Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Using psychophysiological measures to recognize personal music emotional experience*

Le-kai ZHANG[†1,2], Shou-qian SUN[†1], Bai-xi XING[†‡1,2], Rui-ming LUO[†1], Ke-jun ZHANG[1]

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*School of Design, Zhejiang University of Technology, Hangzhou 310023, China*

[†]E-mail: zlkzhang@zju.edu.cn; ssq@zju.edu.cn; sisyxing@gmail.com; joeluo@zju.edu.cn

Received Feb. 9, 2018; Revision accepted Apr. 13, 2018; Crosschecked July 12, 2019

Abstract: Music can trigger human emotion. This is a psychophysiological process. Therefore, using psychophysiological characteristics could be a way to understand individual music emotional experience. In this study, we explore a new method of personal music emotion recognition based on human physiological characteristics. First, we build up a database of features based on emotions related to music and a database based on physiological signals derived from music listening including EDA, PPG, SKT, RSP, and PD variation information. Then linear regression, ridge regression, support vector machines with three different kernels, decision trees, k-nearest neighbors, multi-layer perceptron, and Nu support vector regression (NuSVR) are used to recognize music emotions via a data synthesis of music features and human physiological features. NuSVR outperforms the other methods. The correlation coefficient values are 0.7347 for arousal and 0.7902 for valence, while the mean squared errors are 0.023 23 for arousal and 0.014 85 for valence. Finally, we compare the different data sets and find that the data set with all the features (music features and all physiological features) has the best performance in modeling. The correlation coefficient values are 0.6499 for arousal and 0.7735 for valence, while the mean squared errors are 0.029 32 for arousal and 0.015 76 for valence. We provide an effective way to recognize personal music emotional experience, and the study can be applied to personalized music recommendation.

Key words: Music; Emotion recognition; Physiological signals; Wavelet transform
https://doi.org/10.1631/FITEE.1800101 **CLC number:** TP391.4

1 Introduction

Music is so expressive that it represents a good carrier for human emotion. People often listen to a song to adjust their own state to get rid of bad feelings or motivate themselves (Mori and Iwanaga, 2017). As a matter of fact, feelings about the same song vary from person to person. The evidence of physiological variation stimulated by music is obvious and empirical. Thus, it is favorable for utilization in personal music emotion recognition. We aim to use psychophysiological measures for emotion-based music recognition, since human emotional states can influence physiological changes, and, in turn, such physiological features can be used to reflect human emotions during the experience of listening to music (Maia and Furtado, 2016). The relationship between the physical state of the individual and the music emotional experience has the opportunity to make the computer aware of personal music preference for effective music recommendation.

Many existing studies on the recognition of music emotions focus mainly on the features and characteristics of the music itself as a multimedia modality.

‡ Corresponding author

* Project supported by the Philosophy and Social Science Planning Fund Project of Zhejiang Province, China (No. 20NDQN297YB) and the National Natural Science Foundation of China (No. 61702454)

ORCID: Le-kai ZHANG, http://orcid.org/0000-0002-8136-5882

However, the multimedia features may have limitations or underestimate the recognition results since they ignore human physiological variations induced by listening to music. Fortunately, there are some existing studies and methods that provide significant support for this research. Many researchers have explored recognition of emotions through physiological features (Picard et al., 2001; Rani et al., 2006; de Witte et al., 2017), in which music (Kim and André, 2008), video (Agrafioti et al., 2012), image (Katsis et al., 2011), and advertising (Li et al., 2017) were commonly selected as the stimuli to induce human emotions in the experiments. However, most of the studies were carried out in the lab and the databases were independent without cross-verification. Nevertheless, a breakthrough in application potential was required for industrial platforms.

Thus, this study pursued the possibility of using human psychophysiological features as an index for recognition of music. Instantaneous human psychophysiological states can reveal the exquisite and true feelings about a piece of music. Music features including melfrequency cepstral coefficient (MFCC), centroid, flux, and roll-off, and physiological features including electrodermal activity (EDA), photoplethysmography (PPG), skin temperature (SKT), respiration (RSP), and pupil diameter (PD) variation information were all collected to form a database for emotion recognition modeling. A model comparison was carried out for data synthesis of physiological and music features with the aim of a better understanding of how an individual perceives emotions when listening to music.

2 Related work

Here we would like to provide some selected reviews on emotion recognition with physiological data and multimedia data from various aspects to show the systematic research in this area. There has been much evidence supporting the usefulness of physiological signals in music emotion recognition (Krumhansl, 1997; Nyklíček et al., 1997; Mitterschiffthaler et al., 2007; Li et al., 2016). Different feature data sets have been applied to perform music emotion recognition, and they can be roughly categorized into two groups, the music audio features and the physiological features induced by music. In recent studies, many researchers have built up affective physiological response databases to learn the emotive expression pattern of physiological signals, including electromyogram (EMG), electroencephalography (EEG), electrocardiograph (ECG), PPG, eye tracking, and respiration data. In addition, audio features extracted from music clips such as MFCC, flux, centroid, spectrum, and roll-off can present music emotions effectively. Based on the existing multimodal feature data sets, machine learning approaches have commonly been deployed in emotion pattern exploration.

2.1 Music emotion recognition based on audio data

By music emotion recognition based on audio data, audio feature information is extracted from music clips. Music emotion recognition has been widely used in music emotion retrieval and recommendation. Lu et al. (2010) and Yang and Chen (2011) compared various machine learning approaches for music emotion recognition and proved that music features including MFCC, centroid, flux, and roll-off are useful. Ayadi et al. (2011) compared the classifiers of hidden Markov model (HMM), Gaussian mixture model (GMM), artificial neural network (ANN), and support vector machine (SVM) in speech emotion recognition and considered that the choice of features, the classification scheme, and the preparation of an emotional speech database were crucial for improving the results. It was believed that audio features could convey the emotional content of a music clip, while features of the pitch, energy, timing, and spectrum were the commonly used features within the studies. However, to achieve better results, recognition based on physiological data should also be considered.

2.2 Music emotion recognition based on physiological data

By music emotion recognition based on physiological data, rich feature information is extracted from physiological signals that can reveal the emotions induced by listening to music. In the past decade, there have been a considerable number of studies recognizing emotion by physiological features. First, Picard et al. (2001) recorded the physiological data from blood volume pulse (BVP), heart rate (HR), RSP, EMG, and skin conductance (SC)

of a subject who tried to experience eight affective states, and achieved an emotion recognition rate of 81% by sequential floating forward search (SFFS) and Fisher projection (FP). Kim and André (2008) studied music emotion recognition and reached a recognition rate of 95% for distinct positive and negative emotions using EMG, ECG, SC, and RSP. In Wagner et al. (2005), emotion recognition of anger, surprise, sadness, and pleasure with physiological features from EMG, ECG, SC, and RSP led to a result of 92.05% with a cross-validation method, in which k-nearest neighbors (KNN), linear discriminant function, and multilayer perceptron were compared for the optimal model. A number of physiological features were proved relevant to valence and an arousal emotion response. Specifically, EMG and HR were shown to be more sensitive to valence and SC was more related to arousal (Gerdes et al., 2014). Chandler and Cornes (2012) also presented a report on a physiological analysis solution for a unique individual emotion state recognition, where EMG, galvanic skin response (GSV), facial expressions, and iris features were taken as effective features in emotion measurement. Using the three-dimensional (3D) Gabor feature with principal component analysis (PCA) selection could obtain an emotion classification recognition rate of 77.57% based on facial expression analysis (Yun and Guan, 2013).

It is worth mentioning that HR, GSR, and the first derivative of GSR (FD-GSR) are the most influential physiological signals in the research of emotion recognition based on physiological signals. Additionally, the pupillary response was addressed as an important physiological feature applied in emotion recognition. Ren et al. (2013) explored the affective patterns of stress using pupil variation features to achieve a good performance. The multi-data fusion of ECG, GSR, RSP, and pupillary response has been proved effective and superior to single physiological data sets (Koelstra et al., 2012). Among all the related studies, EMG, ECG, SC, RSP, GSV, and pupillary response were taken to be the strongest affective physiological data according to the literature. In addition, Koelstra et al. (2012) combined the physiological data (GSR, blood volume pressure, RSP, skin temperature, EMG, electrooculogram (EOG), and EEG) with multimedia content feature data (audio features and visual features) to form a fusion data set in emotion analysis, and the results showed that multimodal data fusion generally outperformed all of the single modalities.

2.3 Music emotion recognition based on audio-physiology data fusion

The affective computing research community has witnessed a boom in emotion recognition pattern learning from multimedia data and human physiological data. Regression and classification learning techniques have been widely used in emotion intelligence learning. Examples include Yang and Chen (2011) who used LibSVM to achieve R^2 of 58.3% for arousal and 28.1% for valence based on a regression approach; then they conducted a study and verified the effectiveness of using music emotion in video highlight extraction. A logistic regression method was motivated by the perspective of likelihood computing. Among all the regression methods, margin-based algorithms, like LibSVM, have attracted considerable attention because of their flexibility. SVM has achieved competitive performance in handling complex and high-dimensional data in affective multimedia data and physiological data computing. Wen et al. (2014) conducted a comprehensive review of various classification and regression methods applied in emotion recognition. Among all the existing studies, classification methods have more commonly been used than regression methods when emotions were usually considered as categories.

In a sense, the fusion data structure of physiological features and music features could enrich the emotion identification evidence in the literature overview. Hence, we will collect audio data and physiological data and perform music emotion recognition using different machine learning methods to verify the results.

3 Experiment

3.1 Music stimuli

The affective stimuli material database used for this work consisted of 420 music clips from an instrument music library (Xing et al., 2015). A period of 10 s of a key melody was extracted from each song to create the music stimuli. Each clip was in .wav format with a 16-kHz sampling rate, ensuring music information integrity for feature extraction. The collection of 420 music clips was divided into 21 groups;

each group consisted of 20 music clips for the emotion labeling experiment, and each group contained five happy music clips (high arousal and high valence, e.g., songs of celebration), five soothing music clips (low arousal and high valence, e.g., songs of a river), five sad music clips (low arousal and low valence, e.g., songs of tragedies), five music clips of tense emotion (high arousal and low valence, e.g., songs of a storm), to make a balanced stimuli material set covering all the four quadrants of the valence-arousal space for the listening sessions. Since it would be more flexible and capable of extending the set emotions as a two-dimensional (2D) vector rather than as an independent emotion class, the valence-arousal dimensional model was used in this experiment for its advantage for precise emotion annotation.

3.2 Physiological state measurement and music emotion labeling experiment

The experiment was arranged in two sessions: a music listening session and an emotion labeling session. Fig. 1 presents the physiological signals recorded in a music listening session, and Fig. 2 shows the multimodal sensors placed on the volunteer in the experiment. To examine the feasibility of relating physiological measures to participants' listening experiences, the experiments were conducted in two laboratory rooms with controlled illumination and temperature. One room was for the participants who performed the experiment, and the other room was for the experimenter who recorded the physiological signals.

Each participant was asked to read the instructions for the task flow and required procedure. Once the participants were clear on the experimental process, he/she was led to the laboratory room. After all bio-sensors were placed and their signals checked, the participants were trained to perform a practice trial to familiarize themselves with the system. The detailed process is shown in the following steps:

1. Twenty-one volunteers (aged 20–36 years) were invited to participate in the experiment, and they had no specific music training.

2. The volunteer was situated in a separate room with a constant lighting setting, seated in front of a 21-inch screen, and an ordinary view of scenery was displayed on the screen during the experiment.

3. The volunteer wore the EDA, PPG, SKT, and RSP multi-channel sensors, and calibration was completed on a TOBII X2-30 to obtain the PD variation.

4. Listening session: Each volunteer listened to three groups of 60 music clips, and each group included 20 music clips, with a 20-s break between each clip.

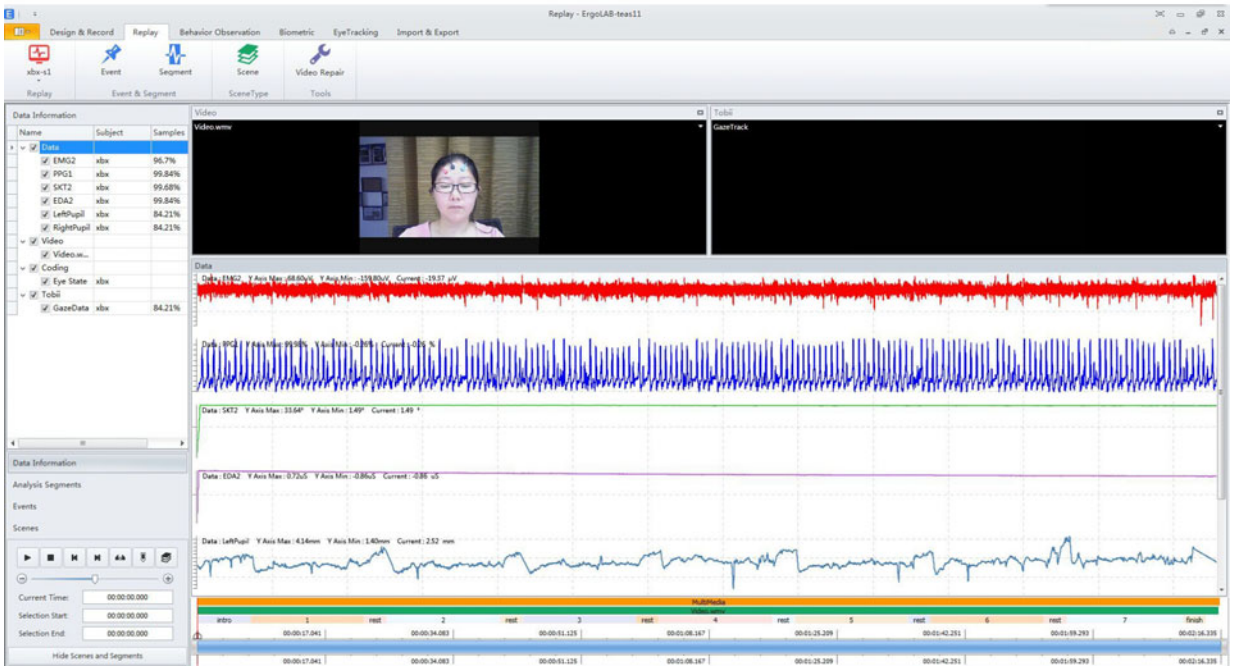

Fig. 1 Multi-channel physiological signals collected in the music listening experiment

5. Emotion labeling session: The volunteer was asked to do the emotion labeling work on the clips immediately on a scale of 0 to 1 for both arousal and valence dimensions according to his/her listening experience after the listening session.

To ensure high quality annotations, we developed a web interface where the volunteer could dynamically annotate the songs on valence and arousal dimensions separately. The self assessment manikin (SAM) questionnaire (Bradley and Lang, 1994) used for the annotations is shown in Fig. 3. The entire experiment lasted about 50 min.

Fig. 2 Multimodal sensors placed on the volunteer in the experiment

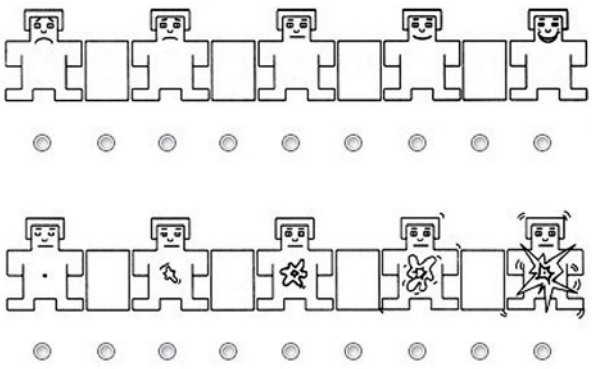

Fig. 3 Screenshot of the web interface for subjective emotion assessment

4 Feature extraction and emotion recognition

4.1 Music feature extraction

We extracted 64 dimensions of music features from each music clip using Marsyas, which is an effective tool (Tzanetakis and Cook, 2000) commonly used in music information analysis, including 52 dimensions of the MFCC feature, 4 dimensions of the spectral centroid feature, 4 dimensions of the spectral roll-off feature, and 4 dimensions of the spectral flux feature (Table 1).

Table 1 Features extracted from music information

Music feature category	Number
Melfrequency cepstral coefficient	52
Centroid feature	4
Roll-off feature	4
Flux feature	4

4.2 Physiological feature analysis

4.2.1 Physiological signal processing

There are four steps in physiological signal processing: (1) signal noise reduction by moving the average filter and wavelet transform; (2) signal segmentation, where each signal data record was divided into 60 segments with a period of 10 s corresponding to each music clip's listening experience; (3) signal decomposition by six levels of Db5 wavelet transform; (4) obtaining the statistical values of the denoised PD variation signals. The signal sequences for EDA, PPG, SKT, RSP, and PD were collected in the experiment by ErgoLab version 2.0 software, which was compatible with the multi-channel physiology instrument and Tobii X2-30.

In signal decomposition, first, the denoised original signal would be filtered into high-pass and low-pass signals, separately. Then the corresponding coefficients of the detailed coefficients from high-pass signals and approximation coefficients from low-pass signals would be obtained as the signal features. Each level of decomposition would produce different coefficients. The mechanism of decomposition was a sampling rate reduction by half in each level so that the coefficients' values would have distinctive differences between levels 1 and 6, while levels 1–6 would have a gradual variation without much value differentiation. Thus, in this experiment, only coefficients of level 1 and level 6 decomposition signals were selected to represent the signal information. The decomposition mechanism is shown in Fig. 4, where $H_1(i)$ is the high-pass filter to obtain the detailed signal D_i and $H_0(i)$ is the low-pass filter to obtain the approximation decomposed signal A_i.

4.2.2 Physiological feature extraction

Due to the good performance of the discrete wavelet transform (DWT) in discrete physiological signal analysis (Ren et al., 2013; Zhang et al., 2016; Wang et al., 2018), we applied a Db5 wavelet with six levels of decomposition to analyze the signal, using the Matlab wavelet toolbox to extract the wavelet features of the physiological signals (Cheng and Liu, 2008; Zhu, 2010). The discrete wavelet transform equation is defined as

$$\mathrm{DWT}_{\psi}x(m,n)=\int_{-\infty}^{\infty}x(t)\psi_{m,n}(t)\mathrm{d}t, \tag{1}$$

where $\psi_{m,n}(t)=2^{m/2}\psi(2^m t-n)$ is the dilated and translated version of the mother wavelet $\psi(t)$.

After DWT processing, we used statistical methods to analyze the original signal, the detailed signals for levels 1 and 6 (DET1 and DET6), the approximation signals for levels 1 and 6 (APP1 and APP6), and the coefficients of DET1, DET6, and APP1. The statistical features were obtained as the values of max, min, mean, range, std, median, MedAD, and MeanAD of the original signals and decomposed signals. Wavelet energy was also extracted as an important feature to form the database. Finally, a total of 438 features from the extracted EDA, PPG, SKT, RSP, and PD variation signals were collected to form the physiological feature data set.

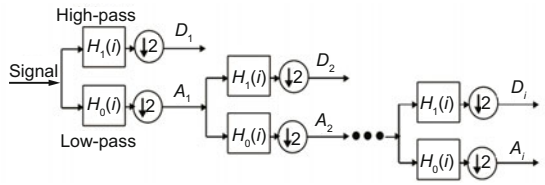

Fig. 4 Wavelet decomposition mechanism. High-pass: high-pass decomposition filter; low-pass: low-pass decomposition filter; ↓2: downsampling operation. A_1, A_2, and A_i are the approximation coefficients of the original signal at levels 1, 2, ..., i, respectively. D_1, D_2, and D_i are the detailed coefficients at levels 1, 2, ..., i, respectively

The physiological feature information is presented in Table 2.

4.3 Emotion recognition algorithms

In this study, we used Scikit-learn (Pedregosa et al., 2011), a free machine learning library for Python, to build the regression model. To verify the efficiency and robustness of the algorithms we chose, a 10×10 cross-validation was applied to build the model.

Scikit-learn provides several popular regression algorithms. We implemented nine of them, including linear regression (LR), ridge regression (RR), support vector machines with three different kernels, decision trees, k-nearest neighbors (KNN), multi-layer perceptron (MLP), and Nu support vector regression (NuSVR). A brief description of each method is as follows:

1. LR fits a linear model with coefficients $\boldsymbol{w}=(w_1,w_2,\ldots,w_p)^{\mathrm{T}}$ to minimize the residual sum of squares between the observed responses in the dataset and the responses predicted by the linear approximation:

$$\min_{\boldsymbol{w}}||\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}||_2^2. \tag{2}$$

2. RR addresses the problems of ordinary least squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares:

$$\min_{\boldsymbol{w}}||\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}||_2^2+\alpha||\boldsymbol{w}||_2^2. \tag{3}$$

Here, α is complexity parameter that controls the amount of shrinkage and $||\boldsymbol{w}||_2$ is the ℓ_2-norm of the parameter vector.

3. Support vector regression is a method extended from support vector classification to solve regression problems. Various kernel functions (linear: $\langle\boldsymbol{x},\boldsymbol{x}'\rangle$; polynomial: $(\gamma\langle\boldsymbol{x},\boldsymbol{x}'\rangle+r)^d$; RBF: $\exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2)$; etc.) can be specified and the

Table 2 Features extracted from physiological signals

Feature	Description	Number
Wavelet statistic	Values of max, min, mean, range, std, median, MedAD, and MeanAD of the original signal and decomposed signals, including APP (levels 1 and 6), DET (levels 1 and 6), APP coefficients (levels 1 and 6), and DET coefficients (levels 1 and 6)	432
Wavelet energy	Wavelet energy of APP signals (levels 1 and 6) and DET signals (levels 1 and 6)	6

APP: approximation signals; DET: detailed signals

free parameters are penalty factor C and relaxation factor epsilon.

4. NuSVR is a variant of SVR which uses a parameter ν, an upper bound on the fraction of training errors, and a lower bound of the fraction of support vectors, to control the number of support vectors.

5. Decision trees is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

6. k-nearest neighbors is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. Neighbor-based regression can be used in cases where the data labels are continuous rather than discrete. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors.

7. Multi-layer perceptron is a supervised learning algorithm that learns a mapping $f(\cdot):\mathbb{R}^i\rightarrow\mathbb{R}^o$ by training on a dataset, where i is the number of dimensions for input and o the number of dimensions for output. Using one nonlinear layer or more than one nonlinear layer, called hidden layer(s), it can learn a nonlinear function approximator for either classification or regression.

5 Results and discussion

5.1 Emotion recognition based on different methods

In the emotion recognition experiment, the valence-arousal (V-A) scoring results and physiological and music features formed the database. To learn an emotion vector regression rule, a training data set sample is typically given. This is produced by the experiment with a balanced distribution in the V-A dimension. Here, the features generated from the signal processing of the physiology and music data constructed the input feature vectors, and the V-A labeling values formed the output label. The data set has 1260 samples, in which a 10×10 cross-validation was used to obtain the average performance result.

The emotion of one physiological signal sequence was defined by the combination of the recognition of the valence and arousal values. Set $\boldsymbol{\theta}_{\mathrm{e}}$ to represent the emotion vectors for each signal, $\boldsymbol{V}_{\mathrm{v}}$ the valence vectors, and $\boldsymbol{V}_{\mathrm{a}}$ the arousal vectors, and then $\boldsymbol{\theta}_{\mathrm{e}}$ could be obtained by the optimal recognition model based on the physiological variation feature vectors:

$$\boldsymbol{\theta}_{\mathrm{e}}=[\boldsymbol{V}_{\mathrm{v}},\boldsymbol{V}_{\mathrm{a}}]. \tag{4}$$

We implemented the database in the ensemble learning methods of LR, RR, support vector machines with three different kernels, MLP, and NuSVR to train the classifier to find the best model. In review of the commonly used methods for affective data analysis, LR, RR, SVR (linear kernel), SVR (RBF kernel), SVR (poly kernel), decision trees, KNN, MLP, and NuSVR were listed as powerful regression tools and had enjoyed success in existing studies on physiological features, and thus these methods were proposed for comparison for the optimal model.

Five steps were taken to achieve the modeling results:

Step 1: We made a collection of 1260 instances of physiological signals.

Step 2: We used Matlab to extract 438 dimensions of physiological features from each signal.

Step 3: The emotion label values, 64 music features, and 438 physiological features were combined to form the emotion database, and a 10×10 cross-validation was applied to run the results.

Step 4: The ANOVA method was applied in feature selection to find the most relevant features with $p<0.001$. As a result, 228 relevant arousal features were selected from 502 features and 226 relevant valence features were selected from all of the 502 originally extracted features. The most relevant features are listed in detail in Tables 3 and 4 for valence and arousal, respectively. Then the principal component analysis (PCA) method was applied to the relevant feature sets. Thus, a 39-principal-feature combination was formed for arousal recognition and a 40-principal-feature combination was generated for valence recognition by PCA.

Step 5: In the experiment, we compared LR, RR, SVR (linear kernel), SVR (RBF kernel), SVR (poly kernel), decision trees, KNN, MLP, and NuSVR with the most relevant feature data set to find the optimal model, separately.

The comparisons of these machine learning approaches are presented in Table 5. Note that all approaches used the default parameter value provided

Table 3 Valence-relevant features from ANOVA analysis ($p < 0.001$)

Signal	Feature content description	Number
EDA	DET level 1, DET level 1 Coe, APP level 6 Coe: mean, max, min, MedAD, range	6
PPG	Original signal, APP levels 1 and 6, APP levels 1 and 6 Coe, DET levels 1 and 6, DET levels 1 and 6 Coe: mean, max, min, range, std, median, MedAD, MeanAD	57
RSP	DET level 6 Coe: max, min, range, std, MedAD, MeanAD	6
SKT	Original signal, APP levels 1 and 6, APP levels 1 and 6 Coe, DET levels 1 and 6, DET levels 1 and 6 Coe: energy, mean, max, min, range, std, median, MedAD, MeanAD	54
PD	Original signal, APP levels 1 and 6, APP levels 1 and 6 Coe, DET level 1, DET level 1 Coe: energy, mean, max, range, std, median, MedAD, MeanAD	40
Music	Spectral centroid, spectral flux, MFCC, Spectral roll-off	63

APP: approximation signals; DET: detailed signals; Coe: coefficients

Table 4 Arousal-relevant features from ANOVA analysis ($p < 0.001$)

Signal	Feature content description	Number
EDA	DET levels 1 and 6 Coe: mean, max, min, median, range	5
PPG	Original signal, APP levels 1 and 6, APP levels 1 and 6 Coe, DET levels 1 and 6, DET levels 1 and 6 Coe: mean, max, min, range, std, median, MedAD, MeanAD	59
RSP	DET level 6 Coe: max, min, range, std, MedAD, MeanAD	6
SKT	Original signal, APP levels 1 and 6, APP levels 1 and 6 Coe, DET levels 1 and 6, DET levels 1 and 6 Coe: energy, mean, max, min, range, std, median, MedAD, MeanAD	65
PD	Original signal, APP levels 1 and 6, APP levels 1 and 6 Coe: energy, mean, max, range, std, median, MedAD, MeanAD	29
Music	Spectral centroid, spectral flux, MFCC, spectral roll-off	64

APP: approximation signals; DET: detailed signals; Coe: coefficients

Table 5 Comparison of different algorithms for emotion recognition results

Algorithm	Arousal		Valence	
	MSE	CC	MSE	CC
Decision tree	0.047 78	0.2395	0.037 27	0.2335
SVR (linear kernel)	0.046 49	0.4371	0.027 29	0.5762
Linear regression	0.043 59	0.4465	0.026 75	0.5812
Ridge regression	0.043 34	0.4474	0.026 74	0.5813
MLP	0.055 73	0.5471	0.035 51	0.6103
SVR (poly kernel)	0.030 99	0.6426	0.025 93	0.6574
k-nearest neighbors	0.027 22	0.6774	0.019 10	0.7179
SVR (RBF kernel)	0.026 76	0.7031	0.017 67	0.7605
NuSVR	**0.023 23**	**0.7347**	**0.014 85**	**0.7902**

by the Scikit-learn function; that is, $\alpha = 1.0$ for RR, $C = 1.0$ and epsilon $= 0.1$ for SVR, $\nu = 0.5$ and the RBF kernel for NuSVR, and $k = 5$ for KNN. As for MLP, there is one hidden layer with 100 neurons; the activation function is relu, the solver for weight optimization is lbfgs, and the learning rate is 0.001.

According to the results, NuSVR apparently outperformed the others, in which the correlation coefficient (CC) value was 0.7347 for arousal, 0.7902 for valence, while the mean squared error (MSE) was 0.023 23 for arousal and 0.014 85 for valence. To compare the different combinations of all the data fusion, we implemented NuSVR to calculate the recognition results.

To further tune ν and C, we set ν from 0.1 to 1.0 with step size 0.1, and calculated the 10×10 cross-validation MSE for arousal and valence, separately. The results are plotted in Fig. 5. Similarly, C was ranged from 0.5 to 5.0 with step size 0.5, and the results are plotted in Fig. 6. In Fig. 5, when $\nu > 0.4$, the MSE curve tends to be stable. In Fig. 6, when $C = 1.0$, NuSVR achieved the lowest MSE for both arousal and valence. Considering model complexity and generalization ability, it is not recommended to

choose large ν and C. Thus, we chose NuSVR with $\nu = 0.5$ and $C = 1.0$ as our regression model.

5.2 Emotion recognition based on multimodal data fusion

We also explored the recognition results of different data fusion sets to see which kinds of modalities contribute most. We implemented NuSVR to calculate the different combinations of all the data fusions. We studied each single modality and multi-modality to provide advice for future application development. The single modality and multi-modality recognition results are shown in Tables 6 and 7, respectively. In all the data fusion comparisons, the data set with all the features had the best performance in modeling. The best single modality was the SKT data set with a model correlation coefficient (CC) value of 0.491 for arousal, 0.539 for valence, and mean squared error (MSE) value of 0.0452 for arousal and 0.0342 for valence.

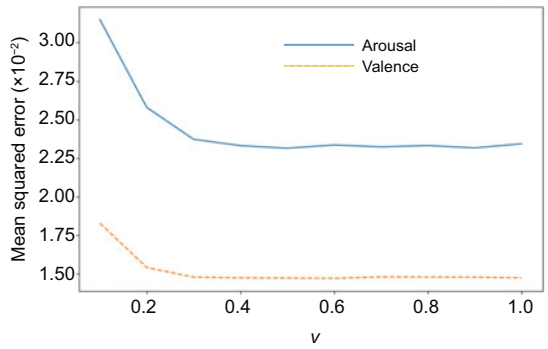

Fig. 5 Mean squared error (MSE) for arousal and valence with different ν

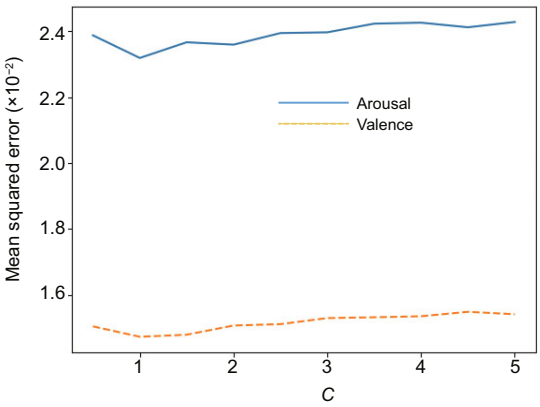

Fig. 6 Mean squared error (MSE) for arousal and valence with different C

When it came to the data fusion modeling experiment, the best three physiological modalities were formed using PPG, SKT, and PD with a CC of 0.553 for arousal and 0.616 for valence, with an MSE of 0.0351 for arousal and 0.0246 for valence. This had surpassed the best result of a single modality. The results indicated that EDA and respiration achieved a relatively low recognition performance in the experiment. However, the performance could not be improved by ignoring these two modalities. The best performance was achieved by the data fusion of all the modalities, in which the CC value was 0.6499 for arousal and 0.7735 for valence, while the MSE was 0.0293 for arousal and 0.0157 for valence. Apparently, data fusion is a promising method and a more extensive range of multimodal data should be grouped for significant recognition of affective states in future studies. In addition, a deeper exploration of

Table 6 Comparison results of single-modality emotion recognition

Feature	Arousal		Valence	
	MSE	CC	MSE	CC
RSP	0.0509	0.202	0.0407	0.205
EDA	0.0468	0.275	0.0368	0.271
PD	0.0421	0.413	0.0367	0.303
Music	0.0437	0.425	0.0373	0.420
PPG	0.0479	0.333	0.0337	0.425
SKT	**0.0452**	**0.491**	**0.0342**	**0.539**

Table 7 Comparison results of multi-modality emotion recognition

Modality fusion	Feature	Arousal		Valence	
		MSE	CC	MSE	CC
All physiological signals	EDA, PPG, R, SKT, PD	0.0363	0.540	0.0256	0.611
Best three physiological signals	PPG, SKT, PD	0.0351	0.553	0.0246	0.616
Best three physiological signals & musical features	PPG, SKT, PD, Music	0.0312	0.623	0.0201	0.721
All physiological signals & musical features	EDA, PPG, R, SKT, PD, Music	**0.0293**	**0.6499**	**0.0157**	**0.7735**

feature mining from the signals and multimedia information could be an opportunity for improvement.

6 Conclusions and future work

We considered emotional state recognition based on music considering how emotion was implemented within the human physiological system and how emotion was expressed in features expressing music information. Physiological variations and musical stimuli could be linked by emotional properties. We strived to explore the linking patterns between physiological variations and musical features and establish a firm foundation for a novel approach in various emotion-driven and intelligent interaction platforms. We discussed a database built upon an emotion experiment procedure, signal processing methods, selection of feature variables, and the choice of learning algorithms for affective computational issues. The results provided a promising way to make the computer aware of personal music preference for effective music recommendation.

The contributions of this study included the following: (1) A physiological feature database and a music emotion feature database were built; (2) On the basis of these two databases, we compared LR, RR, SVR (linear kernel), SVR (RBF kernel), SVR (poly kernel), decision trees, k-nearest neighbors (KNN), MLP, and NuSVR to reveal the emotion patterns in different data fusions, which thus helped achieve the best performance by NuSVR; (3) In all the data fusion comparisons, the data set with all the features (music features and all physiological features) had the best performance in modeling.

In the future, we would like to expand the database to improve the recognition rate. In addition, with the development of wearable computing and mobile computing devices, we will propose an application using physiological and music data fusion to recommend music automatically based on personal preference. It would offer new music recommendation experiences.

Compliance with ethics guidelines

Le-kai ZHANG, Shou-qian SUN, Bai-xi XING, Ruiming LUO, and Ke-jun ZHANG declare that they have no conflict of interest.

References

Agrafioti F, Hatzinakos D, Anderson AK, 2012. ECG pattern analysis for emotion detection. *IEEE Trans Affect Comput*, 3(1):102-115. https://doi.org/10.1109/T-AFFC.2011.28

Ayadi ME, Kamel MS, Karray F, 2011. Survey on speech emotion recognition: features, classification schemes, and databases. *Patt Recogn*, 44(3):572-587. https://doi.org/10.1016/j.patcog.2010.09.020

Bradley MM, Lang PJ, 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *J Behav Ther Exp Psych*, 25(1):49-59. https://doi.org/10.1016/0005-7916(94)90063-9

Chandler C, Cornes R, 2012. Biometric measurement of human emotions. *TECHNIA - Int J Comput Sci Commun Technol*, 4(2):164-168.

Cheng B, Liu GY, 2008. Emotion recognition from surface EMG signal using wavelet transform and neural network. *J Comput Appl*, 28(2):333-335 (in Chinese).

de Witte NAJ, Sütterlin S, Braet C, et al., 2017. Psychophysiological correlates of emotion regulation training in adolescent anxiety: evidence from the novel PIER task. *J Affect Disord*, 214:89-96. https://doi.org/10.1016/j.jad.2017.03.012

Gerdes ABM, Wieser MJ, Alpers GW, 2014. Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. *Front Psychol*, 5:1351. https://doi.org/10.3389/fpsyg.2014.01351

Katsis CD, Katertsidis NS, Fotiadis DI, 2011. An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders. *Biomed Signal Process Contr*, 6(3):261-268. https://doi.org/10.1016/j.bspc.2010.12.001

Kim J, André E, 2008. Emotion recognition based on physiological changes in music listening. *IEEE Trans Patt Anal Mach Intell*, 30(12):2067-2083. https://doi.org/10.1109/TPAMI.2008.26

Koelstra S, Muhl C, Soleymani M, et al., 2012. DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans Affect Comput*, 3(1):18-31. https://doi.org/10.1109/T-AFFC.2011.15

Krumhansl CL, 1997. An exploratory study of musical emotions and psychophysiology. *Can J Exp Psychol*, 51(4): 336-353. https://doi.org/10.1037/1196-1961.51.4.336

Li C, Xu C, Feng ZY, 2016. Analysis of physiological for emotion recognition with the IRS model. *Neurocomputing*, 178:103-111. https://doi.org/10.1016/j.neucom.2015.07.112

Li SS, Walters G, Packer J, et al., 2017. A comparative analysis of self-report and psychophysiological measures of emotion in the context of tourism advertising. *J Travel Res*, 57(8):1078-1092. https://doi.org/10.1177/0047287517733555

Lu Q, Chen XO, Yang DS, et al., 2010. Boosting for multimodal music emotion classification. 11th Int Society for Music Information Retrieval Conf, p.105-110.

Maia CLB, Furtado ES, 2016. A study about psychophysiological measures in user experience monitoring and evaluation. Proc 15th Brazilian Symp on Human Factors in Computing Systems, p.7. https://doi.org/10.1145/3033701.3033708

Mitterschiffthaler MT, Fu CHY, Dalton JA, et al., 2007. A functional MRI study of happy and sad affective states induced by classical music. *Hum Brain Mapp*, 28(11):1150-1162. https://doi.org/10.1002/hbm.20337

Mori K, Iwanaga M, 2017. Two types of peak emotional responses to music: the psychophysiology of chills and tears. *Sci Rep*, 7:46063. https://doi.org/10.1038/srep46063

Nyklíček I, Thayer JF, van Doornen LJP, 1997. Cardiorespiratory differentiation of musically-induced emotions. *J Psychophysiol*, 11(4):304-321.

Pedregosa F, Varoquaux G, Gramfort A, et al., 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res*, 12(10):2825-2830.

Picard RW, Vyzas E, Healey J, 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans Patt Anal Mach Intell*, 23(10):1175-1191. https://doi.org/10.1109/34.954607

Rani P, Liu C, Sarkar N, et al., 2006. An empirical study of machine learning techniques for affect recognition in human-robot interaction. *Patt Anal Appl*, 9(1):58-69. https://doi.org/10.1007/s10044-006-0025-y

Ren P, Barreto A, Gao Y, et al., 2013. Affective assessment by digital processing of the pupil diameter. *IEEE Trans Affect Comput*, 4(1):2-14. https://doi.org/10.1109/T-AFFC.2012.25

Tzanetakis G, Cook P, 2000. MARSYAS: a framework for audio analysis. *Organ Sound*, 4(3):169-175. https://doi.org/10.1017/S1355771800003071

Wagner J, Kim J, Andre E, 2005. From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. IEEE Int Conf on Multimedia and Expo, p.940-943. https://doi.org/10.1109/ICME.2005.1521579

Wang SH, Phillips P, Dong ZC, et al., 2018. Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. *Neurocomputing*, 272:668-676. https://doi.org/10.1016/j.neucom.2017.08.015

Wen WS, Liu GY, Cheng NP, et al., 2014. Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Trans Affect Comput*, 5(2):126-140. https://doi.org/10.1109/TAFFC.2014.2327617

Xing BX, Zhang KJ, Sun SQ, et al., 2015. Emotion-driven Chinese folk music-image retrieval based on DE-SVM. *Neurocomputing*, 148:619-627. https://doi.org/10.1016/j.neucom.2014.08.007

Yang YH, Chen HH, 2011. Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Trans Audio Speech Lang Process*, 19(7):2184-2196. https://doi.org/10.1109/TASL.2011.2118752

Yun T, Guan L, 2013. Human emotional state recognition using real 3D visual features from Gabor library. *Patt Recogn*, 46(2):529-538. https://doi.org/10.1016/j.patcog.2012.08.002

Zhang YD, Yang ZJ, Lu HM, et al., 2016. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access*, 4:8375-8385. https://doi.org/10.1109/ACCESS.2016.2628407

Zhu X, 2010. Emotion recognition of EMG based on BP neural network. Proc 2nd Int Symp on Networking and Network Security, p.227-229.