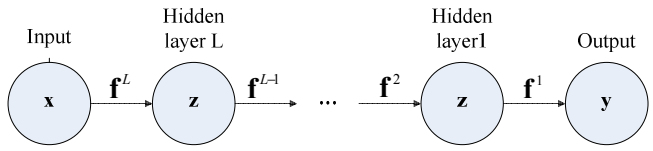

Fig. 3. Graphical model of deep GP.

The direct optimization of the log marginal likelihood $\log p(\mathbf{Y})$ is difficult. The variational inference yields an approximate solution that replaces the analytical solution. Consider a distribution q, the following decomposition holds.

$$\log p(\mathbf{Y})=L(q)+\mathrm{KL}(q\|p) \tag{7}$$

where

$$L(q)=\int_{\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X}} q\log\frac{p(\mathbf{Y},\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X})}{q} \tag{8}$$

$$\mathrm{KL}(q\|p)=-\int_{\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X}} q\log\frac{p(\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X}|\mathbf{Y})}{q} \tag{9}$$

Maximizing the log marginal likelihood $\log p(\mathbf{Y})$ is equivalent to maximizing the lower bound $L(q)$. Note that the joint likelihood in Eq. (8) can be written as follows:

$$p(\mathbf{Y},\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X})=p(\mathbf{Y}|\mathbf{Z}_1)\cdots p(\mathbf{Z}_1|\mathbf{X}) \tag{10}$$

However, the evaluation of integral in Eq. (8) is difficult because of the nonlinear function. Through the GP prior, Eq. (8) can be rewritten as follows:

$$L(q)=\int_{\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{f}^1,\ldots,\mathbf{f}^L,\mathbf{X}} q\log\frac{p(\mathbf{Y},\mathbf{f}^1,\ldots,\mathbf{f}^L,\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X})}{q} \tag{11}$$

where the joint likelihood can be expanded as follows:

$$\begin{aligned}&p(\mathbf{Y},\mathbf{f}^1,\ldots,\mathbf{f}^L,\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{X})\\&=p(\mathbf{Y}|\mathbf{f}^1)p(\mathbf{f}^1|\mathbf{Z}_1)\cdots p(\mathbf{Z}_L|\mathbf{f}^L)p(\mathbf{X})\end{aligned} \tag{12}$$

Equation (11) is still difficult to calculate. Titsias and Lawrence have shown that the problem can be solved by expanding the probability space [11]. Therefore, the joint likelihood in Eq. (11) can be rewritten as follows:

$$\begin{aligned}&p(\mathbf{Y},\mathbf{f}^1,\ldots,\mathbf{f}^L,\mathbf{Z}_1,\ldots,\mathbf{Z}_L,\mathbf{U}^Y,\mathbf{U}^1,\ldots,\mathbf{U}^L,\widetilde{\mathbf{X}},\widetilde{\mathbf{Z}}^1,\ldots\widetilde{\mathbf{Z}}^L,\mathbf{X})\\&=p(\mathbf{Y}|\mathbf{f}^1)p(\mathbf{f}^1|\mathbf{Z}_1,\mathbf{U}^Y)\cdots p(\mathbf{f}^L|\mathbf{X},\mathbf{U}^L)p(\mathbf{U}^L|\widetilde{\mathbf{Z}}^L)p(\mathbf{X})\end{aligned} \tag{13}$$

The lower bound can be maximized by the gradient-based optimization method [11].

V. EXPERIMENTS

A. Dataset Configuration

The database of emotional music in this work refers to 9 classes of emotion (anger, sadness, happiness, boredom, calm, relaxation, nervousness, pleased and peace). The database was constructed by collecting music clips from two websites, All Music Guide [15] and Last.fm [16]. Each class consists of 120 music clips. The sampling rate of each file was 16 kHz and the resolution was 16 bits per sample. The frame size was 256 samples with 50% overlap between adjacent samples.

B. Experimental Results

In the experiment, two-fold cross-validation is used to evaluate the performance of the proposed system. We compare our proposed method to SVM and standard GP. Table II presents the confusion matrix of the proposed system and the baseline. The radial basis function (RBF) is used as a kernel function in the experiment. The accuracy rate can be computed as a ratio between the number of correctly identified data and the number of testing data. In our proposed system (see Table II (A)), the number of hidden layers is 1. From the result of proposed system, we see that the classes-anger and calm are more difficult to be modeled by deep GP (see Table II (A) and (B)). Table II (C) presents the confusion matrix of standard GP. Comparing with deep GP, standard GP has a worse performance than deep GP in the classes- calm,

nervousness, peace and relaxation. Overall, the proposed system has the better performance than SVM and standard GP.

TABLE II
CONFUSION MATRIX (IN %) OF EXPERIMENTAL RESULTS USING DIFFERENT CLASSIFIERS (A) PROPOSED SYSTEM (B) SVM (C) GP

(A)

	ang	bor	cal	hap	ner	pea	ple	rel	sad
ang	**77.5**	3.30	1.70	13.3	0.80	0.00	2.50	0.80	0.00
bor	3.30	**82.5**	0.00	0.80	4.20	1.70	5.80	1.70	0.00
cal	1.70	0.00	**64.2**	2.50	0.00	19.2	0.00	1.70	10.8
hap	16.7	0.00	0.00	**77.5**	0.00	0.80	2.50	0.80	1.70
ner	0.00	6.70	0.80	0.80	**78.3**	0.80	3.30	5.80	3.30
pea	0.00	0.00	13.3	4.20	1.70	**58.3**	1.70	5.00	15.8
ple	0.00	3.30	0.00	1.70	3.30	0.80	**87.5**	2.50	0.80
rel	0.80	5.00	1.70	0.00	10.0	5.00	10.0	**66.7**	0.80
sad	1.70	1.70	15.0	9.20	2.50	11.7	6.70	2.50	**49.2**
Avg.	**71.3**								

(B)

	ang	bor	cal	hap	ner	pea	ple	rel	sad
ang	**90.8**	2.50	0.80	1.70	0.80	1.70	1.70	0.00	0.00
bor	13.3	**72.5**	6.70	0.80	1.70	1.70	0.80	0.00	2.50
cal	2.50	0.80	**94.2**	0.00	0.00	2.50	0.00	0.00	0.00
hap	43.3	1.70	0.80	**52.5**	0.00	0.00	0.00	0.00	1.70
ner	0.80	11.7	7.50	0.80	**74.2**	3.30	0.00	0.80	0.80
pea	0.00	0.80	62.5	0.80	0.80	**30.8**	0.80	0.00	3.30
ple	0.80	14.2	3.30	0.80	0.00	8.30	**70.0**	2.50	0.00
rel	5.00	6.70	5.80	0.00	0.80	17.5	5.00	**59.2**	0.00
sad	1.70	0.80	48.3	5.80	0.00	17.5	1.70	1.70	**22.5**
Avg.	**63.0**								

(C)

	ang	bor	cal	hap	ner	pea	ple	rel	sad
ang	**80.8**	4.20	0.80	13.3	0.80	0.00	0.00	0.00	0.00
bor	1.70	**83.3**	0.00	0.80	5.80	0.80	0.80	6.70	0.00
cal	3.30	0.00	**46.7**	3.30	0.00	25.0	0.00	4.20	17.5
hap	14.2	0.00	0.00	**84.2**	0.00	0.80	0.80	0.00	0.00
ner	4.20	5.00	1.70	0.00	**68.3**	3.30	0.80	13.3	3.30
pea	5.80	0.00	22.5	4.20	0.00	**35.8**	3.30	4.20	24.2
ple	0.00	3.30	0.00	0.00	2.50	0.80	**90.8**	2.50	0.00
rel	4.20	4.20	1.70	0.00	10.8	3.30	10.8	**63.3**	1.70
sad	3.30	0.80	10.0	13.3	0.80	11.7	4.20	2.50	**53.3**
Avg.	**67.4**								

VI. CONCLUSIONS

This paper presents a music emotion recognition system that is based on a deep Gaussian process. The proposed system consists of two major parts- feature extraction and classification. The deep GP-based MER system better captures relationships within highly complex data than does the conventional SVM and standard GP methods. The experimental results revealed better performance than the baseline in average.

In future work, we will apply the deep GP to the MER system in terms of arousal and valence values. Each music data will be annotated with an emotion as a point in the VA plane. We will consider the prediction of VA values as a regression problem. Moreover, deep GP can be inherently extended to a classifier by using an appropriate nonlinear activation function. We would like to use directly the deep GP classifier for emotion recognition.

REFERENCES

[1] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161-1178, 1980.

[2] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent system and Technology*, vol. 3, no. 3, May. 2012.

[3] B. Han, S. Rho, and R. B. Dannenberg, and E. Hwang, "SMERS: Music emotion recognition using support vector regression," in *Proc. Int. Conf. Music Information Retrieval*, Kobe, Japan, 2009.

[4] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 467-476, Feb. 2008.

[5] K. Markov and T. Matsui, "Music genre and emotion recognition using Gaussian processes," in *IEEE Access,* vol. 2, pp. 688-697, Jun. 2014.

[6] A. Daminaou and N. Lawrence, "Deep Gaussian processes," *JMLR*, 31:207-215, 2014.

[7] C. Y. Chang, C. Y. Lo, C. J. Wang, and P. C. Chung, "A music recommendation system with consideration of personal emotion," *Computer Symposium (ICS), 2010 International*, pp.18-23, 16-18 Dec. 2010.

[8] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol.16, pp. 448-457, Feb. 2008.

[9] Y. Ephraim and H. L. Van-Trees, "A signal subspace approach for speech enhancement," in *IEEE Transactions on Speech Audio Processing*, vol. 3, no. 4, pp. 251 -266, Jul. 1995.

[10] C. E. Rasmussen and C. K. I. Williams, "Gaussian Processes in Machine Learning," *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[11] M. Titsias and N. Lawrence, "Bayesian Gaussian process latent variable model," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, May. 2010.

[12] F. Zhu , T. Carpenter , D. R. Gonzalez , M. Atkinsonand J. Wardlaw, "Computed tomography perfusion imaging denoising using Gaussian process regression," *Phys. Med. Biol.*, vol. 57, no. 12, pp.N183 -N198, 2012.

[13] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 65-68.

[14] F. C. Hwang, J. S. Wang, P. C. Chung, and C. F. Yang, "Detecting emotional expression of music with feature selection approach," in *Proc. Int. Conf. Orange Technologies (ICOT)*, March. 2013, pp. 282-286, 12-16.

[15] "The All Music Guide," Available: http://www.allmusic.com

[16] "Last.fm," Available : http://cn.last.fm/home

[17] A. Damianou, C. Ek, M. Titsias, and N. Lawrence, "Manifold relevance determination," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 145–152.

[18] B. Jensen, J. Saez Gallego, and J. Larsen, "A predictive model of music preference using pairwise comparisons," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, March. 2012, pp.1977-1980, 25-30.