Emotional States Associated with Music: Classification, Prediction of Changes, and Consideration in Recommendation

JAMES J. DENG and CLEMENT H. C. LEUNG, Hong Kong Baptist University
ALFREDO MILANI, University of Perugia
LI CHEN, Hong Kong Baptist University

We present several interrelated technical and empirical contributions to the problem of emotion-based music recommendation and show how they can be applied in a possible usage scenario. The contributions are (1) a new three-dimensional resonance-arousal-valence model for the representation of emotion expressed in music, together with methods for automatically classifying a piece of music in terms of this model, using robust regression methods applied to musical/acoustic features; (2) methods for predicting a listener's emotional state on the assumption that the emotional state has been determined entirely by a sequence of pieces of music recently listened to, using conditional random fields and taking into account the decay of emotion intensity over time; and (3) a method for selecting a ranked list of pieces of music that match a particular emotional state, using a minimization iteration method. A series of experiments yield information about the validity of our operationalizations of these contributions. Throughout the article, we refer to an illustrative usage scenario in which all of these contributions can be exploited, where it is assumed that (1) a listener's emotional state is being determined entirely by the music that he or she has been listening to and (2) the listener wants to hear additional music that matches his or her current emotional state. The contributions are intended to be useful in a variety of other scenarios as well.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*; H.5.5 [**Sound and Music Computing**]: Methodologies and Techniques; I.5 [**Pattern Recognition**]: Applications—*Signal processing*

General Terms: Algorithms, Design, Human Factors, Experimentation

Additional Key Words and Phrases: Musical emotion, emotional state, music emotion recognition, affective computing, conditional random fields, music recommendation

ACM Reference Format:
James J. Deng, Clement H. C. Leung, Alfredo Milani, and Li Chen. 2015. Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. ACM Trans. Interact. Intell. Syst. 5, 1, Article 4 (March 2015), 36 pages.
DOI: http://dx.doi.org/10.1145/2723575

1. INTRODUCTION

The goal of our work presented in this article is to provide algorithms and relevant empirical results that will help designers and developers of interactive intelligent systems for music listeners to take into account more effectively the emotions expressed in music and their effects on listeners' emotional states. One scenario in which these contributions can be applied, which we use as a reference scenario throughout this

Authors' addresses: J. J. Deng, C. H.C. Leung, and L. Chen, Department of Computer Science, Hong Kong Baptist University, Hong Kong; emails: {jdeng, clement}@comp.hkbu.edu.hk, lichen@comp.hkbue.du.hk; A. Milani, Department of Mathematics and Computer Science, University of Perugia, Italy; email: milani@unipg.it.

DOI: http://dx.doi.org/10.1145/2723575

article, is one in which a music recommender system predicts the changing emotional state of a listener on the basis of his or her recent listening and which recommends music intended to match the listener's current emotional state. But we believe that creative researchers and designers in this area will think of other scenarios in which our results can be applied, in whole or in part.

With the astounding growth of digital music, music recommendation is extremely important for people discovering music that fits their taste or preferences. Recommender systems have been widely studied in recent years, which are commonly classified into three categories: content-based, collaborative filtering, and hybrid approaches [Adomavicius and Tuzhilin 2005]. Although these approaches have worked well in certain circumstances and for certain items (e.g., movie, music, news), they may not always lead to the best possible designs for affective recommendation systems. Thus, it is necessary to explore new scenarios and solutions for such systems. Music has its own elements (e.g., pitch, dynamics, intensity, rhythm) and structure. Moreover, music is usually regarded as an art form and the soul of language that can make people produce emotional responses or induce their emotions [Juslin and Sloboda 2001; Scherer 2004; Zentner et al. 2008]. Thus, it is constructive and creative to taking into account emotions induced by music in design music recommendation systems. Besides, research on psychology, cognition, and human interaction has indicated that human emotions play a significant role in designing interactive intelligent systems. Therefore, music recommendation systems can benefit from consideration of listeners' dynamic emotions induced by music. However, most current music recommendation systems ignore this key point (emotional state) or simply employ a finite small number of emotional descriptors or labels (e.g., happy, sad, romantic) to analyze and model music, which fail to effectively recommend music to satisfy human high-level needs on the basis of emotion.

Many studies in psychology and physiology have been carried out to recognize and assess emotions induced by music. For example, physiological changes (e.g., heart rate, blood pressure, respiration rate) are often examined to evaluate emotions and their changes in music listening [Agrafioti et al. 2012; Lin et al. 2010; Kim 2008]. Some affective music systems based on physiological response have been designed and developed [Levenson 1988; Wijnalda et al. 2005], and have proved their effectiveness. However, special wearable or medical devices, such as electrocardiogram (ECG) sensors, are required to detect physiological biosignals (e.g., ECG and electromyogram (EMG)), which may result in difficulty and inconvenience in implementing such recommendation systems of music based on emotions detected from listeners. Some researchers in music information retrieval (MIR) seek to detect and recognize emotions induced by music through machine learning and pattern recognition techniques [Yang et al. 2008; Lu et al. 2006; Schmidt et al. 2010; Deng and Leung 2013]. Although the results of these methods have been generally satisfactory in certain circumstances, musical emotions are analyzed and modeled under static consideration without dynamic consideration. To the best of our knowledge, nearly no attempts have been made to analyze and model musical emotions and their changes over time in terms of music recommendation. Therefore, it is significant and encouraging to explore this direction, as such research work is essential to benefit interactive intelligent systems.

Considering that personal data (e.g., music listening history) can be easily collected and accessed, it can be used for designing personalized applications such as recommender systems. Therefore, in this article, we propose using machine learning techniques to predict the listener's dynamic emotional state according to his or her personal historical music listening list and construct an affective recommendation system of music. Before explanation of our work, we give some conditions and statements that allow the design of such an intelligent emotion-based music recommendation system.

—Emotion is a complex psychological and physiological human subjective experience, influenced by many factors such as time, location, mood, and even the listener's life experience, events in the recent past, and interpersonal relationships, thus incorporating all of these factors that affect a listener's emotion in determining his or her emotion is too complex and is beyond the scope of the present study. For the sake of simplicity and controllability, we make assumptions about the listener's emotion elicitation in such a simple scenario: the listener's emotion induced by music is only affected by music listened to, and emotion intensity changes are only associated with time.
—Music can express and reflect a listener's emotion. The listener's personal historical music listening list can be used to track possible changes of emotional state.

As collaborative filtering collects all listeners' data and behaviors to recommend music based on behaviors of the similar listeners, it may be not appropriate to build such a personalized affective recommendation system of music. Moreover, it suffers from problems such as cold start, sparsity, and big data, and thus we suggest applying emotion-based filtering of music content to build such a music recommendation system.

The goal of our work presented in this article is to construct an intelligent music recommendation system based on predicted dynamic emotion of the listener through his or her historical music listening list. To this end, we first need to recognize emotional states associated with music. Despite that some emotion models, such as the OCC model [Ortony et al. 1990], the two-dimensional arousal-valence model [Russell 1980], and the three-dimensional pleasure-arousal-dominance (PAD) model [Mehrabian 1980], have been widely applied to design various emotion-based applications (e.g., musicovery and stereomood), they still have limitations in the context of music and fail to model musical emotions in a computational and adaptive way. Then, we predict changes of the listener's emotion according to his or her historical music listening list. Since music expresses emotions, through the listener's historical music listening list in a session, we apply conditional random fields (CRF) [Lafferty et al. 2001] to compute the probabilities of different emotional states and choose the one with the highest probability as the listener's predicted emotional state. Finally, an affective recommendation system is conducted based on that predicted emotional state of the listener, in which minimization iteration of emotion similarity is carried out to recommend an optimized ranking list of music. To sum up, our main contributions in this article are the following:

(1) We combine discrete emotion theory and dimensional emotion theory to construct a hybrid model of emotion associated with music, along with methods for automatically classifying music in terms of that model. The results of Experiment 1 suggest the usefulness of our proposed emotion representation and classification methods.
(2) We simulate the dynamic changes over time in emotion induced by music, taking into account emotion intensity decay (EID), using particular mathematical functions and CRF. The results from self-reports and objective measurements suggest that our methods yield acceptable accuracy.
(3) We propose an algorithm of music recommendation that identifies music that matches the listener's predicted current emotional state, along with empirical results concerning its accuracy.

The remainder of this article is structured as follows. Section 2 gives a detailed review of the existing literature on emotion models and methods of emotion-based recommendation. Section 3 provides an overview of our methods for music emotion classification, prediction of changes, and recommendation. In Section 4, we introduce our musical emotion model and emotion-relevant musical audio features. The user study and experiment on music emotion recognition are reported in Section 5. In Section 6,

we present an EID model to predict changes of emotions and explain our method for music recommendation based on emotion. The experiments on listener's emotion state prediction and emotion-based music recommendation are given in Sections 7 and 8, respectively. Section 9 presents our conclusion and future work.

2. RELATED RESEARCH

This section first reviews the association between music and emotion, as well as the existing literature on emotion models and computational methods in affective computing. Then, the state of the art on music recommendation based on emotion is presented.

2.1. Music and Emotion

Although music is an art, it owns specific structure—a succession of tones through time—that involves a number of perceptual attributes such as intensity, pitch, rhythm, timbre, melody, tonality, and harmony. In contrast, emotion is a complex subjective and conscious experience, reflected by complicated psychophysiological expressions. Many psychologists attempt to give a clear definition of emotion. For example, Scherer [1993] defined emotion consisting of a synchronization of changes in organismic subsystems such as the central nervous system and the neuroendocrine system. However, to date, the literature is still controversial on the definition of emotion. As a result, although the study between music and emotion dates back hundreds of years, there still exist debates about whether music can induce emotion. The relationship between music and emotion has been further investigated by a number of studies [Bartlett 1996; Juslin and Sloboda 2001; Krumhansl 2002; Scherer et al. 2002; Juslin and Laukka 2004; Konečni 2008]. Some psychologists claim that music cannot induce real human emotions [Juslin and Sloboda 2001; Konečni 2008], whereas others take the opposite view and insist that music can certainly induce emotions [Meyer 1956; Scherer et al. 2002; Coutinho and Cangelosi 2011]. A detailed assessment of whether music can induce emotion is out the scope of this article but can be found in a number of neuroscience and psychological studies supporting the view that listeners can perceive emotions from music and that emotions induced by music are real. Consequently, we take the stance that music can induce emotions. Furthermore, we use the term *musical emotion* to describe emotions induced by music or emotions perceived from music.

Measurements on the expression of emotion in music was introduced much earlier [Seashore 1923]. Many works [Schubert 2004; Webster and Weir 2005] investigated perceptual attributes of music (e.g., pitch, intensity, rhythm, timbre, tonality) contributing to the expression of musical emotion. For example, happy music often has relatively rapid tempos, major modes, and relatively constant ranges of pitch and intensity, whereas sad music usually has slow tempos, minor modes, and fairly constant ranges of pitch and intensity. When people listen to music, these perceptual attributes may have an effect on emotion induction reflected by biophysical changes. The expression of these perceptual attributes are usually expressed by various acoustic features such as zero-crossing, energy, Mel-frequency cepstral coefficients (MFCCs), chroma, statistical spectrum descriptors, and octave-based spectral contrast. The relationship between these acoustic features and their emotional impact was introduced in Schmidt et al. [2010], Kim et al. [2010], and Sezgin et al. [2012]. Consequently, the expression of musical emotion can be represented by a bag of emotion-relevant acoustic features.

In reality, general listeners tend to utilize simple emotional adjectives or labels (e.g., happy, sad, tense) [Hevner 1936] to describe emotions expressed by music, whereas some psychologists attempt to adopt several dimensions to account for representation of musical emotions [Russell 1980; Thayer 1989; Mehrabian 1980]. Furthermore, emotion is assigned an attribute that measures emotion intensity in scale. The intensity of emotion is fundamentally temporal in nature, and it dynamically changes with

psychological response over time [Dean and Bailes 2010]. Several mathematical models (e.g., the inverse exponential model) [Picard 1997; Steunebrink et al. 2008] were used to simulate the process of EID. Through researchers' efforts, there have been plenty of emotion models and computational methods in affective computing applied in the past several decades. A detailed description will be given in the following section.

2.2. Emotion Models in Previous Work

Emotion is a complex psychological and physiological human experience, and thus various emotion representations or models have been proposed by different domain experts. These models are usually rooted in two emotion theories: discrete emotion theory and dimensional emotion theory [Scherer 2004]. A detailed comparison of discrete and dimensional emotion models is presented in Eerola and Vuoskoski [2010]. We review several widely used emotion models from these two theories as follows:

—*Discrete emotion theory* suggests employing a number of emotional descriptors or adjectives to express basic human emotions (e.g., joy, sadness, anger, contempt, happiness) [Izard and Malatesta 1987]. The classic study by Hevner [1936] investigated the relationship between music and emotion, and developed an adjective circle consisting of eight clusters totaling 67 emotional adjective terms to depict musical emotions. Ortony et al. [1990] proposed an emotion cognitive model, commonly known as the OCC model, to hierarchically describe 22 emotion descriptors. This emotion model benefits emotion classification in light of events, actions of agents, and aspects of objects. In the research community of Music Information Retrieval Evaluation eXchange (MIREX), five clusters of different emotion labels for music are gathered and widely used in the task of music mood classification. For the sake of simplicity, some researchers [Pohle et al. 2005] tend to express musical emotions by coarser-grained partition such as (soft, neutral, aggression) or (happy, neutral, sad) in musical emotion classification. Although the preceding discrete emotion models are easy to understand and employ, they still have shortcomings that limit their efficacy and adaptive capacities. First, a small number of primary emotion adjectives is not able to adaptively describe the enormous musical emotions. Second, the distinction between different emotional effects expressed by music is not always readily apparent (e.g., sadness, mourning), and thus it is difficult to accurately express emotions using these ambiguous emotion terms or labels. Finally, emotions induced by music are unlike the discrete psychological response. Therefore, some researchers attempt to build dimensional emotion models to overcome the limitations mentioned previously.

—*Dimensional emotion theory* states that emotion should be depicted in a psychological dimensional space. This theory benefits by representing a wide range of emotions not necessarily depicted by specific emotion descriptors. At present, there is no consensus on the dimensions of emotion [Gunes et al. 2011]. Significantly, Russell [1980] proposed representing emotion by linear combinations of two independent dimensions: arousal and valence. Arousal refers to the level of activation in stimuli, with a range from sleepy to aroused. Valence accounts for pleasantness and unpleasantness. This two-dimensional arousal-valence emotion model cannot very well distinguish finitely generated emotions by music such as calm and bored [Collier 2007]. Fontaine et al. [2007] indicated that emotion was not only two-dimensional. Thayer [1989] rotated arousal-valence dimensions 45 degrees to produce two separate physiological arousal dimensions: energetic arousal and tense arousal, which are convenient to measure subjective experience. Eerola et al. [2009] applied Thayer's emotion model to predict multidimensional emotional ratings from music and obtained a satisfactory result for music emotion classification. However, some researchers argue that energetic

arousal and tense arousal are mixtures of a single activation dimension and valence dimension [Yik et al. 1999], whereas others insist that they are not mixtures of valence and activation [Schimmack and Rainer 2002]. The PAD emotion model [Mehrabian 1980] employs three dimensions to represent emotion. The dimensions of arousal and valence are identical to Russell's model, and the dimension of dominance represents the control and dominance of the emotion, which is particularly useful in distinguishing among emotions having similar magnitude of arousal and valence. For example, anger is a dominant emotion, whereas fear is a submissive emotion. Compared to the arousal-valence model, the PAD model outperforms to distinguish them [Mehrabian 1996; Hanjalic 2006]. Gebhard [2005] constructed a layered structure that maps discrete emotions represented by the OCC emotion model into a three-dimensional PAD space, which is widely applied in affective computing. However, in the context of musical emotions, arousal and valence have high validity for expressing emotions induced by music, whereas dominance suffers from a problem that emotions induced by music are not appropriate to be described by dominant or submissive stimulus [MacDorman and Ho 2007]. Fortunately, Bigand et al. [2005] found that the third dimension seemed have an emotional character that can be measured by music characteristics like continuity–discontinuity or melodic–harmonic contrast, which inspired us to represent musical emotion in this way.

Apart from emotion models, emotion computation is another challenging task in affective computing. For example, how long an emotion lasts and the influence it has are always taken into account. Marsella et al. [2010] gave a review of recent work on computational models of emotion, where emotions were modeled as a continuous time-varying process in a specific period of time. Levenson [1988] analyzed initiation and completion of emotion by intensity and time through subjective self-report experiments. Gebhard [2005] used Euclidean distance to compute emotion intensity upon the PAD emotion model. Emotion plays an important role in decision making and behaviors, and the aim of *affective computing* is to design intelligent interactive systems that can understand and respond to human emotions. Picard [1997] attempted to apply mathematical models to simulate emotion decay by the inverse exponential model and emotion saturation by the logarithmic model. Furthermore, Picard proposed a computational framework for affective learning and decision making. Relying on the OCC model, qualitative and quantitative analysis of emotion were performed by Steunebrink et al. [2008]. In summary, these computational models and quantitative methods provide enlightenment for modeling time-varying musical emotions.

2.3. Current Methods for Emotion-Based Music Recommendation

Current music recommendation systems are usually classified into three categories: content based, collaborative filtering, and hybrid [Adomavicius and Tuzhilin 2005]. For example, Pandora radio adopts a content-based approach to recommend music with similar music characteristics to the ones that the listener has provided, whereas another popular music Web site, Last.fm, applies collaborative filtering to recommend music based on the listening behaviors of similar listeners. Netflix employs a hybrid strategy of combining content based and collaborative filtering to recommend movies. However, these recommendation systems fail to take into account emotion or sentiment induced or expressed by music. Here, we review the state of the art in music recommendation systems that take emotion into consideration.

Due to the impact of musical emotions and users' affective states, there is a considerable amount of research in this area. Tkalcic et al. [2011] provided a unifying

framework consisting of three stages to recommend music: detecting emotion state, entry stage, and consumption state. Most approaches are broadly classified into two groups according to the way of emotion detection or recognition: physiological measurement and subjective self-report measurement. As for physiological measurement, physiological changes of biosignals (e.g., heart rate, blood pressure) are often used to assess and recognize emotions in music listening. Many psychologists have made great efforts in emotion recognition from physiological biosignals [Agrafioti et al. 2012; Lin et al. 2010], and there are many works and systems related to emotion recognition from physiological signals. For example, Kim [2008] recognized emotion based on physiological changes acquired by EMG, skin conductivity, ECG, and respiration in music listening. Wijnalda et al. [2005] developed a personalized music system named "IM4Sports" for sports performance, where the user could select music with a rhythm matching the heart rate or running pace. The affective DJ [Healey et al. 1998] utilized physiological changes such as skin conductance to perceive the affective state changes of the user. Moreover, Janssen et al. [2009] built a personalized affective music player that inferred the user's current affective state from physiological signals and would select music based on the target affective state and current affective state [Janssen et al. 2012]. However, it is still challenging to accurately recognize emotions from biosignals. Moreover, capturing these complex signals requires specific wearable devices, which limits feasibility and efficacy.

In the field of MIR and pattern recognition, through subjective self-report measurement of musical emotion in the training phase, many works focus on applying machine learning techniques to recognize emotion expressed by music through mining massive music data based on acoustic or tag features [Ilie and Thompson 2006; Kim et al. 2010; Deng and Leung 2013]. For example, the regression model [Yang et al. 2008; Schmidt et al. 2010], Gaussian mixture model [Lu et al. 2006], and hidden Markov model (HMM) [Cheng et al. 2008] are usually applied to recognize emotion from music. Yang and Chen [2011] constructed a ranking-based emotion recognition system that used a two-dimensional valence and arousal space to represent musical emotions and calculate emotion similarities of music. Jun et al. [2010] utilized the Smith-Waterman algorithm to measure the similarity between mood sequences and recommended a list of music based on this similarity. Han et al. [2010] proposed an emotion state transition (EST) model to represent human emotions and their transitions, and gave a context-based music recommendation ontology for modeling the user's musical preferences and context. Furthermore, a personalized hybrid music recommendation system was built on listeners' interests and music contents, combining content-based, collaborative filtering, and emotion-based approaches [Lu and Tseng 2009]. However, these approaches ignore the influence of time on emotion and fail to track listeners' time-varying emotions and dynamically recommend music fitting their emotions.

Considering that it is rather complex and difficult to recognize emotion from physiological measurement, we suggest predicting the listener's current emotion based on his or her music listening list within a short-term session. Many research works on emotion prediction in time series apply regression or Markov models [Schmidt and Kim 2010], but most suffer from accuracy and reliability problems. The Markov model can infer the current emotion state depending on the previous emotion state, whereas the CRF [Lafferty et al. 2001] model can predict the current state depending not only on the previous series of states but also on successive states. The CRF model outperforms the HMM in many applications (e.g., speech recognition, segmentation). Schmidt and Kim [2011] have applied CRF to model the relationship between acoustic data and emotion parameters over time, showing good performance on emotion prediction. Consequently, we attempt to adopt the CRF model to predict the listener's time-varying emotion state based on music sequence.

3. OVERVIEW OF THE PROPOSED METHOD FOR DYNAMIC EMOTION-BASED MUSIC RECOMMENDATION

To get started, we give some assumptions and descriptions on emotion elicitation and perception in our scenario. As mentioned previously, the listener's emotional state is influenced by many factors, such as time, location, and mood, and even the listener's life experience, events in the recent past, and interpersonal relationships; thus, it is quite complex and difficult to build computational models by incorporating all of these factors for emotion elicitation, which is beyond the scope of this article. Moreover, perception of emotion from music is a complex cognitive process, and the interaction between these factors is still not well studied in the field of psychology. To simplify the conditions for emotion elicitation and perception, we only consider the following simple scenario and assumption: the listener's emotion elicitation is only affected by music listened to. In other words, the listener's emotion state will evoke to S_i when listening to a piece of music expressing emotion state S_i. For example, the listener becomes sad while listening to a sad-sounding music and the listener becomes joyful while listening to a joyful-sounding music. Hence, if the condition and circumstance of emotion elicitation are not of our making, the perception of emotion from music may not necessarily mirror what a listener is actually feeling.

The ultimate goal of this article is to build an intelligent music recommendation scheme based on the listener's dynamic emotion states. To this end, a novel method is proposed under the preceding assumptions. We suppose that the music listening list of listener $\boldsymbol{u}$ is denoted by $\boldsymbol{M_u} = \{\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_N}\}$, where $\boldsymbol{m_i}$ refers to the i-th piece of music and $\boldsymbol{N}$ is the total number of music pieces in the list. The paradigm of our method for such a music recommendation system comprises the following three steps:

—First, $\forall$ music sequence $\boldsymbol{M_u}$, the music emotion recognition module is performed to recognize emotion states of music pieces in this sequence, and thus the corresponding sequence of emotion states is produced, denoted by $\boldsymbol{S_u} = \{\boldsymbol{s_u^{t_1}}, \boldsymbol{s_u^{t_2}}, \ldots, \boldsymbol{s_u^{t_N}}\}$.
—Next, the emotion state prediction module is performed by using CRF, and thus the predicted emotion state of listener $\mathbf{u}$ can be expressed by $\hat{s} = \arg\max\ \boldsymbol{P}(\boldsymbol{s_i}|\boldsymbol{S_u})$, where $\boldsymbol{s_i}$ is a candidate emotion state ($\boldsymbol{s_i} \in \boldsymbol{S}$, and $\boldsymbol{S}$ is the collection of finite emotion states), and $\boldsymbol{P}(\boldsymbol{S_u})$ is the Markov chain on emotion states with transition probabilities. The outcome of this step is the predicted emotion state $\hat{s}$ of listener $\mathbf{u}$.
—Finally, relying on predicted emotion state $\hat{\mathbf{s}}$ of listener $\mathbf{u}$, a recommendation base $\boldsymbol{S_{u,\hat{s}}}$ is established and influence weight $\boldsymbol{\Lambda}$ of $\boldsymbol{S_{u,\hat{s}}}$ is calculated. The recommendation algorithm is performed to generate an output of the optimal ranked music list $\boldsymbol{M_u^r} = \{\boldsymbol{m_1'}, \boldsymbol{m_2'}, \ldots, \boldsymbol{m_{\mathbb{N}}'}\}$ that has the highest emotion similarities to $\boldsymbol{S_{u,\hat{s}}}$.

The overall structure of our proposed intelligent emotion-based music recommendation system is illustrated in Figure 1. There are three modules that handle the recommendation task: the music emotion recognition module, emotion state prediction module, and recommendation module. A brief explanation of each module is given next.

In the module of music emotion recognition, relying on the resonance-arousal-valence (RAV) musical emotion representation, we first extract arousal-based features, valence-based features, and resonance-based features from training music data, then select the most emotion-relevant features to represent musical emotion. Next, we construct robust regression models to recognize emotion of novel pieces of music, and the results of dimension reduction are used for computation of emotion similarity. As for the emotion state prediction module, considering that emotion can reflect a short-term psychological response of the listener, our intelligent system is able to learn the listener's emotion state from his or her historical music listening list within a listening session. Each piece

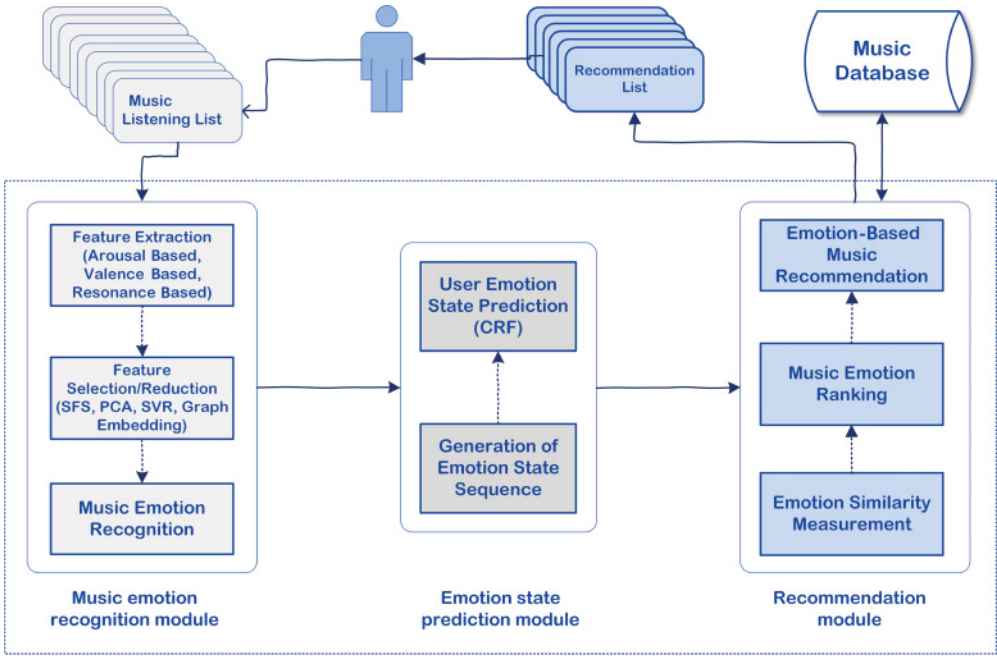

Fig. 1. Overview of the proposed intelligent dynamic emotion-based music recommendation system. There are three modules from left to right: music emotion recognition, emotion state prediction, and recommendation.

of music in the historical listening list expresses an emotion state that has influence on the listener's current emotion elicitation, and this influence is modeled as an EID process. After that, the CRF method will be used to predict the listener's current emotion state. In the recommendation module, the intelligent recommendation system recommends music based on the predicted emotion state (that the listener may want to listen to) and recommendation base. The musical emotion similarity measurement and related ranking algorithm are applied, and finally an optimal ranked music list will be produced to achieve our goal. In the following sections, each step of the proposed method is described in further detail.

4. MUSIC EMOTION MODEL

In this section, we shall introduce our approach to effectively represent musical emotion in a flexible and efficient manner. As mentioned in the review, the two-dimensional arousal-valence emotion model is insufficient to adequately represent emotion information and has shortcomings in distinguishing fine-grained emotions such as fear and anger [Han et al. 2009]. As for the three-dimensional PAD emotion model, the dimension of dominance is not suitable for indicating emotion induced by music, whether dominant or submissive. Fortunately, apart from measurement of two dimensions (arousal and valence), Bigand et al. [2005] showed that the third dimension implied an association with an emotional character measured by musical characteristics such as continuity–discontinuity and melodic–harmonic contrast. Consequently, we apply this finding to construct a novel three-dimensional representation of musical emotion. The first and the second dimensions are identical to the arousal-valence model, and the third dimension reflects particular characteristics of music (e.g., continuity–discontinuity,

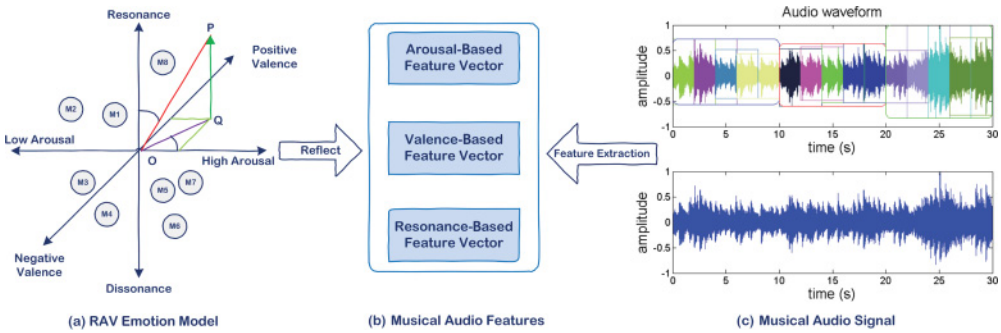

Fig. 2. Three-dimensional music emotion model by arousal, valence, and resonance in Cartesian coordinates. The expression of emotion in music is represented by feature vectors extracted from musical audio.

melodic–harmonic contrast). Considering that the third dimension shows a relation with music attributes such as resonance and dissonance, for simplicity of notation, we hence use the term *resonance* to denote the third dimension. It certainly is possible to give other more suitable names to denote this dimension. However, it is unnecessary to validate whether the given name is appropriate. As a result, we can construct a novel representation—RAV—to model musical emotion.

4.1. Music Emotion Representation

The RAV musical emotion model is illustrated by a three-dimensional Cartesian coordinate system shown in Figure 2(a), which comprises eight octants separated by three planes. The expression for musical emotion of an entire instance of music $\boldsymbol{m_i}$ is represented by $\boldsymbol{E_i} = \langle \boldsymbol{s_i}, \boldsymbol{rav_i} \rangle$, where $\boldsymbol{s_i}$ indicates a specific emotion state, and $\boldsymbol{rav_i}$ refers to its ground truth of musical emotion, denoted by $\boldsymbol{rav} = (\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\nu})$. Corresponding to the RAV musical emotion model, $\boldsymbol{\gamma}$ represents the resonance-based feature vector, $\boldsymbol{\alpha}$ refers to the arousal-based feature vector, and $\boldsymbol{\nu}$ denotes the valence-based feature vector. An emotion state is described as an octant within the RAV space. Therefore, using this two-tuple expression of musical emotion, we can combine both discrete and continuous properties for measuring and expressing emotion flexibly. The ground truth of musical emotion $\boldsymbol{rav_i}$ can be calculated in two ways: regression and dimension reduction, where regression is applied in music emotion state classification and the results of dimension reduction are used for computation of emotion similarity in recommendation.

4.1.1. Regression. The aim of regression is to estimate musical emotion $\boldsymbol{rav}$ by using musical/acoustic features and self-report emotion ratings. To this end, many regression models (e.g., MLR, PLR, SVR) can be applied. In our scenario, we regress $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\nu}$ on resonance-based features, arousal-based features, and valence-based features, respectively, and set all their ranges from –1 to 1. Thus, we can see from Figure 2(a) that each piece of music is located at a point in RAV space. Suppose that a piece of music $\boldsymbol{m_i}$ is located at point $P(\boldsymbol{r}, \boldsymbol{a}, \boldsymbol{v})$; we define the initial intensity of emotion induced by music based on its Euclidean distance $|OP| = \rho$ from zero point $O(0, 0, 0)$ to point P. Therefore, the maximum value of $|OP|$ equals 1.7321. Taking a piece of music (Beethoven's "Ode to Joy") as an example, suppose that the expression of emotion is given by $E = \langle s_1, (0.7615, 0.8158, 0.7357) \rangle$; the initial emotion intensity of this piece of music is calculated as 1.3367. Additionally, to simply describe emotion intensity without scale values, the emotion intensity can be divided into three uniform segments, referring to slight, moderate, and strong levels, respectively.

4.1.2. Dimension Reduction. Considering that there may be a large number of musical/acoustic features, we can employ dimension reduction techniques to improve efficiency and accuracy for emotion recognition. Given N pieces of music, we formulate a music-to-audio feature matrix $X_{K\times N} = [\mathbf{R} \mid \mathbf{A} \mid \mathbf{V}]^{\mathrm{T}}$, where K is the total number of audio features of music. We aim to find an optimal matrix $\mathcal{H}_{K\times d}$ to project higher-dimension audio features of music to a lower-dimension space $\mathbb{R}^d$: $Y = X^{\mathrm{T}}\mathcal{H}$, where $Y = \{y_1, \ldots, y_N\}$ with $y_i \in \mathbb{R}^d$, such that the reconstruction error summed by $\sum_{i=1}^{N} \parallel x_i - \tilde{x_i} \parallel^2$ is minimized. We concentrate on graph embedding (GE), but any other option may be applied as well. GE attempts to find the projective map that optimally preserves the neighborhood structure of the original dataset [Yan et al. 2007]. The general procedure for the formulation of GE is described as follows. Given a graph $G = \{X, W\}$ with N vertices $X = \{x_1, \ldots, x_N\}$, each vertex x_i refers to a piece of music. Let W be a symmetric adjacency matrix, where $W_{i,j}$ is the weight of the edge (x_i, x_j). This graph G preserves the topology and the relationship among the original music objects. We aim to find the optimal low-dimensional representation for graph G by minimizing weighted least square error: $\sum_{i,j} \|y_i - y_j\|^2 W_{i,j}$. Thus, the optimal projection matrix $\mathcal{H}$ can be obtained by solving the following optimization problem:

$$\begin{aligned} \underset{\mathcal{H}}{\text{minimize}} \quad & \text{trace}(\mathcal{H}^{\mathrm{T}} X L X^{\mathrm{T}} \mathcal{H}) \\ \text{subject to} \quad & \mathcal{H}^{\mathrm{T}} X D X^{\mathrm{T}} \mathcal{H} = I, \end{aligned} \tag{1}$$

where D is a diagonal matrix with $D_{i,i} = \Sigma_j W_{i,j}$, $L = D - W$ is the graph Laplacian matrix, and I is the identity matrix. Through computation of eigenvalues by the equation $XLX^{\mathrm{T}}\mathcal{H} = \Lambda XDX^{\mathrm{T}}\mathcal{H}$, the optimal projection matrix $\hat{\mathcal{H}}$ are regarded as the eigenvectors corresponding to the maximum eigenvalue. As the similarity matrix W can be formulated by different similarity criteria such as Euclidean distance and cosine similarity, other similarity criteria applied in dimension reduction methods (e.g., Laplacian eigenmap, locally preserving projections, and ISOMAP) may also be employed. To simplify the computation, we adopt similarity computation by the form of $X^{\mathrm{T}}X$ to measure all pairwise emotional similarities of music and then set an appropriate threshold for nearest neighbors to formulate adjacency matrix W.

4.2. Emotion-Relevant Feature Extraction

Corresponding to the RAV musical emotion model, three types of musical audio feature sets are extracted from musical audio signals as emotion-relevant audio features.

4.2.1. Arousal-Based Features. This feature set comprises music characteristics that tend to arouse emotion. Intensity or dynamics represents loudness or volume of a sound, which contributes mostly to predict arouse [Schubert 1999]. For instance, high intensity seems to arouse excited or joyful feelings or emotions, whereas low dynamics seems to arouse neutral or depressed emotions [Lu et al. 2006]. Energy is often used to measure dynamics or intensity in acoustics. The average energy of a given piece of music is computed by the root-mean-square (RMS) method [Lartillot and Toiviainen 2007]. Low and high energy are commonly used to express the percentage of frames contrasted with average energy. High-frequency energy measures the amount of energy above a certain cutoff frequency, which reflects the extent of brightness, whereas low-frequency energy is to the contrary [Orio 2006]. As entropy is a useful metric to measure information, relative entropy of spectrum is also utilized to measure the degree of emotion aroused. Moreover, pitch represents the fundamental frequency of a sound and has influence on activating a listener's emotion. High or low pitch can reflect different emotional expression (e.g., active or inactive). Furthermore, chroma are often utilized to describe energy distribution by 12 semitone (from A to G#) in Western music,

which is associated with arousal changes [Muller and Ewert 2010]. Thus, the musical and acoustic features mentioned previously are comprised of arousal-based features.

4.2.2. Valence-Based Features. This feature set is related to music characteristics that tend to induce valence (unpleasant–pleasant) response. Timbre is a key and comprehensive factor for expressing different emotional feelings. For example, a special timbre may inspire a valence response or unpleasant feelings [Lu et al. 2006]. The acoustic features often utilized to represent timbre are MFCCs and statistical spectrum descriptors (spectral shape and spectral contrast). Spectral shape features are usually obtained by short, overlapping frames through a Hanning window and discrete Fourier transform (DFT). Spectral shape features consist of spectral centroid, flux, flatness, and rolloff that represent the frequency less than a specific proportion of spectral distribution. Spectral contrast features describe the comparison or correlation of spectrum, such as spectral kurtosis, valley, skewness, regularity, spread, and zero-crossing rate [Schmidt et al. 2010]. Thus, timbre represented by using MFCCs and the previously mentioned spectrum characteristics is regarded as an important impact factor for reflecting valence. Rhythm also has an important influence on invoking emotional feelings of pleasure or displeasure. For example, a particular rhythmic music may inspire valence or pleasurable feelings, whereas nonrhythmic music often appears as boredom or not feeling good. Rhythm reflects a different duration over a steady background of the beat, which is expressed by characteristics [Schmidt et al. 2010] such as beat spectrum, beat onsets, onset rate, silence rate, fluctuation, event density, and tempo. To sum up, both timbre and rhythm-related musical and acoustic features are extracted to express valence-based features.

4.2.3. Resonance-Based Features. This feature set consists of music attributes such as resonance or dissonance, represented by characteristics like continuity–discontinuity and melodic–harmonic-related features. The term *melodic intervals* usually refers to separately played musical notes, described by ascending or descending musical intervals [Magalhães and de Haas 2011]. Harmony refers to simultaneously performed tones or chords, representing mixture sounds such as muddy, sharp, and smooth. The flux of six-dimensional tonal centroid vector is applied to detect harmonic changes [Muller and Ewert 2010]. A harmony chord (e.g., Chopin's "Nocturne in D flat major, Op. 27") often reflects consonance, whereas a dissonant chord (e.g., Mozart's "Adagio and Fugue in C Minor, K. 546") often reflects dissonance, and thus they have influence on invoking resonant or dissonant emotional feelings. Consonance features are often defined by the peaks of spectrum and their space of these spectral peaks, and thus we use spectral peaks and roughness to describe the music consonance attribute. Tonality describes the hierarchical pitch relationship among center keys, consisting of key and key clarity. The major mode often indicates emotional resonance, whereas the minor mode often indicates disgust or dissonant emotional feelings [Scherer and Oshinsky 1977]. Therefore, the previously mentioned melodic–harmonic-related features are comprised of resonance-based features.

4.2.4. Feature Selection. Apart from the preceding three types of feature sets, some of their statistical features (e.g., mean and variance) are also calculated and selected as part of musical emotion-relevant features. Moreover, the first differences of segments and the second differences of frames are regarded as statistical features of specific acoustic features [Picard et al. 2001]. In the procedure of feature extraction, global features of an entire music excerpt not only are extracted but also local features produced by segmentation and short-term window frame decomposition are also taken into account. Figure 3 illustrates an example of a part of global and local features extracted from musical audio.

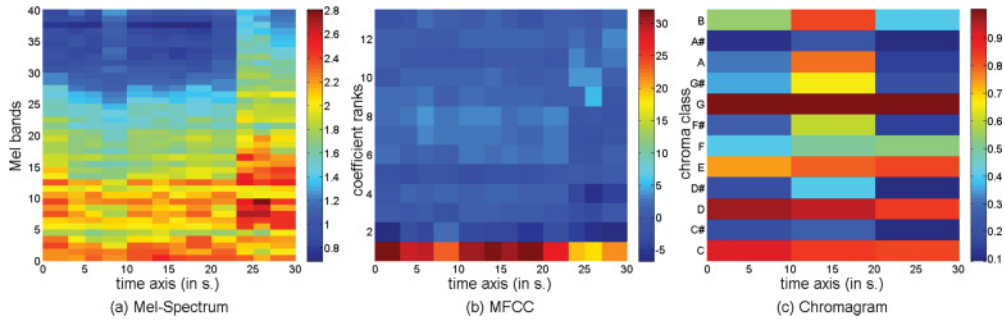

Fig. 3. Part of local features extracted from music "Ode to Joy": spectrum of segmented and frame decomposed audio waveform (a), MFCCs (b), and chromagram for segmented audio waveform (c).

Table I. The Best Emotion-Relevant Features Extracted from Musical Audio Data

Dimensions	Best Emotion-Relevant Features Selected
Arousal	rms_mean, rms_std, lowenergy, brightness_mean, entropy_mean, entropy_std, pitch_mean, pitch_std, chroma, diff_rms_mean, diff_entropy_mean
Valence	zerocross_mean, flux_mean, centroid_mean, spread_mean, mfcc, rolloff_mean, regularity_mean, skewness_mean, flatness_mean, kurtosis_mean, beatspectrum, eventdensity, td, acd, onsets_times, fluctuation_mean, diff_onsets_times, diff_fluctuation_mean
Resonance	roughness_mean, roughness_std, key, keyclarity, keystrength, tonalcentroid, mode, hcdf_mean, hcdf_std, inharmonicity, diff_roughness_mean, diff_hcdf_mean

As there are many musical and acoustic features that can be applied in statistical analysis, it is necessary to select the most emotion-relevant features to formulate expression of musical emotion. To this end, the sequential floating forward search (SFFS) [Pudil et al. 1994] can be applied because of its consistent success in feature selection. We thereby also adopt the sequential forward selection (SFS) technique to reduce redundant statistical features for acoustic features. In summary, Table I presents our selected emotion-relevant features extracted from musical audio signals.

5. EXPERIMENT 1: MUSIC EMOTION RECOGNITION

The objective of this experiment is to evaluate the effectiveness of our utilized three-dimensional RAV musical emotion model for music emotion recognition. Specifically, we relate musical audio features and their emotional impact based on this model, and we apply different dimension reduction methods (e.g., PCA, GE) and regression methods (e.g., support vector regression (SVR), Bayesian regression). Furthermore, in a user study for evaluating emotions induced by music, both subjective (self-report) and objective (those involving physiological change) evaluations have been applied. Moreover, evaluation metrics and experimental results are presented at the end of this section.

5.1. Method

5.1.1. Participants. There were a total of 65 participants (37 men and 28 women) from the Hong Kong Baptist University involved in this experiment, and all of them were undergraduate or graduate students. The average age was 24 years, with a range from 18 to 30 years. We divided these participants into two groups based on their music knowledge and background. Twenty-three participants received some level of musical education or had a music-relevant background and are referred to as the musicians

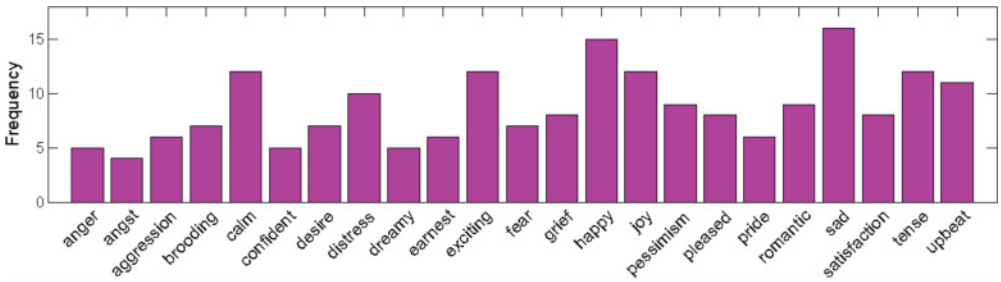

Fig. 4. The overall 23 emotion labels and their frequency distribution for selected training music excerpts.

group. The other 42 participants did not have basic musical knowledge and background and are referred to as the nonmusicians group.

5.1.2. Music Dataset. We chose 275 Western classical music excerpts from the Amazon MP3 store, Last.fm, StockMusic,[1] and the RWC classical musicdatabase.[2] The music excerpts contained concerto, sonata, symphony, and string quartet styles, and the composers of these classical music included Bach, Beethoven, Brahms, Chopin, Haydn, Mozart, Schubert, Schumann, Tchaikovsky, and Vivaldi. The selected music excerpts could illustrate a wide range of emotional feelings distributed throughout RAV musical emotion space to induce differentiated emotions from the listeners. We chose those representative music excerpts thatare able to induce emotions expected in the experiment. For instance, a selected excerpt from "Mozart's Fantasia in D Minor" was used to evoke the listener's emotion of sadness. Additionally, all music audio excerpts were converted to a uniform format, with sampling rate 22,050Hz, 16 bits, 705 bit rate, and stereo channel. Each music excerpt was truncated to 30 seconds length to avoid a long music excerpt associated with multiple emotions.

5.1.3. Experimental Design and Procedure. All participants came to the laboratory and were explained dimensions of the RAV musical emotion model: arousal, valence, and resonance. Participants were required to fill out online self-report questionnaires to collect their ratings of arousal, valence, and resonance from given music excerpts after they listened to them. Specifically, the range of the rating scale for arousal was from 0 to 10. They were told that higher arousal values corresponded to feeling emotions more intensely or activating emotions more strongly, whereas lower arousal values were associated with those emotions with less intensity. The range of the rating scale for valence was from -5 to 5. Participants were told that positive valence values represented pleasurable emotions (e.g., joy, excitement), whereas negative valence values represented displeasurable emotions (e.g., anger, sadness); 0 referred to neutral feelings of valence. Similarly, the range of the rating scale for resonance was also from -5 to 5, where positive values represented music consonance and negative values represented music dissonance. Participants were asked to rate resonance values that sound more consonant, like beautiful, euphonious in melodic-harmonic characteristics, whereas are referred to as dissonance. All participants were allowed to listen to each music excerpt more than once before making ratings. In addition, after participants listened to each given excerpt of music, they were also required to choose one label that best described the perceived emotion. The 23 predefined emotion labels were carefully selected from the OCC model and other commonly used music emotion labels [Hu et al. 2009], as shown in Figure 4. The whole rating and labeling procedure was

[1]http://www.stockmusic.net.
[2]https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-c.html.

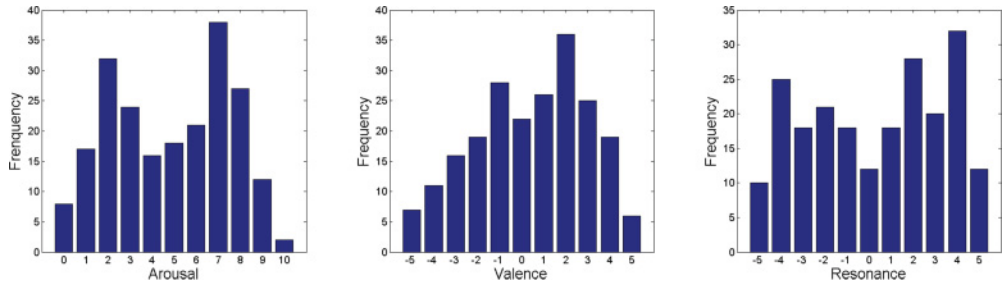

Fig. 5. Self-reported frequency distribution for arousal, valence, and resonance. The distribution for arousal ratings (low and high) and resonance ratings (consonance and dissonance) tend to the binomial distribution, whereas the distribution for valence ratings is more focused on positive values.

Table II. Comparison of the Means and Standard Deviations of Self-Report Ratings of Emotion between Musician and Nonmusician Groups

	Arousal		Valence		Resonance	
	Mean	SD	Mean	SD	Mean	SD
Nonmusicians	1.78	0.43	2.17	0.59	2.82	1.28
Musicians	1.05	0.22	1.76	0.36	2.34	0.93

that each participant listened to each given music excerpt by headphone, and then he or she gave ratings of arousal, valence, and resonance, respectively. After that, the participant would choose one emotion label that best expressed the emotion perceived. Each music excerpt was guaranteed to be listened to, rated, and labeled by at least 10 participants. Figure 4 shows the frequency distribution of emotion labels chosen by the participants.

5.1.4. Data Analysis. After all training data of music excerpts were rated and labeled, we acquired an amount of self-reported ratings from all participants. Frequency distributions of arousal, valence, and resonance for training music excerpts were collected as shown in Figure 5. We can see that there are twin peaks in the frequency distribution of arousal corresponding to low and high arousal ratings. The frequency distribution of valence is approximately normal, and the frequency distribution of resonance is rated toward consonance and dissonance. The rating quality of both groups (musicians and nonmusicians) given the same training music excerpts were compared as shown in Table II. We notice that the musician group leads to a better rating quality and reliability than the nonmusician group, because the former has smaller standard deviations (SD) than the latter. Furthermore, as we can see as well in the two groups, SD(arousal) $<$ SD(valence) $<$ SD(resonance), because rating arousal is much easier than rating valence and resonance for all participants. The rating of resonance shows the largest standard deviation (SD $= 0.93$), because consonance measured by music intervals and continuity–discontinuity characteristics is difficult to perceive and assess, especially for the nonmusicians group. Relying on statistics of emotion labels for given music excerpts from all participants, the most commonly selected emotion label for each given music excerpt is regarded as the ground truth emotion label for that music excerpt, and each music excerpt is associated with only one emotion label from the predefined emotion label list.

We assumed the real musical emotion values of arousal, valence, and resonance for each excerpt of music by calculating the weighted mean across all ratings. After that, all mean rating values of arousal, valence, and resonance were normalized to a continuous scale from -1 to 1, representing the ground truth musical emotion values. Figure 6 shows the musical emotion distribution of the dataset in three-dimensional

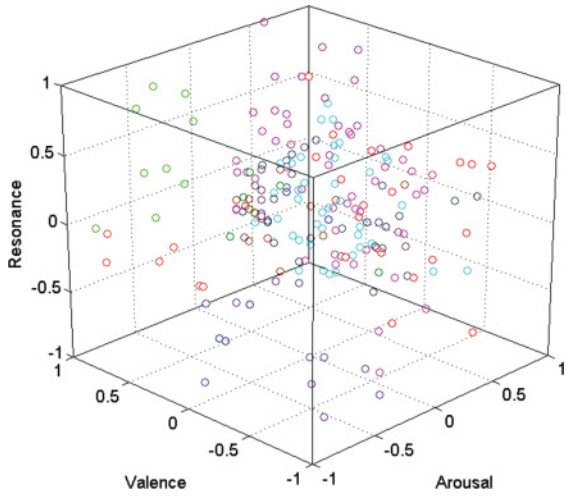

Fig. 6. Emotion distribution of selected music excerpts in terms of the RAV musical emotion model.

RAV musical emotion space. To predict musical emotions of novel excerpts of music, we constructed three distinct regressors to generate arousal, valence, and resonance values, respectively. The regressors should be trained perfectly, thus we evaluated three different regression approaches, such as SVR, sparse Bayesian regression (SBR), and variational Bayesian regression (VBR), then selected the optimal regressors to predict musical emotion values. According to previously mentioned emotion-relevant musical audio features, we utilized the MIR toolbox [Lartillot and Toiviainen 2007] and Marsyas [Tzanetakis and Cook 2000] to extract arousal-based features, valence-based features, and resonance-based features, respectively. Each 30-second music excerpt was segmented (10 seconds) and then frame decomposed (4 seconds) with half overlapping carried out. For each frame, we extracted arousal-based features, valence-based features, and resonance-based features, respectively. Moreover, some statistical values of these features, such as mean, variance, the first difference of segments, and the second difference of frames, were also calculated and selected as corresponding emotion-relevant features. Therefore, there were a total of 64 arousal-based features, 88 valence-based features, and 49 resonance-based features extracted in our experiment. We carefully chose 150 representative excerpts of music with small rating variances to train three distinct regressors. We applied different tools, such as the LIBSVM [Chang and Lin 2011] library for SVM and the relevance vector machine (RVM) [Tipping 2001] for SBR and VBR, to train different regressors.

5.1.5. Evaluation Criteria. R-squared is a commonly used statistical metric to determine how well the regression model fits the data. Moreover, the explained variance score also can be applied to evaluate the performance of different regression methods. The expression of explained variance score can be represented by

$$\Upsilon(y_t, y_p) = \frac{\mathrm{Var}(y_t) - \mathrm{Var}(y_t - y_p)}{\mathrm{Var}(y_t)}, \tag{2}$$

where y_t refers to the test value, y_p represents the predicted value, and Var denotes the variance. When $\Upsilon(y_t, y_p) = 1$, the result of regression prediction is perfect. The smaller $\Upsilon(y_t, y_p)$ is, the worse is the regression prediction produced. Furthermore, if $\Upsilon(y_t, y_p) < 0$, the performance of regression for prediction is poor.

Table III. Comparison of Coefficient Determination (R-Squared) and Explained Variance Score (Υ) for Arousal, Valence, and Resonance Obtained by Three Different Regression Approaches (SVR, SBR, VBR)

	Arousal		Valence		Resonance	
	R-Squared	Υ_α	R-Squared	Υ_v	R-Squared	Υ_γ
SVR	0.7249	0.7556	0.6119	0.6142	0.5374	0.5496
SBR	0.7381	0.7395	0.6296	0.6376	0.5456	0.5558
VBR	0.7108	0.7415	0.6328	0.6340	0.5554	0.5639

Note: Υ_α, Υ_v, Υ_γ are explained variance scores for arousal, valence, and resonance, respectively. If Υ is closer to 1, the regression performance will be better, and if $\Upsilon < 0$, the regression performance is poor.

To evaluate performance of the RAV musical emotion model for emotion recognition from music, normalized emotion similarity of two excerpts of music is given by

$$\Delta(m_i, m_j) = \frac{||e_i - e_j||_2}{||e_i||_2}, \tag{3}$$

where e_i and e_j are emotion values of music excerpt i and j, respectively. In a finite music dataset, suppose that $||e_i||_2$ is the largest, and thus Δ is confined by $0 \leq \Delta < 1$. If Δ gets close to 1, the performance of the model becomes poor. Otherwise, the performance is good. When $\Delta = 0$, the performance is perfect. To compare with other emotion models, we chose a two-dimensional arousal-valence emotion model as the benchmark, where resonance-based features were aggregated into a valence-based feature set. Given a test set, the mean and variance of Δ were calculated by different models.

5.2. Results

We now provide a breakdown of the results from our experiments and observations for music emotion regression and recognition based on the RAV musical emotion model.

5.2.1. Regression of Musical Emotion. Under the same configuration and operating conditions for music emotion regression, Table III shows the experimental results of three different regression approaches (SVR, SBR, VBR) for predicting musical emotion values. Overall, there is no significant difference in performance among these three regression approaches. However, slight discrepancies are observed. SVR is the regression benchmark. R-squared and the explained variance score Υ of SBR were slightly larger than that of the baseline. Moreover, R-squared and Υ of VBR were slightly larger than that of SBR, because VBR is a robust regression method based on SBR. Therefore, we can regard that VBR performs best for musical emotion regression. Furthermore, compared with arousal, valence, and resonance, R-squared and Υ for arousal were the largest, and R-squared and Υ for resonance were the smallest ($\Upsilon_\alpha > \Upsilon_v > \Upsilon_\gamma$). This phenomenon means that arousal is predicted the best and resonance is predicted the worst. In summary, the average regression accuracy for VBR can reach R-squared $= 0.6330$ and $\Upsilon = 0.6465$. Consequently, the results of VBR for predicting musical emotion values were applied in the following experiments. In the designing of regressors, particularly if we have more good-quality training data of musical excerpts, it may obtain better predictions of musical emotion values.

5.2.2. Music Emotion Recognition. We compared the performance of the three-dimensional RAV musical emotion model with the two-dimensional emotion model. Cross validation was carried out to assess model performance, and the averaged values of Δ with respect to two representation approaches were calculated. Table IV shows the comparison results of different emotion representations for music emotion recognition. From the table, we can see that the mean value of Δ for the RAV model ($\Delta = 0.40$) is

Table IV. Comparison of Normalized Emotional Similarity between Two Music Emotion Representations

Music Emotion Representation	Mean Δ	SD Δ
Arousal-Valence	0.49	0.33
Resonance-Arousal-Valence	0.40	0.25

Note: If Δ gets close to 1, the performance of the model is poor. Conversely, if Δ is closer to 0, the performance of the model will be better.

Table V. Average Accuracy of Music Emotion Recognition by PCA, SVR, and GE

Dimension Reduction	Subset 1	Subset 2	Mean
PCA	77.3%	80.7%	79.0%
SVR	65.3%	71.3%	68.3%
GE	81.3%	85.3%	83.3%

smaller than that of the arousal-valence model ($\Delta = 0.49$). In addition, the standard deviations (SD = 0.33) of Δ for the arousal-valence model is bigger than that of the RAV model (SD = 0.25). This is because the third dimension, "resonance," also has a positive effect on distinguishing emotion conveyed by music, whereas the two-dimensional arousal-valence representation is insufficient for distinguishing some emotions. Consequently, the RAV musical emotion representation demonstrated better performance than the arousal-valence representation in the task of music emotion recognition.

We have evaluated the performance of different dimension reduction approaches for music emotion recognition. We selected PCA as the baseline to compare with the performance of SVR and GE. The experiments were evaluated on the two subsets of different size. One subset, named "subset 1," contains 75 training music excerpts, and the other training set, named "subset 2," contains 150 training music excerpts. Relying on the ground truth emotion labels for music excerpts obtained in the user study, we adopted accuracy of music emotion recognition to evaluate the performance of dimension reduction approaches. Table V shows the accuracy of music emotion recognition by PCA, SVR, and GE. GE displays the highest music emotion recognition accuracy in both training subsets (the accuracy is greater than 80%), which means that it outperforms PCA and SVR in music emotion recognition. SVR shows the lowest music emotion recognition accuracy, as the results of regression have the largest variance. Furthermore, the number of training datasets also has influence on the accuracy of music emotion recognition. We can see that subset 2 performs better than subset 1, which means that the larger the training dataset, the better is the training effectiveness that may be achieved. Consequently, the recognition performance of music emotion may be improved as well.

6. COMPUTATIONAL MODELS FOR EMOTION IN RECOMMENDATION SYSTEMS

Affective computing aims to recognize, interpret, process, and simulate human emotions, which is capable of contributing to improve intelligence and adaptivity of recommendation systems. In the remainder of this section, we shall first apply the theory of affective computing to explain computational models for emotion, and then we present our proposed method for building such an affective recommendation system of music.

6.1. Emotion State and Transition

As described in Section 4.1, relying on the RAV musical emotion model, we can naturally generate eight octants by three planes. For simplicity of notation, we denote $+R$ for resonance, $-R$ for dissonance, $+A$ for aroused, $-A$ for unaroused, $+V$ for valence, and $-V$ for not valence. In a very fine scale, the emotion state can populate a near-continuous spectrum, and hence there can be many emotional states. A similar

Table VI. Emotion States (ES) Based on the RAV Musical Emotion Model

Emotion State	Octant	Emotion State	Octant
ES_1	+R +A +V	ES_5	−R +A +V
ES_2	+R +A −V	ES_6	−R +A −V
ES_3	+R −A −V	ES_7	−R −A −V
ES_4	+R −A +V	ES_8	−R −A +V

situation seems to occur in color vision. Color can also populate a vast spectrum, but it is nevertheless useful to start with a relatively small number of discrete descriptors. Consequently, emotion states are controlled in a fine-scale finite element space. Lu et al. [2006] applied the two-dimensional arousal-valence emotion model to classify music into four distinct states, whereby we generate more emotion states with respect to the RAV musical emotion model. For the sake of simplicity, each octant corresponding to RAV emotion space can be regarded as an emotion state. A clear definition of an emotion state is given in Definition 6.1. There are eight finite emotion states in total, which are abstract and conceptual. In other words, these emotion states do not directly indicate some specific emotion labels such as happiness or sadness. Table VI shows all eight emotion states corresponding to the RAV musical emotion model.

Definition 6.1 (*Emotion State*). An *emotion state* is a product of musical emotion changes and is modeled as a discrete state that stands for an octant in the three-dimensional RAV musical emotion space. Since there are eight octants, there are consequently eight emotion states, denoted by a collection $\mathrm{ES} = \{ES_1, \ldots, ES_8\}$.

In line with the assumption of emotion elicitation and perception from music given in Section 3, we give an example of the paradigm of emotion influence of a historical music listening list to the current emotion state. Suppose that a listener u and his or her emotion were induced by music at timestep t; we denote his or her emotion state produced at this timestep by ES_t^u, where $ES_t^u \in ES$. Therefore, relying on his or her historical listening list during a listening session, we can implicitly obtain a corresponding sequence of emotion states expressed by this historical listening list, which has different degrees of influence on elicitation of successive emotion state ES_{t+i}^u at timestep $(t + i)$. In psychology and emotion theory, emotion involves an intensity attribute that often is assessed by an emotional intensity scale (EIS) [Talarico et al. 1994]. Consequently, time-varying influence of emotion induced by music can be modeled by an EID process.

Emotion intensity tends to wax and wane with time. Picard [1997] has pointed out that emotion intensity dynamically changes during its life cycle, and the intensity of an emotion will be weak through time after it is generated. Furthermore, she proposed to computationally model the process of EID by using an inverse of the exponential function. Many other researchers have also taken this view and have applied the same or similar expressions for modeling EID for different purposes [Dias and Paiva 2005; Steunebrink et al. 2008], and have proved it to be effective in affective computing. The original expression for the relationship between the intensity of emotion e and the duration time t was given by Picard as follows:

$$\mathrm{Intensity}(e, t) = \mathrm{Intensity}(e, t_0) \times \exp(-\sigma \cdot t), \tag{4}$$

where the constant factor σ represents the intensity decay factor of given emotion e, determining the speed of EID. The limitation of the preceding expression is that it is underfitting the ground truth emotion intensities over time. To solve this problem and improve fitting of the EID model, Steunebrink et al. [2008] utilized inverse sigmoid function and introduced a parameter called *half-life*, which is the amount of time required for emotion intensity falling to one half of its initial scale. We adopt this strategy to better quantify the EID process, and the expression of intensity of musical

emotion e and duration time t is rewritten as follows:

$$\text{Intensity}(q_i, t, t_0, \mu, \sigma) = \frac{q_i}{1 + \exp(t - t_0 - \mu)\sigma}, \tag{5}$$

where q_i refers to the initial emotion intensity triggered at time t_0, which is the onset time of the emotion generated, the notation σ is the decay factor of the emotion, and the notation μ is the half-life time of that emotion. If σ is small, the emotion intensity will decay slowly. Otherwise, the emotion intensity will decay faster. Through experience in practice, we find that some negative emotions (e.g., sadness and pessimism) decay slowly, whereas some positive emotions (e.g., pleased and joy) decay rapidly. There are eight emotion states in the RAV musical emotion model, and thus there are eight corresponding distinct EID factors. As the emotion is a short-term experience that lasts seconds to several minutes at most [Ekman 1994], for the sake of simplicity, in the experiment, we can assume that the half-life μ is a small time constant (e.g., empirically from 90 seconds to 6 minutes).

Emotion is a short-term physiological response, and it moves flexibly to response to various stimuli [Juslin and Sloboda 2001]. Normally, and naturally, each emotion state is prone to remain unchanged, with a higher probability than transition to other emotion states. Furthermore, some emotion states are more likely to transit to certain emotion states, whereas others are less likely to transition. For instance, without any other external influence, the listener with an emotion of happy is more likely to transit to the emotion of calm than transition to the emotion of sad. Conversely, the emotion transition from calm to happy is not the reverse process of transition from happy to calm. To model the EST process, we apply the theory of directed graph and Markov chain to give a clear definition of EST.

Definition 6.2 (*Emotion State Transition*). An *emotion state transition* refers to an emotional change of the listener from one emotion state to another on an emotion state space. EST is represented by a directed graph $G(ES, DG)$, in which ES is the set of emotion states and DG is the set of directed edges. $ES_i \rightarrow ES_j$ denotes EST from ES_i to ES_j. $P(ES_i \rightarrow ES_j)$ represents the probability of EST from ES_i to ES_j. Given a finite number N of emotion states, EST can be represented by an EST matrix $P(ES)_{N\times N}$ whose elements are the EST probabilities $\{P(ES_i \rightarrow ES_j)\}$:

$$P(ES)_{N\times N} = \begin{pmatrix} ES_{1,1} & ES_{1,2} & \cdots & ES_{1,N} \\ ES_{2,1} & ES_{2,2} & \cdots & ES_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ ES_{N,1} & ES_{N,2} & \cdots & ES_{N,N} \end{pmatrix},$$

where elements in each row of $P(ES)_{N\times N}$ sum to one, $\sum_{j=1}^{N} ES_{i,j} = 1$, and $i = 1, \ldots, N$. $P(ES)_{N\times N}$ is a stochastic matrix. As EST from ES_i to ES_j is nonsymmetric ($ES_{i,j} \neq ES_{j,i}$), $P(ES)_{N\times N}$ is a nonsymmetric matrix.

Since there are eight finite emotion states in our scenario, we hence obtain an 8×8 EST matrix $P(ES)_{8\times 8} = \{ES_{i,j} \mid 1 \leq i, j \leq 8\}$, where $0 \leq ES_{i,j} \leq 1$. We can initialize prior probabilities of EST $P(ES_{i,j})$ to represent transition from emotion state ES_i to ES_j, where $P(ES_{i,i})$ refers to the probability of the self-loop with the same starting and ending emotion state.

6.2. Emotion State Prediction

We propose recommending music based on emotion perceived by the listener's historical listening list within a session (a short period of time). Detecting the listener's

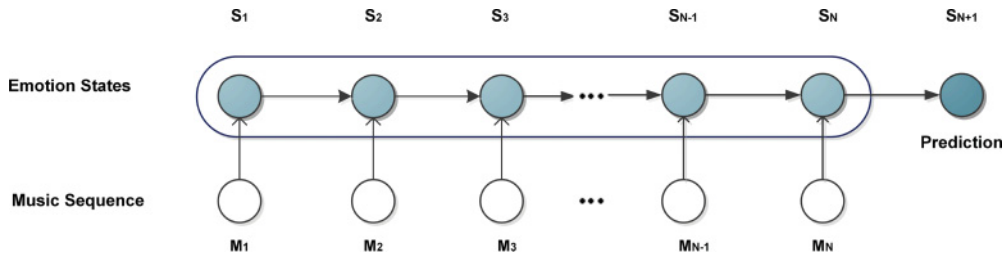

Fig. 7. Graphical model of linear-chain CRFs for a listener's emotion state prediction.

dynamic emotion is a challenge because emotion is a complex physiological response of human beings, influenced by many factors such as location, the listener's life experience, and character. However, it is rather difficult to incorporate all of these factors to predict the listener's dynamic emotion. Moreover, capturing the listener's physiological response such as heart rate, skin conductivity, and even brain wave activity is quite difficult and complex in practice. For simplicity, here we consider a simple scenario: the most possible changes of the listener's emotion state based on his or her music listening list in a session. The paradigm for this process is presented as follows: given a listening list $M_u = \{m_1, m_2, \ldots, m_N\}$ of the listener u in a listening session, a sequence of corresponding emotion states can be implicitly obtained by the RAV musical emotion model, denoted by $ES^u = \{ES_1^u, ES_2^u, \ldots, ES_N^u\}$. This sequence of emotion states has influence on elicitation of the successive emotion state ES_{N+i}^u, where i is the index of emotion states starting from ES_N^u. According to the computational model of EID described previously, the influence of each piece of music in list M_u to the current emotion state can be calculated sensibly. Consequently, we propose adopting discriminative and conditional probability theory to model dynamic emotion state prediction, computing the probabilities of the current possible emotion state of the listener given implicit observations ES^u. The influence weight of each emotion state in ES^u is given by $P(ES_{j+k}^u|ES_j^u)$, where $1 \leq j \leq N$, $1 \leq j+k \leq N$, and $\sum_{j=1}^{N} P(\hat{ES}^u|ES_j^u) = 1$.

6.2.1. Conditional Random Fields. We apply CRF [Lafferty et al. 2001] to predict the listener's emotion state based on his or her previous music listening list in a session. CRF is an undirected graphical and discriminative model that models conditional probability $P(Y|X)$ of a particular label sequence Y, given observation sequence X. Therefore, the task of a listener's dynamic emotion state prediction can be expressed as

$$\hat{ES}^u = \arg\max \; P(ES_i|ES^u), \tag{6}$$

where i is the index of emotion states corresponding to the RAV musical emotion space and $1 \leq i \leq 8$. $\boldsymbol{ES_i}$ is a candidate emotion state for prediction. Through computation of $P(ES_i|ES^u)$, the most likely current emotion state $\hat{ES}^u$ for prediction is decoded.

The illustration of emotion state prediction by CRF is shown in Figure 7. The log-linear formula of CRF in our scenario is represented by

$$P(ES_i|ES^u) = \frac{1}{Z(ES^u)} \exp\left\{\sum_{k=1}^{K} \lambda_k F_k(ES_i, ES^u)\right\}, \tag{7}$$

where k is the index of feature function, and F_k represents the feature function that corresponds to EID function. Further, feature function F_k is weighted by parameter λ_k, and K is the total number of feature functions in this sequence ES^u, where $1 \leq K \leq 8$. The constant $Z(ES^u)$ represents a normalization factor for the sequence ES^u,

partitioning this sequence and guaranteeing the distribution sum equal to one. Therefore, the expression of $Z(ES^u)$ yields the following form:

$$Z(ES^u) = \sum_{ES} \exp\left\{\sum_{k=1}^{K} \lambda_k F_k(ES_i, ES^u)\right\}. \tag{8}$$

Considering that emotion states in sequence ES^u have different degrees of influence on current emotion state ES^u_{N+1}, the EID function given in Equation (5) is utilized to represent feature function F_k. Therefore, aggregated feature function $F_k(ES_i, ES^u)$ is represented by the following formula:

$$F_k(ES_i, ES^u) = \sum_{j=1}^{T-1} f_k(ES_i, t_j) = \sum_{j=1}^{T-1} \frac{q_j}{1 + \exp(t_j - t_0 - \mu)\sigma_k}, \tag{9}$$

where T is the current timestep, and $f_k(t)$ is the k-th EID function with decay factor σ_k. The index j represents the timestep t_j of music with emotion state ES_i, and q_i represents its initial emotion intensity. Consequently, through computation of $P(ES_i|ES^u)$ by CRF, the listener's current emotion $\hat{s}$ can be predicted by choosing the emotion state with the largest probability.

6.3. Emotion-Based Music Recommendation

Our proposed music recommendation scheme is based on the predicted listener's current emotion state. Using the implicit sequence of emotion states obtained from the listener's music listening list, we not only can predict the listener's current emotion state, but we also can separate it into subsequences based on the same emotion states. The general paradigm can be outlined as follows. Given a listener's listening list in a session denoted by $M_u = \{m_1, m_2, \ldots, m_N\}$, we can implicitly obtain the corresponding emotion state sequence $ES^u = \{ES^u_1, ES^u_2, \ldots, ES^u_N\}$ through music emotion recognition. Supposing that the predicted emotion state is ES_i and that there are k pieces of music belonging to this emotion state, we thus select this subsequence denoted by $M_u(ES_i)$ as the base for recommendation. Therefore, our proposed recommendation system attempts to make a decision that selects an optimized ranked music list that has the highest emotion similarities to $M_u(ES_i)$.

6.3.1. Music Emotion Similarity. Relying on the comparison of music emotion recognition performance by using PCA, SVR, and GE given in Table V, GE is selected to handle musical audio features to map them into the RAV musical emotion space to formulate a d-dimensional musical emotion representation. Through model learning of GE described in Section 4.1.2, we can obtain the optimal projection matrix $\mathcal{H}_{K\times d}$ and then project the K-dimensional musical audio features X to a d-dimensional RAV musical emotion space by $X^{\mathrm{T}}\mathcal{H}$. Thus, the music emotion value rav can be expressed by $rav = X^{\mathrm{T}}\mathcal{H}$. Therefore, given two music excerpts m_i and m_j, the emotion similarity of these two music excerpts can be calculated by the inner product as follows:

$$\mathcal{F}(m_i, m_j) = \left(X_i^{\mathrm{T}}\mathcal{H}\right)\left(X_j^{\mathrm{T}}\mathcal{H}\right)^{\mathrm{T}} = X_i^{\mathrm{T}}\mathcal{H}\mathcal{H}^{\mathrm{T}}X_j. \tag{10}$$

Moreover, corresponding to the regression approaches to handling musical audio features, we can obtain a three-dimensional musical emotion value rav. Therefore, we can use Euclidean distance or cosine distance to compute music emotional similarity. Given N pieces of music, their emotional similarity can be expressed by an $N \times N$ emotional similarity matrix denoted by $F_{N\times N}$, where each generic entry $F(i, j)$ is nonnegative and represents the emotional similarity between music m_i and music m_j, and F is

a symmetric matrix ($F = F^{\mathrm{T}}$). This emotional similarity matrix can be used to form clusters with respect to musical emotion in recommendation.

6.3.2. Music Emotion Ranking. The ultimate goal of our proposed recommendation system is to recommend a music list that best matches music subsequence $M_u(ES_i)$ formed by the predicted emotion state ES_i. Considering that each piece of music in the subsequence $M_u(ES_i)$ has some degree of influence on current emotion state ES_{i+1}^u, we define the influence weight factor λ_{m_i} for music m_i in $M_u(ES_i)$ as follows:

$$\lambda_{m_i} = \frac{\text{Intensity}(q_i, t_{m_i}, t_0, \mu, \sigma)}{\sum_{k=1}^{n} \text{Intensity}(q_k, t_{m_k}, t_0, \mu, \sigma)}, \tag{11}$$

where t_{m_i} refers to the time when the listener produces emotion evoked by music m_i, q_i is the initial musical emotion intensity obtained by ℓ^2-norm in RAV musical emotion space, and n is the the number of music excerpts in subsequence $M_u(ES_i)$.

We attempt to obtain an optimized ranked music list that corresponds to a predicted listener's emotion state. In other words, the recommended music list should have the highest emotional similarities to subsequence $M_u(ES_i)$. Consequently, we convert this ranking problem to the following optimization problem by minimizing dissimilarity of musical emotion between $M_u(ES_i)$ and $\mathbb{N}$ candidate pieces of music, with quadratic regularization terms:

$$\begin{aligned} \underset{\Theta}{\text{minimize}} \quad & \Phi(\Theta) = \left\{ \sum_{j=1}^{\mathbb{N}} \Lambda \mathcal{F} + \frac{\eta_1}{2} \|\Theta\|_2^2 + \frac{\eta_2}{2} \|\mathcal{H}\|_2^2 \right\} \\ \text{subject to} \quad & \begin{cases} \Lambda = [\lambda_{m_1}, \lambda_{m_2}, \ldots, \lambda_{m_k}] \\ \Theta = [\theta_1, \theta_2, \ldots, \theta_{\mathbb{N}}]^{\mathrm{T}} \\ \mathcal{F} = [\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k]^{\mathrm{T}}, \ \mathcal{F}_i = \sum_{j=1}^{\mathrm{k}} \mathcal{F}(m_i', m_j), \end{cases} \end{aligned} \tag{12}$$

where $\eta_1, \eta_2 > 0$ refer to regularization parameters, k is the size of sequence $M_u(ES_i)$, and $\theta_i = m_i'$ refers to a candidate recommendation excerpt. We can apply a minimization iteration strategy to solve this optimal ranking problem. Influence weight vector Λ and regulation parameters η_1 and η_2 have significant effect on performance of our proposed emotion-based music recommendation algorithm. The details of music emotion ranking and the recommendation procedure are provided in Algorithm 1, which consists of the *initialization* process and the *iterative* process of emotion similarity minimization. Given the input sequence $M_u(ES_i)$, projection matrix $\mathcal{H}$, and regularization parameters η_1 and η_2, the algorithm outputs a optimal ranked music list Θ_M with the highest emotional similarities to sequence $M_u(ES_i)$. In the initialization process, we first compute the influence weight factor λ_{m_i} by Equation (11) and the music emotion similarity matrix by Equation (10). Then, the algorithm randomly chooses $\mathbb{N}$ candidate music objects by function RandomCandidateSet($\mathbb{N}$) to compose a candidate music recommendation list Θ_M. Next, the composed music list Θ_M is sorted by music emotion similarity matrix F. In the iteration process, the algorithm adopts minimization iteration method to minimize cumulated music emotion similarity $\Phi(\Theta_M)$ until all music objects in the database are processed. Finally, an optimal ranked music list $\Theta_M = \{m_1', m_2', \ldots, m_{\mathbb{N}}'\}$ is the output of our proposed recommendation algorithm.

7. EXPERIMENT 2: THE LISTENER'S EMOTION STATE PREDICTION BY CRF

The first experiment demonstrated that we were able to recognize emotions expressed by music. The aim of this experiment is to evaluate the effectiveness and performance of a listener's dynamic emotion state prediction by using CRF.

ALGORITHM 1: Emotion-Based Music Recommendation Algorithm

Input: A subsequence $M_u(ES_i)$ obtained from music listening sequence ES^u, the optimal projection matrix $\mathcal{H}$ obtained by GE, and regularization parameters η_1 and η_2.
Output: A ranked music list Θ_M with the highest emotional similarities to subsequence $M_u(ES_i)$ based on the predicted emotion state ES_i.

Compute influence weight factor λ_{m_i} with $i = 1, \ldots, N$, by Equation (11)
Initialization of a recommendation list with size $\mathbb{N}$: $\Theta_M \leftarrow \text{RandomCandidateSet}(\mathbb{N})$
for $j = 1 \rightarrow \text{Size}(\Theta_M)$ **do**
    Compute emotion similarity $\mathcal{F}(\mathrm{m_i}, \mathrm{m_j}) \leftarrow \mathrm{X_i^T}\mathcal{H}\mathcal{H}^{\mathrm{T}}\mathrm{X_j}$
end for
Sort the list Θ_M by music emotion similarity $\Theta_M \leftarrow \text{Sort}(\Theta_M)$
Cumulation of music emotion similarity with regularization terms:

$$\Phi(\Theta) = \left\{ \sum_{j=1}^{\mathbb{N}} \Lambda\mathcal{F} + \frac{\eta_1}{2}\|\Theta\|_2^2 + \frac{\eta_2}{2}\|\mathcal{H}\|_2^2 \right\}$$

Update candidate music recommendation list Θ_M:
for $i = 1 \rightarrow N$ **do**
    if $\Phi(\Theta_{M_i}) < \Phi(\Theta_M)$ **then**
        Replace music excerpt $\theta_{M_i} \leftarrow \theta_i$
        Update candidate list $\Theta_M \leftarrow \text{Sort}(\Theta_{M_i})$
    end if
end for

7.1. Method

7.1.1. Participants. There were 32 participants from the Hong Kong Baptist University taking part in this experiment. All of them were undergraduate or graduate students, and the average age was 25 years, with a range from 22 to 30 years. All participants needed to report their dynamic emotion intensity after they listened to the given music excerpts by self-report and physiological measurement.

7.1.2. Music Materials. The stimulus materials consist of 48 Western classical music excerpts chosen by four participants from the musician group. To guarantee that these selected music excerpts involving all eight emotion states corresponding to the RAV musical emotion model, each emotion state involved three representative training music excerpts that obviously evoked its expected emotion. Therefore, we formed our stimuli music dataset to estimate the musical EID factors and EST probability matrix in this experiment.

7.1.3. Experimental Design and Procedure. The participants came to the laboratory and participated in CRF model parameter estimation for the listener's dynamic emotion state prediction. Two key CRF model parameters were estimated from self-report and physiological data collected: EID factor σ_i and EST probabilities $P(ES)_{N\times N}$. To estimate EID factors, two estimation measurements were carried out. The first measurement was undertaken by self-report. We explained the concepts of the RAV musical emotion model and EID to all participants before the experiment. Similar to the procedure of self-report rating of RAV musical emotion in Experiment 5, we also applied online self-report questionnaires to record the participants' emotion intensity changes over time after they listened to the given music excerpts. To accurately quantify emotion intensity changes over time, we adopted the Borg Category Ratio (CR10) scale [Borg 1998] standard to measure musical emotion intensity range in scale from low to medium to high intensity. The numerical Borg CR10 scale ranges from 0 to 10, where 0 stands for nothing at all and 10 stands for extremely strong emotion intensity, which is regarded

Table VII. Self-Report Emotional Intensity Scale (EIS)

EIS	Scale Description	EIS	Scale Description
0	Nothing at all	5	Strong
0.5	Extremely weak	6	
1	Very weak	7	Very strong
2	Weak	8	
3	Moderate	9	
4	Somewhat strong	10	Extremely strong

Note: The listeners subjectively estimate their emotion intensities by giving some specific real numbers, such as 0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10.

as the maximum value. Table VII shows the details of musical EIS based on a modified Borg CR10 scale measurement in this experiment. The participants were asked to listen to each given music excerpt from the training dataset, and then we applied the experience sampling method (ESM), which required participants to stop at certain times (30-second intervals) and make reports of their emotional experience by real numbers (e.g., 0, 0.5, 1, 2, . . . , 10) in real time to record the intensities of their perceived emotions in time series. Moreover, in line with the RAV musical emotion model, there were eight different EID factors with respect to eight emotion states that were required to be estimated. The evaluation for each training music excerpt was guaranteed to be carried out by at least 10 participants to make the ratings more objective and reliable.

Considering that the limitation of self-report measurement was subjective, the second measurement undertaken by physiological evaluation was considered. In physiological measurement, the participants' physiological changes (e.g., heart rate, skin conductance) can reflect the changes in intensity of aroused emotion [Kim 2008]. For example, when people experience the emotion of anger (or excitement), they will also experience some physiological reactions such as increased heart rate and more rapid breathing. Especially considering that incorporating all physiological response variables is too complex and is beyond the scope of this work, to simplify the physiological analysis we employed heart rate variability (HRV) biofeedback to measure musical emotion intensity changes over time, which could be used as a supplement to the self-report measurement. To collect the participants' heart rates after they listened to the given music excerpts, a software application named Instant Heart Rate developed by Azumio was used. The whole procedure involved participants listening to the given music excerpt and then placing the tips of their index fingers gently on the camera lens to completely cover it. Each participant held it steady for at least 10 seconds, then the Instant Heart Rate software detected heart rate values. Each detected his or her heart rate at different timesteps (10-second intervals) repeatedly. Therefore, through the recording of heart rate produced by Instant Heart Rate given in the timeline data, these dynamic physiological changes (HRV) were applied to objectively analyze musical EID factors.

To estimate EST probability, we also designed an experiment to test our hypothesis on EST. After the participants listened to the given music excerpt with respect to a particular emotion state, they were asked to give their confidence scores of transition from the current emotion state to another emotion state. The confidence score was employed in the self-report evaluation, which took an integer value from 0 to 10, where the minimum value 0 stood for nothing at all and the maximum value 10 stood for absolute transition from the current emotion state to another emotion state with 100% probability. Each pairwise emotion state tested on self-reported measurement of EST was guaranteed to be carried out by at least 10 participants in this experiment. Therefore, through multiple transition tests on different training music excerpts associated

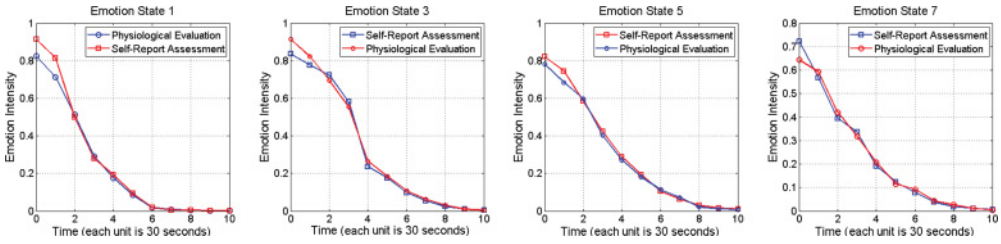

Fig. 8. Self-report and physiological measurement of four EID processes. The vertical axis refers to normalized emotion intensity, and the horizontal axis is time (each time unit is 30 seconds).

with particular emotion states, we hence obtained intuitive confidence scores from all ESTs.

7.1.4. Data Processing and Analysis. For each given training music excerpt, the collected self-report data was averaged first, then the mean value at each time point was normalized to range from 0 to 1. Finally, these normalized emotion intensities at each time point were used to analyze emotion intensity changes. At this point, the collected participants' physiological raw data (heart rate) would be preprocessed to filter out the individual differences. The baseline of heart rate data for each participant had been recorded before he or she listened to the music. The relative deviations from baseline of each participant were normalized to a continuous scale ranging from 0 to 1, then they were used to analyze emotion intensity changes. The distance between time-varying emotion intensities obtained by two measurements were calculated, and noneffective data was eliminated by setting an appropriate distance threshold. Finally, the weighted mean data of time-varying emotion intensities from self-report and physiological measurement was applied to estimate the EID factors. Figure 8 shows the normalized emotion intensity changes in time series for four selected emotion states.

We applied the maximum likelihood estimation (MLE) method to estimate EID factor $\boldsymbol{\sigma_i}$. Given a sample of N records of time-varying emotional intensities with respect to emotion state ES_i, denoted by $D = \{I_{t_1}, I_{t_2}, \ldots, I_{t_N}\}$, suppose that the density function of this sample distribution complies with EID function of ES_i given in Equation (5). Therefore, the likelihood function $L(\sigma_i|D)$ from this sample can be represented by the following formula:

$$L(\sigma_i|D) = \prod_{j=1}^{N} \text{Intensity}(t_j, \sigma_i) = \prod_{j=1}^{N} \frac{q_j}{1 + \exp(t_j - t_0 - \mu)\sigma_i}. \tag{13}$$

As log-likelihood is often used to make computation convenient, MLE usually estimates σ_i by maximizing $\ln L(\sigma_i)$. After taking a derivative with respect to σ_i and setting it to zero, we can get maximum likelihood estimator $\hat{\sigma}_{iML}$, which can be regarded as the ground truth EID factor for ES_i.

To formulate the EST probability matrix, we averaged all confidence scores of the same pairwise EST, then calculated the probability of EST from state ES_i to state ES_j by

$$P(ES_{i,j}) = C_{i,j} / \sum_j C_{i,j}, \tag{14}$$

where $C_{i,j}$ refers to the average confidence score value of the transition from emotion state ES_i to ES_j, and $P(ES_{i,j})$ is confined by $\sum_{j=1}^{8} P(ES_{i,j}) = 1$, $0 \le i, j \le 8$.

7.1.5. Evaluation Criteria. To evaluate the reliability and effectiveness of learned probability matrix P(ES) for EST obtained by Equation (14), we prepared another group of participants to give their confidence scores for EST, and these collected data were regarded as test data. Kullback Leibler (KL) divergence was adopted to measure the difference between two discrete probability distributions of the EST model, where the distribution of training data was denoted by P and the distribution of test data was denoted by Q. Thus, the expression of KL divergence of Q with respect to P is given by

$$D_{KL}(P||Q) = \sum_i P(i) \ln \left(\frac{P(i)}{Q(i)} \right), \tag{15}$$

where $P(i) = \sum_j P(ES_{i,j})$ and $Q(i) = \sum_j Q(ES_{i,j})$. By the Gibbs inequality, the result of Equation (15) is always nonnegative, $D_{KL}(P||Q) \geq 0$. If and only if $P = Q$ does $D_{KL}(P||Q) = 0$. The smaller the KL divergence, the more similar are the compared distributions P and Q. Otherwise, the greater the KL divergence, the less similar are the two distributions. Therefore, we can set a threshold of KL divergence to control their similarity. If the KL divergence is below this threshold, the distributions P and Q are considered similar (the trained EST model is valid); otherwise, they are considered dissimilar (the trained EST model is regarded as invalid).

The accuracy of listeners' dynamic emotion state prediction is evaluated in terms of average error rate (AER) of emotion state predicted by CRF. A smaller AER depicts a better prediction performance; otherwise, the prediction is worse. Moreover, a receiver operating characteristic (ROC) is used to visualize emotion state prediction performance by CRF. Music with three-level emotion intensity (slight, moderate, strong) is used to train CRF separately. A ROC space is defined by true positive rate (TPR) and false positive rate (FPR), which depicts benefits (true positive) and costs (false positive) in prediction. Additionally, the area under the curve (AUC) of ROC was calculated to summarize the ROC result. The larger the AUC, the better the prediction performance. Otherwise, the smaller AUC values correspond to worse prediction performance.

7.2. Results

We applied the MLE approach to estimate eight different EID factors $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_8\}$ by eight different subsets of the training music dataset. Considering that emotion is a short-term experience, we assume that the half-life time of emotion is 5 time units (each time unit is 30 seconds), thus $\mu = 5$ in the experiment. An example of the paradigm for EID factor calculation is given as follows. As for the learning process of decay factor for emotion state ES_4, the emotion intensity at time t_0 is normalized to initial 0.902, at time t_1 the emotion intensity is 0.7031, at time t_2 the emotion intensity is 0.5054, at time t_3 the emotion intensity is 0.2033, at time t_4 the emotion intensity is 0.055, and so on until timestep at t_{10}. Finally, we calculated decay factor σ_4 of ES_4 by MLE, and σ_4 equals 0.4112. The other seven decay factors can be calculated by following a similar computation process. Figure 9 shows eight emotion decay factors learned by MLE. The vertical axis represents normalized emotion intensity ranging from 0 to 1, and the horizontal axis represents the time unit (each time unit is 30 seconds). As we can see, emotion state ES_4 has the smallest decay factor with $\sigma_4 = 0.4112$, indicating that the intensity of ES_4 decays the most slowly. Emotion state ES_7 has the largest decay factor with $\sigma_7 = 0.9335$, indicating that the intensity of ES_7 decays the most rapidly. From the observation, we confirmed that negative emotional feelings decay slower than positive emotional feelings. For example, the emotion of sadness lasts longer than happiness.

We also evaluated the effectiveness of the learned probability matrix for EST by analyzing the KL divergence between P and Q. We calculated the results of KL divergence

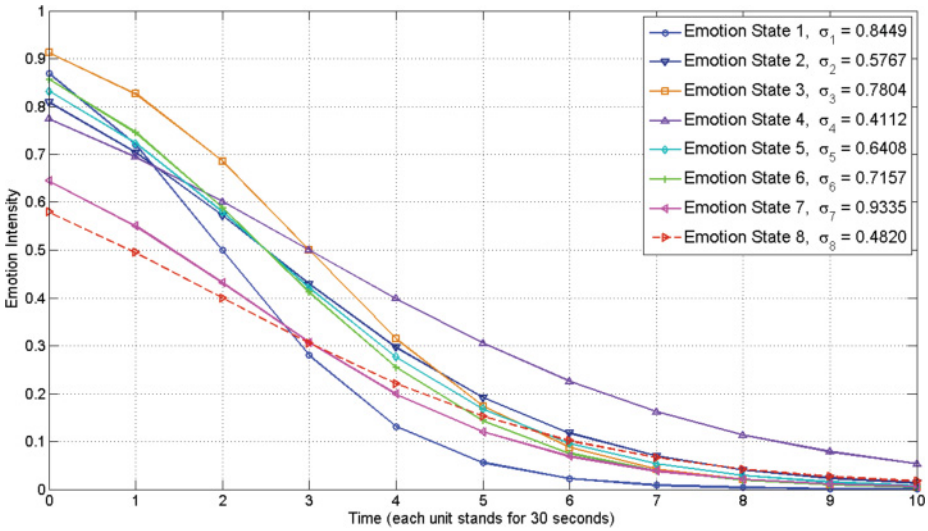

Fig. 9. Illustration of eight EID factors corresponding to emotion states in RAV space. The vertical axis refers to normalized emotion intensity, and the horizontal axis represents time.

between P and Q on increasing the size of test data, illustrated in Figure 10(a). These results tell us how well the smoothed EST performs in the learning process. We can see that at the beginning with the size of test data increase, the KL divergence between P and Q will be decreased, then it will become stable after the size of test data Q is up to 24. Furthermore, when the size of test data is 40, $D_{KL}(P||Q) = 0.153$. This means that the learned EST model is able to effectively estimate the real transition probabilities, which can be applied at CRF. After that, we calculated each EST probability; Table VIII shows results of the learned EST matrix. We can see that the diagonal element $P(ES_{i,i})$ is the largest in each row, because when a listener is in a specific emotion state ES_i, he or she most likely tends to remain in the same emotion state. The largest probability and the smallest probability in each row of the EST matrix always reflect that corresponding emotion states are contradictory in the RAV model. For instance, emotion state ES_1 is contradictory to ES_7; meanwhile, $P(ES_{1,1})$ is the largest with 0.523, and $P(ES_{1,7})$ is the smallest with 0.002. Moreover, the EST matrix illustrates that EST is nonsymmetric. For instance, the probability of EST from ES_1 to ES_5 is denoted by $P(ES_{1,5}) = 0.196$, whereas the opposite transition $P(ES_{5,1}) = 0.124$. Further, we find that EST tends to make an easier transition from a positive emotion state to a negative emotion state than the opposite transition, because of $P(ES_{1,5}) > P(ES_{5,1})$.

To predict the listener's emotion state, we constructed the CRF model to obtain the listener's emotion state with the maximum possibility, then evaluated the predicted emotion state with the listener's real emotion state. Therefore, we regarded this prediction problem as a binary (true or false) decision. The size of sequence in CRF was of critical importance to overall performance. This issue was investigated, and the results are illustrated in Figure 10(b). The vertical axis is the AER of emotion state prediction, and the horizontal axis is the size of the sequence in CRF. From the observation, with the size of sequence increase, the AER will be increased, and when the size of the sequence is up to 10, the AER will become stable, because we assume that the life cycle of each emotion state is 10 time units (each units is 30 seconds). Furthermore, we can see that the AER is less than 15%, thus the trained CRF model is effective in emotion state prediction within a certain time period for affective recommendation systems.

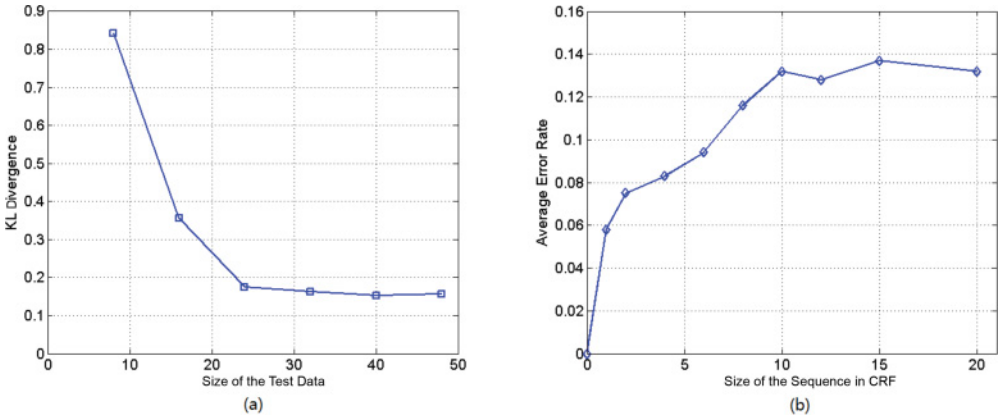

Fig. 10. (a) KL divergence for EST learning on training dataset P and test data set Q. With the size of test data increase, the KL divergence will be decreased. When the size of test data is 40, $D_{KL}(P||Q) = 0.153$. (b) AER of emotion state prediction on different sequences by CRF. With the size of sequence increase, the AER will increase gradually until the size of the sequence is up to 10, where AER = 0.132.

Table VIII. Learned Emotion State Transition Probability Matrix

$P(ES_{i,j})$	ES_1	ES_2	ES_3	ES_4	ES_5	ES_6	ES_7	ES_8
ES_1	0.523	0.075	0.005	0.129	0.196	0.027	0.002	0.043
ES_2	0.092	0.563	0.046	0.027	0.071	0.154	0.038	0.009
ES_3	0.021	0.118	0.486	0.065	0.017	0.083	0.163	0.037
ES_4	0.082	0.016	0.051	0.611	0.062	0.007	0.025	0.146
ES_5	0.124	0.028	0.014	0.058	0.647	0.037	0.019	0.073
ES_6	0.036	0.159	0.042	0.014	0.057	0.581	0.094	0.027
ES_7	0.013	0.056	0.163	0.021	0.018	0.082	0.615	0.032
ES_8	0.056	0.005	0.038	0.210	0.091	0.017	0.047	0.537

Note: $P(ES_{i,j})$ is the probability of EST from ES_i to ES_j, and the diagonal entry $P(ES_{i,i})$ represents the probability of the emotion state remaining the same. The EST probability matrix is nonsymmetric, thus $P(ES_{i,j})$ and $P(ES_{j,i})$ are different. For example, the probability of EST from ES_1 to ES_5 is calculated to $P(ES_{1,5}) = 0.196$, whereas $P(ES_{5,1}) = 0.124$.

We carried out a ROC analysis on CRF in three different groups of music data. The first data group, which we denote as dataset 1, is a set that includes all music excerpts that convey strong emotion intensity. The second data group, denoted as dataset 2, is a set that includes all music excerpts that convey moderate emotion intensity. The third data group, denoted as dataset 3, is a set that includes all music excerpts that convey slight emotion intensity. In each data group, we formed different test sequences of music excerpts. Furthermore, we still formed some test sequences of music excerpts from a different data group, denoted as dataset 4. Figure 11 shows the results of ROC analysis performed on the different datasets. The closer the AUC is to 1, the better the emotion state prediction performance. We can see that $\text{AUC}_1 > \text{AUC}_2 > \text{AUC}_3$, thus emotion state prediction conducted in dataset 1 has the best performance, whereas emotion state prediction conducted in dataset 3 has the worst performance. This is because music with strong emotion intensity, it is easier to influence the successive emotion than that of music with slight emotion intensity.

Table IX shows the overall accuracy of CRF applied to emotion state prediction over different datasets. As seen in this table, the performance of CRF depends on its sequence composition. The AER of dataset 1 is the smallest (0.1156), followed by the

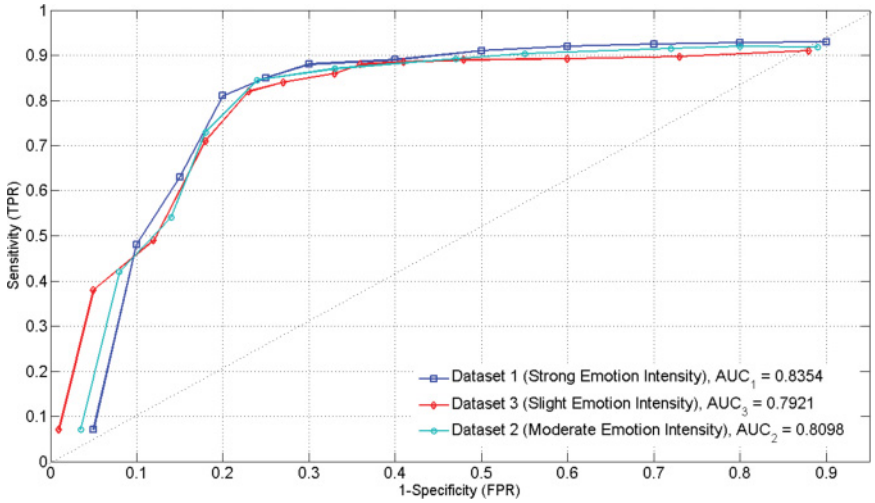

Fig. 11. ROC curve for emotion state prediction over different subdatasets. AUC_1, AUC_2, and AUC_3 refer to the AUC of datasets 1, 2, and 3, respectively. The larger the AUC, the better the emotion state prediction performance. The AUC of ROC summaries ROC result, in which $AUC_1 > AUC_2 > AUC_3$.

Table IX. Overall Evaluation Results (AER and AUC) of Listeners' Emotion State Prediction by CRF over Four Different Datasets

Evaluation	Dataset 1	Dataset 2	Dataset 3	Dataset 4
AER	0.12	0.15	0.18	0.16
AUC	0.84	0.81	0.79	0.82

Note: The smaller AER depicts a better prediction performance. The larger the AUC, the better the emotion state prediction. Dataset 1 with strong emotion intensity shows the lowest AER and the highest AUC.

AER of dataset 2, and the AER of dataset 3 is the largest (0.1821). It was suggested that music conveying strong emotion intensity leads to a greater effect than music conveying moderate or slight emotion intensity. The AER of dataset 4 is between the largest and the smallest AER. Furthermore, the AUC of dataset 1 is the largest (0.8354), which also proves that the sequence consisting of music conveys strong emotion intensity results in best prediction performance. The AUC of dataset 4 is 0.8217, which is between the largest AUC of dataset 1 and the smallest AUC of dataset 3. This indicates that usually the CRF model is effective for emotion state prediction. In summary, the overall average emotion prediction accuracy is greater than 80% in the experiment, which is satisfactory in tracking the listener's dynamic emotion state.

8. EXPERIMENT 3: EMOTION-BASED MUSIC RECOMMENDATION

The first and second experiments demonstrated that music emotion recognition and detecting the listener's emotion state was effective and consistent. The aim of emotion-based music recommendation is to recommend a ranked music list relying on the listener's emotion state. Experiment 3 evaluates our recommendation algorithm.

8.1. Evaluation Criteria

We used a top-$\mathbb{N}$ recommendation strategy to evaluate the quality and performance of our proposed emotion-based music recommendation algorithm. Thus, mean average precision (MAP) was adopted to measure the performance of emotion-based music ranking for recommendation. MAP is based on the precision at position k, denoted

Table X. Average Accuracy of Dynamic Emotion-Based Music Recommendation: (1) by Precision at Position k and (2) with Influence Weight ($\exists\Lambda$) and without Influence Weight ($\nexists\Lambda$)

Precision	$\mathbb{N} = 5$	$\mathbb{N} = 10$	$\exists\Lambda$	$\nexists\Lambda$
$P@1$	0.88	0.85	0.86	0.78
$P@2$	0.81	0.79	0.80	0.73
$P@3$	0.75	0.73	0.74	0.68
$P@4$	0.72	0.69	0.70	0.67
$P@5$	0.65	0.61	0.63	0.61
MAP	0.76	0.73	0.75	0.70

Note: The notation $\mathbb{N}$ is the size of recommendation list. With increase of $\mathbb{N}$, accumulated emotion dissimilarity of the whole list will increase, whereas the AP will decrease.

by $P@k$ and average precision (AP). *Emotional relevance*, which measures how well a recommended music excerpt meets the emotion need of the listener, is applied to judge recommendation effectiveness. Suppose that there are r_k music excerpts of emotional relevance in the top-n recommended music list; thus, we can calculate $P@k = \frac{r_k}{n}$, and $AP(q) = \frac{1}{r_q}\sum_{k=1}^{n}\{P@k \times rel(k)\}$, where $rel(k)$ is an indicator function equaling 1 if music excerpt at rank k is emotionally relevant, and otherwise zero. MAP is calculated by $\sum_{q=1}^{\mathcal{Q}} AP(q)/\mathcal{Q}$, where $\mathcal{Q}$ is the number of recommendation attempts of the case.

8.2. Results

We evaluated the performance of the emotion-based music recommendation algorithm using the dataset given in Experiment 1, then randomly split the dataset into 80 training and 160 test music excerpts. Through multiple tests of the recommendation algorithm for different instances, we obtained the following experimental results.

8.2.1. Effect of Regulation Parameter. The selection of regulation parameter has influence on performance of recommendation. To study the effect of regulation parameter η in recommendation, we randomly selected some music excerpts and calculated emotion similarity with candidate music excerpts, then investigated variations in regulation parameter estimation. Through evaluation of regulation parameter in a specific range, it was reasonable to set $\eta = 0.01$ under the RAV musical emotion space.

8.2.2. Effect of Influence Weight Vector. Each music excerpt in the listening list has a different degree of influence weight in similarity computation of our proposed recommendation algorithm, because emotion evoked by music dynamically changes over time. Therefore, we investigated the effect of influence weight and evaluated music recommendation algorithm performance depending on two ways to calculate emotion similarity: with influence weight vector (denoted by $\exists\Lambda$) and without influence weight vector (denoted by $\nexists\Lambda$). The right part of Table X gives the comparison results between them. Emotion similarity calculation with influence weight vector for music recommendation has a higher precision than that without influence weight vector. Meanwhile, MAP of similarity computation with influence weight vector is larger than that without influence weight vector, with MAP increasing by 5%.

8.2.3. Model Size Sensitivity. Recall from Section 6.3 that emotion-based music recommendation attempts to find an optimal ranked music list (size $= \mathbb{N}$) that has the highest emotional similarity to subsequence $M_u(ES_i)$. To evaluate the sensitivity of our recommendation algorithm on the value of $\mathbb{N}$, we conducted experiments using a value of $\eta = 0.01$ to compare the recommendation accuracies when $\mathbb{N} = 5$ and 10. The comparison result is given in the left part of Table X. As we can see from this table, the overall recommendation accuracy tends to decrease when we increase the value of $\mathbb{N}$.

From the observations, it shows that the smaller position k is, the higher is the AP achieved, which results from the ranking music list. We find that as the size of the ranked list increases, the accuracy will decrease. MAP for recommendation with size $\mathbb{N} = 5$ (MAP = 0.76) is higher than that with size $\mathbb{N} = 10$ (MAP = 0.73). This is because the accumulated emotion dissimilarity for the recommendation list will increase with the growing number of size. In summary, we can control the size of the recommendation list in a reasonable range to obtain the satisfied result.

9. CONCLUSION AND FUTURE WORK

We have presented several interrelated technical and empirical contributions intended to support the development of emotion-aware interactive music applications:

(1) An effective hybrid musical emotion model—RAV—was constructed, and the mapping between musical/acoustic features and their emotional impact according to this model was presented, along with robust regression methods for recognizing emotion from music. The results of Experiment 1 suggest that our musical emotion representation yields results that are competitive with those obtained by existing representations.
(2) We have presented a method for predicting a listener's changing emotional state on the basis of his or her historical music listening list in a session, which is based on the simplifying assumption that the listener's emotional state is determined only by the music recently listened to. To this end, an EID model and an EST model have been learned on the basis of data from a user study, and CRF has been used to predict the listener's emotional state. The strategy underlying the prediction process is that probabilities of different emotional states are calculated first, then the emotional state with the highest probability is chosen as the listener's predicted emotional state. The results of Experiment 2 suggest that this is a promising approach.
(3) We have presented a method, using minimization iteration, for computing a ranked list of music intended to match a particular assumed emotional state (e.g., which can be useful in a situation where it can be assumed that a user would like to hear more music that matches his or her current emotional state). The results of Experiment 3 show that the regulation parameter, the influence weight, and the size of the recommendation list all have effects on recommendation performance.

It is our hope that these contributions will support future efforts to build high-performance intelligent emotion-aware music recommender systems, as well as other music applications that take into account the emotions induced by music and listeners' desires to hear music associated with particular emotions (which may or may not correspond to their current emotional state). The illustrative reference scenario that we have discussed is one of many possible usage scenarios.

One major challenge is to take into account the fact that a listener's emotional state can be influenced by many factors other than the music that he or she has been listening to recently. In such cases, other inputs besides recently played pieces of music need to be taken into account in predicting a person's emotional state (e.g., biosignals such as those represented in an EMG or an ECG) [Healey et al. 1998; Wijnalda et al. 2005; Kim 2008; Lin et al. 2010; Janssen et al. 2012]. Our methods for the representation of emotions and the prediction of intensity decay may be applicable with some adaptation to these cases. Another assumption that often does not apply is that a listener wants to hear music that matches his or her current emotional state. To take other cases into account, a system might, for example, enable the user to specify the desired emotional character of the music explicitly, such as by pointing to a piece of music that expresses that emotion and requesting recommendations of other music that does so as well

[Kuo et al. 2005; Jun et al. 2010; Yang and Chen 2011; Deng and Leung 2012]. There are other ways in which interactive intelligent systems for presenting music can take emotion into account, some of which have probably not even been thought of yet. It is our hope that the contributions in this article will be found useful in various ways as research in this challenging area continues to progress.

ACKNOWLEDGMENTS

This work was done at the Department of Computer Science at Hong Kong Baptist University. The reviewing of the article was managed by associate editor Michael Young and coeditor-in-chief Anthony Jameson. The authors would especially like to thank chief editors Anthony Jameson and Krzysztof Gajos for their valuable advice and generous assistance with this manuscript. Additionally, the authors would like to thank several other reviewers for their helpful comments. Last, the authors would like to thank classmates and students in Hong Kong Baptist University for their participation in data acquisition and experiments.

REFERENCES

G. Adomavicius and A. Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 6, 734–749. DOI:http://dx.doi.org/10.1109/TKDE.2005.99

F. Agrafioti, D. Hatzinakos, and A. K. Anderson. 2012. ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing* 3, 1, 102–115.

D. L. Bartlett. 1996. Physiological responses to music and sound stimuli. In *Handbook of Music Psychology*, Vol. 2, D. Hodges (Ed.). Mmb Music, 343–85.

E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. 2005. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion* 19, 8, 1113–1139.

G. Borg. 1998. *Borg's Perceived Exertion and Pain Scales*. Human Kinetics.

C. C. Chang and C. J. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3, 27.

H.-T. Cheng, Y.-H. Yang, Y.-C. Lin, I.-B. Liao, and H. H. Chen. 2008. Automatic chord recognition for music classification and retrieval. In *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo*. 1505–1508.

G. L. Collier. 2007. Beyond valence and activity in the emotional connotations of music. *Psychology of Music* 35, 1, 110–131.

E. Coutinho and A. Cangelosi. 2011. Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion* 11, 4, 921.

R. T. Dean and F. Bailes. 2010. Time series analysis as a method to examine acoustical influences on real-time perception of music. *Empirical Musicology Review* 5, 4, 152–175.

J. J. Deng and C. Leung. 2012. Emotion-based music recommendation using audio features and user playlist. In *Proceedings of the 6th International Conference on New Trends in Information Science and Service Science and Data Mining (ISSDM)*. 796–801.

J. J. Deng and C. H. C. Leung. 2013. Music retrieval in joint emotion space using audio features and emotional tags. In *Advances in Multimedia Modeling*. Lecture Notes in Computer Science, Vol. 7732. Springer, 524–534.

J. Dias and A. Paiva. 2005. Feeling and reasoning: A computational model for emotional characters. In *Proceedings of the 12th Portuguese Conference on Progress in Artificial Intelligence*. 127–140. DOI:http://dx.doi.org/10.1007/11595014_13

T. Eerola, O. Lartillot, and P. Toiviainen. 2009. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proceedings of the International Conference on Music Information Retrieval*. 621–626.

T. Eerola and J. K. Vuoskoski. 2010. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39, 1, 18–49. http://pom.sagepub.com/cgi/doi/10.1177/0305735610362821

P. Ekman. 1994. Moods, emotions, and traits. In *The Nature of Emotions: Fundamental Questions*, P. Eckman and R. J. Davidson (Eds). Oxford University Press, 56–58.

J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological Science* 18, 12, 1050–1057.

P. Gebhard. 2005. ALMA: A layered model of affect. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems*. 29–36. DOI:http://dx.doi.org/10.1145/1082473.1082478

H. Gunes, B. Schuller, M. Pantic, and R. Cowie. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG '11)*. 827–834.

B. Han, S. Rho, R. B. Dannenberg, and E. Hwang. 2009. SMERS: Music emotion recognition using support vector regression.. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*. 651–656. http://dblp.uni-trier.de/db/conf/ismir/ismir2009.html#HanRDH09.

B. Han, S. Rho, S. Jun, and E. Hwang. 2010. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications* 47, 3, 433–460.

A. Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* 23, 2, 90–100.

J. Healey, R. Picard, and F. Dabek. 1998. A new affect-perceiving interface and its application to personalized music selection. In *Proceedings of the 1998 Workshop on Perceptual User Interfaces*. 4–6.

K. Hevner. 1936. Experimental studies of the elements of expression in music. *American Journal of Psychology* 48, 2, 246–268.

X. Hu, J. S. Downie, and A. F. Ehmann. 2009. Lyric text mining in music mood classification. *American Music* 183, 5,049, 2–209.

G. Ilie and W. F. Thompson. 2006. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception* 23, 4, 319–330.

C. E. Izard and C. Z. Malatesta. 1987. Perspectives on emotional development I: Differential emotions theory of early emotional development. In *Handbook of Infant Development* (2nd ed.), J. D. Osofsky (Ed.). Wiley, 494–554.

J. H. Janssen, E. L. van den Broek, and J. H. D. M. Westerink. 2009. Personalized affective music player. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*. 1–6.

J. H. Janssen, E. L. van den Broek, and J. H. D. M. Westerink. 2012. Tune in to your emotions: A robust personalized affective music player. *User Modeling and User-Adapted Interaction* 22, 3, 255–279.

S. Jun, S. Rho, and E. Hwang. 2010. Music retrieval and recommendation scheme based on varying mood sequences. *International Journal on Semantic Web and Information Systems* 6, 2, 1–16.

P. N. Juslin and P. Laukka. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research* 33, 3, 217–238.

P. N. Juslin and J. A. Sloboda. 2001. *Music and Emotion: Theory and Research*, Vol. 20. Oxford University Press. http://nursinglibrary.org/Portal/main.aspx?pageid=36&sku=80380&ProductPrice=89.5000

J. Kim. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 12, 2067–2083.

Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proceedings of the International Conference on Music Information Retrieval*. 255–266.

V. J. Konečni. 2008. Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts* 2, 2, 115.

C. L. Krumhansl. 2002. Music: A link between cognition and emotion. *Current Directions in Psychological Science* 11, 2, 45–50.

F. F. Kuo, M. F. Chiang, M. K. Shan, and S. Y. Lee. 2005. Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*. 507–510.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. 282–289. http://dl.acm.org/citation.cfm?id=645530.655813

O. Lartillot and P. Toiviainen. 2007. A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects*. 237–244.

R. W. Levenson. 1988. Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In *Social Psychophysiology and Emotion: Theory and Clinical Applications*, H. Wagner (Ed.). John Wiley and Sons, 17–42.

Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen. 2010. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering* 57, 7, 1798–1806.

C. C. Lu and V. S. Tseng. 2009. A novel method for personalized music recommendation. *Expert Systems with Applications* 36, 6, 10035–10044.

L. Lu, D. Liu, and H. J. Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1, 5–18.

K. F. MacDorman and S. O. C. C. Ho. 2007. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research* 36, 4, 281–299.

J. P. Magalhães and W. B. de Haas. 2011. Functional modelling of musical harmony: An experience report. In *Proceedings of the 16th International Conference on Functional Programming*. 156–162.

S. Marsella, J. Gratch, and P. Petta. 2010. Computational models of emotion. In *A Blueprint for Affective Computing: A Sourcebook and Manual,* K. R. Scherer, T. Banziger, and E. Roesch (Eds.). Oxford University Press, 21–46.

A. Mehrabian. 1980. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Oelgeschlager, Gunn, and Hain, Cambridge, MA.

A. Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14, 4, 261–292.

L. B. Meyer. 1956. *Emotion and Meaning in Music*. University of Chicago Press.

M. Muller and S. Ewert. 2010. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 3, 649–662.

N. Orio. 2006. *Music Retrieval: A Tutorial and Review*, Vol. 1. Now Pub.

A. Ortony, G. L. Clore, and A. Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press.

R. W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, 167–170.

R. W. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10, 1175–1191.

T. Pohle, E. Pampalk, and G. Widmer. 2005. Evaluation of frequently used audio features for classification of music into perceptual categories. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing*.

P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler. 1994. Floating search methods for feature selection with nonmonotonic criterion functions. In *Proceedings of the 12th International Conference on Pattern Recognition*. 279–283.

J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6, 1161.

K. R. Scherer. 1993. Neuroscience projections to current debates in emotion psychology. *Cognition and Emotion* 7, 1, 1–41.

K. R. Scherer. 2004. Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research* 33, 3, 239–251.

K. R. Scherer and J. S. Oshinsky. 1977. Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion* 1, 4, 331–346.

K. R. Scherer, M. R. Zentner, and A. Schacht. 2002. Emotional states generated by music: An exploratory study of music experts. *Musicae Scientiae* 5, 1, 149–171.

U. Schimmack and R. Rainer. 2002. Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion* 2, 4, 412.

E. M. Schmidt and Y. E. Kim. 2010. Prediction of time-varying musical mood distributions from audio. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*. 465–470.

E. M. Schmidt and Y. E. Kim. 2011. Modeling musical emotion dynamics with conditional random fields. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*. 777–782.

E. M. Schmidt, D. Turnbull, and Y. E. Kim. 2010. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the International Conference on Multimedia Information Retrieval*. 267–274.

E. Schubert. 1999. Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology* 51, 3, 154–165.

E. Schubert. 2004. Modeling perceived emotion with continuous musical features. *Music Perception* 21, 4, 561–585.

C. E. Seashore. 1923. Measurements on the expression of emotion in music. *Proceedings of the National Academy of Sciences of the United States of America* 9, 9, 323.

M. C. Sezgin, B. Gunsel, and G. Karabulut Kurt. 2012. Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing* 2012, 1, 16.

B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer. 2008. A formal model of emotions: Integrating qualitative and quantitative aspects. In *Proceedings of the 18th European Conference on Artificial Intelligence*. 256–260. http://dl.acm.org/citation.cfm?id=1567281.1567340

J. M. Talarico, K. S. LaBar, and D. C. Rubin. 1994. Emotional intensity predicts autobiographical memory experience. *Memory and Cognition* 32, 7, 1118–1132.

R. E. Thayer. 1989. *The Biopsychology of Mood and Arousal*. Oxford University Press.

M. E. Tipping. 2001. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* 1, 211–244.

M. Tkalcic, A. Kosir, and J. Tasic. 2011. Affective recommender systems: The role of emotions in recommender systems. In *Proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems*. 9–13.

G. Tzanetakis and P. Cook. 2000. Marsyas: A framework for audio analysis. *Organised Sound* 4, 3, 169–175.

G. D. Webster and C. G. Weir. 2005. Emotional responses to music: Interactive effects of mode, texture, and tempo. *Motivation and Emotion* 29, 1, 19–39.

G. Wijnalda, S. Pauws, F. Vignoli, and H. Stuckenschmidt. 2005. A personalized music system for motivation in sport performance. *Pervasive Computing* 4, 3, 26–32.

S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1, 40–51.

Y. H. Yang and H. H. Chen. 2011. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4, 762–774.

Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen. 2008. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 2, 448–457.

M. S. M. Yik, J. A. Russell, and L. F. Barrett. 1999. Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology* 77, 3, 600.

M. Zentner, D. Grandjean, and K. R. Scherer. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion* 8, 4, 494.

Received February 2012; revised December 2014; accepted January 2015