# Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning

**Article** · November 2014

DOI: 10.1145/2647868.2654904



**4 authors**, including:

Erheng Zhong
The Hong Kong University of Science and Technology
**32** PUBLICATIONS **1,576** CITATIONS
SEE PROFILE

Andrew Horner
The Hong Kong University of Science and Technology
**153** PUBLICATIONS **1,724** CITATIONS
SEE PROFILE

# Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning

Bin Wu[1], Erheng Zhong[1], Andrew Horner[1], Qiang Yang[1,2]
[1]Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong
[2]Noah's Ark Lab, Huawei, Hong Kong
{bwuaa,ezhong,horner,qyang}@cse.ust.hk

## ABSTRACT

Music emotion recognition, which aims to automatically recognize the affective content of a piece of music, has become one of the key components of music searching, exploring, and social networking applications. Although researchers have given more and more attention to music emotion recognition studies, the recognition performance has come to a bottleneck in recent years. One major reason is that experts' labels for music emotion are mostly song-level, while music emotion usually varies within a song. Traditional methods have considered each song as a single instance and have built models based on song-level features. However, they ignored the dynamics of music emotion and failed to capture accurate emotion-feature correlations. In this paper, we model music emotion recognition as a novel *multi-label multi-layer multi-instance multi-view learning* problem: music is formulated as a hierarchical multi-instance structure (e.g., song-segment-sentence) where multiple emotion labels correspond to at least one of the instances with multiple views of each layer. We propose a Hierarchical Music Emotion Recognition model (HMER) – a novel hierarchical Bayesian model using sentence-level music and lyrics features. It captures music emotion dynamics with a song-segment-sentence hierarchical structure. HMER also considers emotion correlations between both music segments and sentences. Experimental results show that HMER outperforms several state-of-the-art methods in terms of $F_1$ score and mean average precision.

## Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Methodologies and Techniques, Systems

## Keywords

Music Emotion Recognition; Multi-label Multi-layer Multi-instance Multi-view Learning


*MM'14,* November 03–07, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2647868.2654904.

## 1. INTRODUCTION

The famous $19^{th}$ century Russian novelist Leo Tolstoy once said, "Music is the shorthand of emotion". Music is the most direct artistic form of expressing and riding the roller-coaster of emotion[1]. Researchers have given increasing attention to this area because of many interesting and important applications of music emotion. For example, *Sensbeat*[2] is a social network where users connect with one another by sharing music and emotions, and *Stereomood*[3] is a music website which enables users to explore music based on emotion queries.

Music emotion recognition, which aims to automatically recognize the affective content of a piece of music, is the key component of the above-mentioned music emotion applications. Many previous studies have been devoted to improve the accuracy of music emotion recognition. However, the recognition accuracy is still far from satisfactory and has come to a bottleneck in recent years despite sufficient labeled data [28]. As reported in MIREX – an annual Audio Music Mood Classification Evaluation Exchange[4], the best classification accuracies have been around 68% on a five-emotion classification task since 2011, and no improvement was made in 2013.

One major reason for the performance bottleneck is that previous methods considered each song as a single instance and assumed that emotions are consistent over the entire songs [6]. However, music emotion is time-varying [19, 15]. For example, theme A of the song "I believe I can fly" by R. Kelly expresses a "reflective" emotion, while theme B conveys an "ambitious" emotion. More specifically, based on our observations from data, we propose a structure of music emotion dynamics as follows:

- Music emotion can vary greatly between music segments.
- Music emotion is mostly consistent within each segment.
- Music emotion is almost always consistent within each sentence.

Previous methods were not aware of such a structure. They extract song-level features by averaging features over the entire song, which may lead to inaccurate feature representation, and may therefore degrade recognition performance. Attempts have been made to capture the dynamics

[1]"Emotion" and "mood" are often used interchangeably. In this paper, we use "emotion".
[2]`https://itunes.apple.com/hk/app/sensbeat/id725472587?mt=8`
[3]`https://itunes.apple.com/hk/app/stereomood-tuning-my-emotions/id524634435?mt=8`
[4]http://www.music-ir.org/mirex/wiki

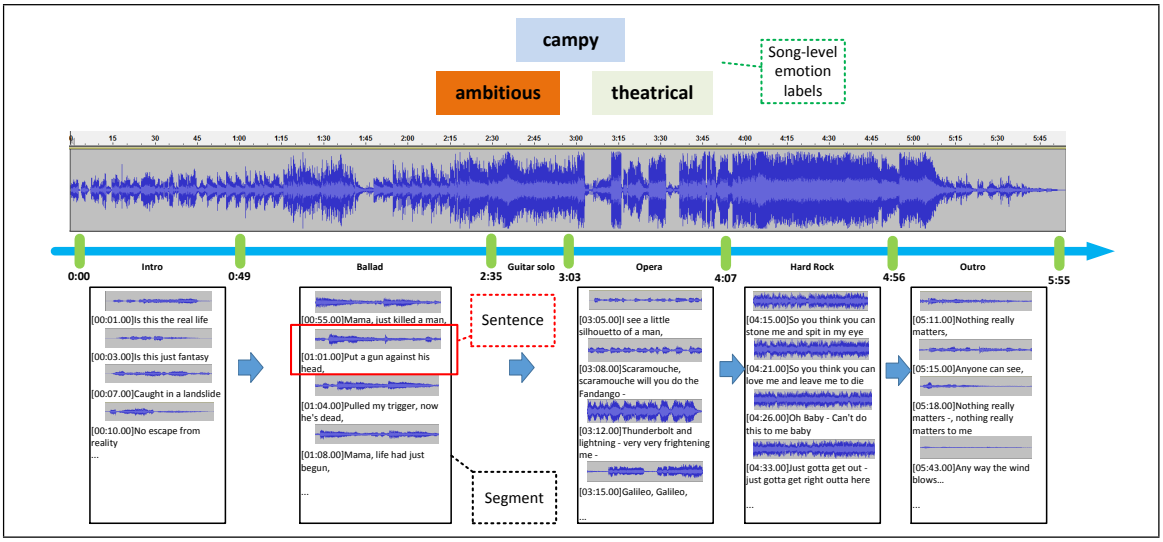


**Figure 1: Music structure of the song "Bohemian Rhapsody" by Queen. This song has three song-level emotion labels. Emotion varies between segments, but is mostly consistent within each segment, and almost always consistent within each sentence.**

of music emotion to some extent. For example, Schmidt *et al.* [20] and Lu *et al.* [15] exploited segment-level labeled data of music, assuming that emotions being conveyed within the same segment were usually consistent. Their approaches, however, is not scalable because labeling each segment requires a great deal of effort by music experts, which is difficult to come by in practice. Therefore, other label sources are desired.

Fortunately, the large number of song-level labels can be utilized without making strong assumptions like traditional single-instance methods. For example, multi-instance learning methods assume that each song is a bag of instances (e.g., segments, sentences, or music notes), where the song-level labels correspond to at least one of the instances [16]. This assumption allows different music emotions to correspond to different music segments.

Also, lyrics are another good source of labels in describing the dynamics of music emotion [9, 13]. Lyrics and music are always highly synchronized and carefully composed to match each other in terms of emotion. LRC formatted lyrics files, which are available for many songs on the Internet, provide sentence-level synchronized lyrics by indicating start and end playback timestamps for each sentence. Using this information may allow us to obtain more fine-grained and accurate music-lyrics correlations and improve feature representation.

Motivated by the structure of music emotion dynamics, as well as the two sources of labels, we propose a novel *multi-label multi-layer multi-instance multi-view learning* problem that formulates music emotion recognition as follows: "multi-label" means that a song usually has multiple emotions [21]; "multi-layer multi-instance" means that the multi-instance structure is multi-layer (e.g., song-segment-sentence), and the multiple labels correspond to at least one of the instances; "multi-view" means that multiple views (such as music and lyrics) co-exist in each instance layer (e.g., song, segment, and sentence).

To approach music emotion recognition as the novel *multi-label multi-layer multi-instance multi-view learning* problem, we propose a Hierarchical Music Emotion Recognition model (HMER) – a novel hierarchical Bayesian model which accurately captures the dynamic structure of music emotion. More specifically, HMER requires lyrics with timestamps, uses sentence-level music and lyrics features, and captures music emotion dynamics by a song-segment-sentence hierarchical structure. Also, HMER not only considers emotion correlations within a song by assuming its segments follow correlated emotion distributions, but also considers stricter emotion correlations within each segment by assuming its sentences follow the same emotion distributions. The experimental results show that HMER outperforms several state-of-the-art methods on both mean average precision and $F_1$ scores. Note that our model is aimed for offline processing so computational time is a secondary issue.

We summarize our main contributions as follows:

- We propose a novel hierarchical Bayesian model to approach music emotion recognition as a novel *multi-label multi-layer multi-instance multi-view learning* problem.
- We empirically prove the effectiveness of our proposed model on music emotion recognition using a real-world dataset.

## 2. PROBLEM DEFINITION

In this section, we first provide an example to illustrate the features and structure of music emotion dynamics. Then, we give the formal problem definition and formulation motivated by the observations from the example. We also provide a preliminary introduction to the concepts of *multi-label multi-layer multi-instance multi-view learning*.

### 2.1 An Illustrating Example

Figure 1 presents an example to illustrate the problem using the song "Bohemian Rhapsody" by Queen[5]. This song

[5]Music video of the song can be found at `http://www.youtube.com/watch?v=fJ9rUzIMcZQ`. LRC file can be down-

was labeled by AllMusic as "campy", "ambitious", and "theatrical" [1]. The full waveform is shown in the figure for illustration. The song has six segments: Intro, Ballad, Guitar solo, Opera, Hard rock, and Outro [2]. For each segment, we show part of the lyrics. Using the corresponding LRC file, we can locate the start timestamp for each sentence. We also extracted the waveform of each sentence for illustration.

Overall, we can observe from the complete waveform that the amplitude changes dramatically over the song (amplitude significantly correlates with strength of emotion [12]). More specifically, amplitude is weakest during the Intro and Outro, medium in the Ballad and Opera, and strongest in the Hard rock, indicating emotion varies greatly in different segments. For each segment, we can see that the amplitude of each sentence is of a similar level, implying that emotions are generally consistent within the same segment. Nevertheless, we can observe some special cases. For example, consider the sentence at 03:12 in Opera, the amplitude level is obviously larger than the others. This is because composers sometimes make use of contrasts to express emotion. Furthermore, amplitude level is almost always consistent within each sentence.

In summary, three assumptions can be made: 1) music emotion varies greatly between music segments; 2) music emotion is mostly consistent within each segment; 3) music emotion is almost always consistent within each sentence. Based on these assumptions, the three emotion labels "campy", "ambitious", and "theatrical", which are not similar with one another, are probably describing different segments. Intuitively, this is problematic for traditional single-instance methods which assume that multiple labels simultaneously describe song-level features.

## 2.2 Problem Formulation

Consider a set of songs $X$ of size $D$. Each song $X_d$ has a set of labels $Y_d$ of size $|Y_d|$, and $I_d$ segments. The $i^{th}$ segment $X_{di}$ has $S_d^{(i)}$ sentences. The $s^{th}$ sentence $X_{dis}$ has two perspectives/views: a music view and a lyrics view, where the corresponding music and lyrics tokens are denoted as $m_{dis}$ and $l_{dis}$, respectively. Following [31], each music token is represented as a bag of prototypes $\mathcal{C} = \{c_1, \ldots, c_C\}$ for the music corpus of size $C$ by clustering acoustic features. Each lyrics token is represented as a bag of words $\mathcal{W} = \{w_1, \ldots, w_W\}$ of size $W$ by constructing a word corpus. The objective is to predict labels for testing songs $X_{test}$ given a set of training songs and labels $\{X_{train}, Y_{train}\}$. Notations are listed in Table 1.

In this paper, we formulate music emotion recognition as a novel *multi-label multi-layer multi-instance multi-view learning* problem as follows: the multiple-labeled nature of music emotion makes it a multi-label classification problem; the fact that song-level labels correspond to at least one of the song segments and sentences fits this problem to a multi-layer multi-instance learning problem; finally, since music and lyrics are synchronized, we always have two views for each layer: from song-layer to music note layer. However, since LRC data only provides synchronization on the sentence-layer and emotions are usually consistent within a sentence as observed in Figure 1, we only consider song, segment, and sentence layers in this paper.

loaded at `http://music.baidu.com/data2/lrc/1831720/1831720.lrc`

**Table 1: Notation.**

| Notation | Definition |
|---|---|
| $D$ | Number of songs (bags). |
| $K$ | Number of topics. |
| $X_d$ | Song $d$, including music and lyrics content. |
| $Y$ | Label set. |
| $Y_d$ | Labels of song $d$. |
| $\|Y_d\|$ | Number of labels. |
| $y_{dn}$ | The $n$-th label in $Y_d$. |
| $\theta_d$ | Topic distritution of song $X_d$. |
| $\gamma$ | Dirichlet prior of $\theta_d$. |
| $\varphi$ | Label distribution for topics. |
| $\beta$ | Dirichlet prior of $\varphi$. |
| $z_{dn}$ | Topic assignments of label $y_{dn}$ in song $X_d$. |
| $I_d$ | Number of segments (instances) in song $d$. |
| $\theta_d^{(i)'}$ | Label distribution of the $i$-th segment. |
| $\alpha_d^{(i)}$ | Dirichlet prior of $\theta'_{di}$. |
| $\xi_d^{(i)}$ | Parameter for song $d$ and instance $i$ which co-decides $\alpha_d^{(i)}$. |
| $\xi$ | Fixed parameter for initializing $\xi_d^{(i)}$. |
| $\eta$ | Fixed parameter which co-decides $\alpha_d^{(i)}$. |
| $S_d^{(i)}$ | Number of sentences in the $i$-th segment of song $X_d$. |
| $v'_{dis}$ | Label assignment of the song $X_d$, $i$-th segment, $s$-th sentence. |
| $m_{dis}$ | Music token of song $X_d$, $i$-th segment, $s$-th sentence. |
| $l_{dis}$ | Lyrics token of song $X_d$, $i$-th segment, $s$-th sentence. |
| $\varphi^{(m)}$ | Music token distribution over label assignment. |
| $\varphi^{(l)}$ | Lyrics token distribution over label assignment. |
| $\beta^{(m)}$ | Dirichlet prior of $\varphi^{(m)}$. |
| $\beta^{(l)}$ | Dirichlet prior of $\varphi^{(l)}$. |
| $C$ | Number of music prototypes. |
| $W$ | Number of lyrics words. |

## 2.3 Preliminary

In this subsection, we give a preliminary introduction to *multi-label multi-layer multi-instance multi-view learning*: 1) Multi-label learning aims to classify an instance into multiple labels simultaneously; 2) Multi-view learning aims to learn from multiple perspectives/views/modals of instances to improve accuracy [27]; 3) Multi-instance learning assumes that labels are provided on bags of instances to cope with the fact that labels may correspond to different instances instead of the entire bag of instances [31].

Multi-instance learning is usually also a multi-label learning problem, i.e., multi-label multi-instance [31, 30, 17]. Multi-instance learning sometimes have multiple layers of instances, i.e., multi-layer multi-instance [8]. Nevertheless, to the best of our knowledge, multi-view learning has only been applied to bag-level (i.e., multi-view multi-instance learning) [17], but neither to instance-level nor to multi-layer instance-level in a multi-instance learning setting (i.e., (multi-layer) multi-instance multi-view learning). *Multi-label multi-layer multi-instance multi-view learning* is what we propose and focus on in this paper.

# 3. HIERARCHICAL MUSIC EMOTION RECOGNITION MODEL

We elaborate our proposed Hierarchical Music Emotion Recognition (HMER) Model in this section. We first introduce our proposed generative model of music emotion.

Then, following the generative process, parameters such as emotion distribution over music and lyrics can be trained given the emotion labels and sentence-level music-lyrics. Finally, we can predict emotions of songs given the trained parameters and observed music and lyrics content in the test set.

## 3.1 The Generative Model

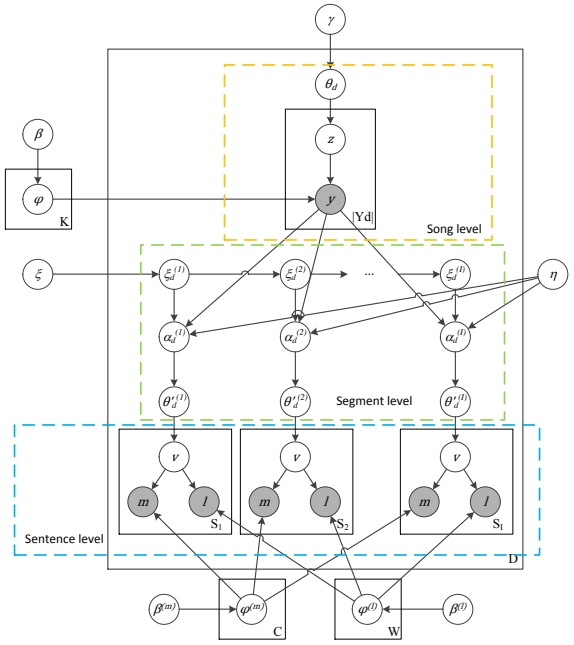


**Figure 2: Graphical model of the proposed model.**

We consider music and emotion as a top-down generative model as follows. First, composers are inspired by some musical ideas/topics perhaps based on non-musical associations. Then, when creating music and lyrics, composers design the segment structure (i.e., emotions expressed in each segment) to express the overall emotions. Next, within each segment, multiple sentences with music and lyrics content are generated by the underlying segment emotions[6]. Note that the above generative model is a probabilistic assumption for modeling purposes instead of what takes place in the real-life composition process.

More specifically, HMER represents emotions for each segment $i$ as a probability distribution $\theta_d^{(i)}$ because emotion can vary greatly between segments as observed in Figure 1. Also, since the segment emotions are correlated within the same song, we require emotion distributions of the segments to be similar using a Markov chain $\boldsymbol{\xi_d}$ ($\boldsymbol{\xi_d}$ affects how similar the Dirichlet prior between segments are, thereby how similar the distributions are). Moreover, since emotions of sentences can sometimes vary but are mostly consistent within the same segment, we assume sentences in the same segment follow the same emotion distribution, which restricts sentences to be highly similar. Finally, each sentence is restricted to a single emotion because emotion is almost always consistent within each sentence.

The generative process is described in Figure 2 as a graphical model and stated in Algorithm 1, where we refer to the topic-label generation process as the topic-label part, and the other processes as the label-token part (i.e., label-music and label-lyrics). Notations of variables in Algorithm 1 are listed in Table 1.

**Algorithm 1:** Generative process of the proposed model.

**for** *each song* $X_d$ **do**
- –Sample a topic distribution of labels $\theta_d \in Dirichlet(\gamma)$; $\theta_d$ is a $K$-dimensional Dirichlet distribution parameterized by $\gamma$.
- **for** *each label* $y_{dn}$ **do**
  - –Sample a topic assignment $z_{dn} \sim Multinominal(\theta_d)$.
  - –Sample a label from $p(y_{dn}|z_{dn}, \varphi_{z_{dn}}) = Multinominal(\varphi_{z_{dn}})$ from the topic $z_{dn}$.
- –Compute the label priors for $X_d$: $\boldsymbol{\alpha_d^{(i)}} = \{\eta \times N_1/|Y_d| + f(\xi_{d1}^{(i)}), \ldots, \eta \times N_{|Y_d|}/|Y_d| + f(\xi_{d|Y_d|}^{(i)})\}$, where $N_n$ is the number of $y_{dn}$ in $Y_d$; $\xi_{dn}^{(i)} \sim \mathcal{N}(\xi_{dn}^{(i-1)}, \sigma^2)$, and $\xi_d^{(0)} = \xi$.
- **for** *each segment* $\boldsymbol{x_{di}} \in X_d$ **do**
  - **for** *each sentence* $\boldsymbol{x_{dis}} \in \boldsymbol{x_{di}}$ **do**
    - –Sample a label assignment $v_{dis} \sim Multinomial(\theta_d^{(i)'})$.
    - –Sample a music sentence $\boldsymbol{m_{dis}}$ from $p(\boldsymbol{m}|v_{dis}, \varphi^{(m)}) = \prod_{c=1}^{C} (\varphi^m_{c,v_{dis}})^{\boldsymbol{m_c}}$; $\varphi^{(m)}_{\cdot,v_{dis}}$ is a $C$-dimensional Multinominal for the label $v_{dis}$.
    - –Sample a lyrics sentence $\boldsymbol{l_{dis}}$ from $p(\boldsymbol{l}|v_{dis}, \varphi^{(l)}) = \prod_{w=1}^{W} (\varphi^l_{w,v_{dis}})^{\boldsymbol{l_w}}$; $\varphi^{(l)}_{\cdot,v_{dis}}$ is a $W$-dimensional Multinominal for the label $v_{dis}$.

The innovations of HMER are three-folds:

1) Music and lyrics are generated simultaneously to express the same emotions on the sentence level (i.e., music and lyrics tokens $m$ and $l$ are generated by the sentence emotion $v$), while other models only assume music and lyrics convey the same emotions on the song-level [21, 18, 17];

2) The song-segment-sentence hierarchical structure captures the dynamics of music emotion, which cannot be handled by previous methods since they are one layer multi-instance models (i.e., either segments or sentences) [18, 17], while HMER has two layers (i.e., both segments and sentences).

3) Emotion correlations between music segments in the same song are captured by $\xi_d^{(i)}$ as a Markov process (i.e., emotions of music segment $\theta_d^{(i)}$ are generated for both song-level emotions $Y_d$ and the preceding segment's emotions $\theta_d^{(i-1)}$), while music segments are considered exchangeable by other multi-instance learning methods [18, 17].

Following the generative process, HMER is first trained by iteratively assigning labels to each sentence and updating topic-song distribution $\theta_d$, the label-segment distribution $\theta_d^{(i)}$ and its prior $\alpha_d^{(i)}$, the topic-label distribution $\varphi$, the music-label distribution $\varphi^{(m)}$, and the lyrics-label distribution $\varphi^{(l)}$ accordingly. Then, fixing the trained $\varphi$, $\varphi^{(m)}$, and $\varphi^{(l)}$, we can obtain label distributions of test songs by sampling label assignments through the generative process

[6]The generative process can continue until the music note level, but we ignore these in this paper to simplify the problem because emotion is usually consistent within a sentence and LRC files only provide sentence-level music-lyrics synchronization.

iteratively. Details of the training and testing processes are elaborated in the following subsections.

## 3.2 Training Process

Optimizing the segment label distribution's Dirichlet prior $\alpha_d^{(i)}$ and the other parameters jointly is computationally difficult. Therefore, we use an alternative optimization process, where $\varphi$, $\varphi^{(m)}$, and $\varphi^{(l)}$ are estimated using Gibbs sampling, and $\alpha_d^{(i)}$ is then optimized given the current model likelihood estimation.

***Fixing $\boldsymbol{\alpha}$ and maximizing the joint probability $p(X, Y, v, z)$.***

Given $\boldsymbol{\alpha}$, the topic-label part and the label-token part are decoupled. Therefore, the two parts can be estimated separately [18, 17]. The topic-label part is a standard Latent Dirichlet Allocation model [4], which can be solved by existing methods.

In the following, we estimate the label-token part using a collapsed Gibbs Sampling [7]. We denote $N_{ydi}$ as the number of times label $y$ is assigned to segment $x_{di}$, $N_{cy}^{(m)}$ as the number of times label $y$ is assigned to music prototype $c$, and $N_{wy}^{(l)}$ as the number of times label $y$ is assigned to word $w$. Then, we use the Gibbs sampling equation to update the label assignment of $\boldsymbol{x_{dis}}$ as follows:

$$\begin{aligned} &P(v_{dis} = y | \boldsymbol{x_{dis}}, \boldsymbol{x_{di-s}}, \boldsymbol{v_{di-s}}, \boldsymbol{\alpha}_d^{(i)}, \beta^{(m)}, \beta^{(l)}) \\ &\propto \frac{N_{ydi,-s} + \alpha_{dy}^{(i)}}{N_{\cdot di,-s} + \Sigma_{y'} \alpha_{dy'}^{(i)}} \\ &\times \frac{\prod_{c=1}^{C} \prod_{p=1}^{m_{disc}} (N_{cy,-s} + m_{disc} - p + \beta^{(m)})}{\prod_{r=1}^{N_{dis}} (N_{\cdot y,-s}) + N_{dis} - r + C\beta^{(m)}} \\ &\times \frac{\prod_{w=1}^{W} \prod_{q=1}^{l_{disw}} (N_{wy,-s} + l_{disw} - q + \beta^{(l)})}{\prod_{r=1}^{N_{dis}} (N_{\cdot y,-s}) + N_{dis} - r + W\beta^{(l)}}, \end{aligned} \tag{1}$$

where

$$\alpha_{dy}^{(i)} = \eta \times N_{yd} / |Y_d| + f(\xi_{dy}^{(i)}), \tag{2}$$

and $f()$ is a problem-specific transformation function that can take many forms, such as $f(x) = x$, $f(x) = e^x$, and $f(x) = log[1 + e^x]$.

***Optimizing $\boldsymbol{\xi}$.***

The joint probability of $\boldsymbol{\xi}$ and the label assignment $v$ for each sentence is:

$$\begin{aligned} &P(\boldsymbol{v}, \boldsymbol{\xi}) = p(\boldsymbol{v}|\boldsymbol{\xi}) p(\boldsymbol{\xi}) \\ &= \prod_{d=1}^{D} \prod_{i=1}^{I_d} \left\{ \frac{\Gamma(\Sigma_y \alpha_{dy}^{(i)})}{\Gamma(N_{\cdot di} + \Sigma_y \alpha_{dy}^{(i)})} \prod_{y=1}^{|Y_d|} \frac{\Gamma(\alpha_{dy}^{(i)} + N_{ydi})}{\Gamma(\alpha_{dy}^{(i)})} \right\} \\ &\times \prod_{d=1}^{D} \prod_{i=1}^{I_d} \prod_{y=1}^{|Y_d|} p(\xi_{dy}^{(i)} | \xi_{dy}^{(i-1)}), \end{aligned} \tag{3}$$

where $\xi_{dy}^{(0)} = \xi_y$, $\xi_{dy}^{(i)} \sim \mathcal{N}(\xi_{dy}^{(i-1)}, \sigma^2)$, and $\xi_y$ is a fixed parameter for each label $y$. Let $\mathcal{L} = log(P(\boldsymbol{v}, \boldsymbol{\xi}))$, for each segment $i > 0$ we have:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \xi_{dy}^{(i)}} = &\Big[ \Psi(\Sigma_{y'} \alpha_{dy'}^{(i)}) - \Psi(N_{\cdot di} + \Sigma_{y'} \alpha_{dy'}^{(i)}) \\ &+ \Psi(\alpha_{dy}^{(i)} + N_{ydi}) - \Psi(\alpha_{dy}^{(i)}) \Big] \times \frac{\partial f(\xi_{dy}^{(i)})}{\xi_{dy}^{(i)}} \\ &- \frac{\xi_{dy}^{(i)} - \xi_{dy}^{(i-1)}}{\sigma^2}, \end{aligned} \tag{4}$$

where $\Psi$ is the digamma function. Then, $\xi_{dy}^{(i)}$ is updated as:

$$\xi_{dy}^{(i)} = \xi_{dy}^{(i)} - \lambda * \frac{\partial \mathcal{L}}{\partial \xi_{dy}^{(i)}}, \tag{5}$$

where $\lambda$ is the learning rate. $f()$ is defined as $f(x) = x$ in this paper. Note that the choice of $f()$ depends on the problem and data. $f(x) = e^x$ and $f(x) = log[1 + e^x]$ can be adopted if updating $\xi_{dy}^{(i)}$ leads to negative $\alpha_{dy}^{(i)}$. After training, the posterior estimates of $\hat{\varphi}$, $\hat{\varphi}^{(m)}$, and $\hat{\varphi}^{(l)}$ are:

$$\hat{\varphi}_{yk} = \frac{N_{yk} + \beta}{N_{\cdot k} + |Y_d|\beta}, \tag{6}$$

$$\hat{\varphi}_{cy}^{(m)} = \frac{N_{cy}^{(m)} + \beta^{(m)}}{N_{\cdot y}^{(m)} + C\beta^{(m)}}, \tag{7}$$

$$\hat{\varphi}_{wy}^{(l)} = \frac{N_{wy}^{(l)} + \beta^{(l)}}{N_{\cdot y}^{(l)} + W\beta^{(l)}}, \tag{8}$$

where $\hat{\varphi}_{yk}$ is the estimated probability of label $y$ given topic $k$, $\hat{\varphi}_{cy}^{(m)}$ is the estimated probability of music prototype $c$ given label $y$, and $\hat{\varphi}_{wy}^{(l)}$ is the estimated probability of word $w$ given label $y$. The training algorithm is shown in Algorithm 2.

**Algorithm 2:** Training algorithm.

**Input**: Song set $\boldsymbol{X}$; label set $Y_d$ and segmentation for each song $X_d$; sentence-level aligned music and lyrics observation $\boldsymbol{x}_{dis} = \{\boldsymbol{m}_{dis}, \boldsymbol{l}_{dis}\}$ for the $i$-th segment and $s$-th sentence.

**Output**: Topic-label distribution $\varphi$; label-music token distribution $\varphi^{(m)}$; label-lyrics word distribution $\varphi^{(l)}$.

**for** *each song* $X_d$ **do**
- –Sample $\boldsymbol{z}$ using a Gibbs sampler [7] and obtain $\varphi$.
- **for** *each segment* $x_{di}$ **do**
  - $\xi_{dy}^{(i)} \sim \mathcal{N}(\xi_{dy}^{(i-1)}, \sigma^2)$.
  - $\alpha_{dy}^{(i)} = \eta \times N_{yd} / |Y_d| + f(\xi_{dy}^{(i)})$.
  - **repeat**
    - **for** *each observed sentence* $\{\boldsymbol{m}_{dis}, \boldsymbol{l}_{dis}\}$ **do**
      - –Sample $v_{dis}$ using Eq. 1 for several iterations.
    - –Update $\xi_{dy}^{(i)}$ using Eq. 4 and 5 until convergence.
    - –Compute $\boldsymbol{\alpha}$ using Eq. 2.
  - **until** *several iterations*;

## 3.3 Testing Process

In the testing process, since the labels for each song are no longer available, the topic-label and label-token parts are estimated in a combined way. To make the procedure more efficient, we use the fast inference method [18], where in each iteration $\boldsymbol{\alpha}_d^{(i)}$ is directly computed as:

$$\boldsymbol{\alpha}_d^{(i)} = \eta(\hat{\boldsymbol{\theta}}_d \boldsymbol{\varphi}) + f(\boldsymbol{\xi}_d^{(i)}) \tag{9}$$

Other steps are similar to the training process. The testing algorithm is listed in Algorithm 3. The point estimate of song $d$'s topic distribution $\hat{\theta}_d$ and segment $i$'s label distribution $\hat{\theta}_d^{\prime(i)}$ are as follows:

$$\hat{\theta}_{dk} = \frac{N_{dk} + \gamma}{N_{d\cdot} + K\gamma} \tag{10}$$

$$\hat{\theta}'^{(i)}_{dy} = \frac{N_{ydi} + \alpha^{(i)}_{dy}}{N_{\cdot di} + |Y|\alpha^{(i)}_{dy}} \quad (11)$$

Then, the label distribution $\mathcal{L}_d$ of song $d$ can be predicted by:

$$\mathcal{L}_d = \frac{\sum_{i=1}^{I_d} \hat{\boldsymbol{\theta}}'^{(i)}_{\boldsymbol{d}}}{I_d} \quad (12)$$

**Algorithm 3:** Testing algorithm.

**Input**: Song set $\boldsymbol{X}$; Segmentation for each song $X_d$; sentence-level aligned music and lyrics observation $\boldsymbol{x}_{dis} = \{\boldsymbol{m}_{dis}, \boldsymbol{l}_{dis}\}$ for the $i$-th segment and $s$-th sentence. Trained topic-label distribution $\varphi$; trained label-music token distribution $\varphi^{(m)}$; and trained label-lyrics word distribution $\varphi^{(l)}$.

**Output**: Topic distribution $\hat{\boldsymbol{\theta}}_d$; label distribution $\hat{\boldsymbol{\theta}}'^{(i)}_d$ over segments, and song $d$'s label distribution $\mathcal{L}_d$.

**for** *each song* $X_d$ **do**
- –Initialize $\boldsymbol{\theta}_d$.
- **repeat**
  - **for** *each segment* $x_{di}$ **do**
    - $\boldsymbol{\xi}^{(i)}_d \sim \mathcal{N}(\boldsymbol{\xi}^{(i-1)}_d, \sigma^2 \boldsymbol{I})$.
    - $\boldsymbol{\alpha}^{(i)}_d = \eta(\hat{\boldsymbol{\theta}}_d \boldsymbol{\varphi}) + f(\boldsymbol{\xi}^{(i)}_d)$.
    - **repeat**
      - **for** *each observed sentence* $\{\boldsymbol{m}_{dis}, \boldsymbol{l}_{dis}\}$ **do**
        - –Sample $v_{dis}$ using Eq. 1 for several iterations.
      - –Update $\xi^{(i)}_{dy}$ using Eq. 4 and 5 until convergence.
    - **until** *several iterations*;
- **until** *a few iterations*;

## 3.4 Computational Complexity

Consider when $\boldsymbol{\alpha}^{(i)}_d$ is fixed, the computational complexity of the topic-label part model (i.e., LDA) is $\mathcal{O}(|Y_d|K)$ for song $d$ per iteration. As for the label-token part, the complexity of updating segment $i$ is bounded by $\mathcal{O}(|Y_d|(C+W))$. Therefore, the complexity of the label-token part is $\mathcal{O}(|Y_d|I_d(C+W))$, and the complete complexity given $\boldsymbol{\alpha}^{(i)}_d$ is $\mathcal{O}(\sum_{d=1}^{D}|Y_d|[K + I_d(C+W)])$. The complexity of updating $\boldsymbol{\alpha}^{(i)}_d$ is $\mathcal{O}(\sum_{d=1}^{D} I_d|Y_d|)$. In sum, the total complexity is $\mathcal{O}(\sum_{d=1}^{D}|Y_d|[K + I_d(C+W)]) + \mathcal{O}(\sum_{d=1}^{D} I_d|Y_d|) = \mathcal{O}(\sum_{d=1}^{D}|Y_d|[K + I_d(C+W)])$. In practice, the complexity can be much lower since the size of the music prototypes and lyrics words are usually much smaller than $C$ and $W$ in each song (e.g., a 15 second music segment usually contains 3 music prototypes and 10 lyrics words)

# 4. EXPERIMENTS

We empirically answer the following questions in this section: 1) Does the proposed *multi-label multi-layer multi-instance multi-view* formulation fit the music emotion recognition problem and improve the recognition performance? 2) How do parameters such as learning rate affect the performance? 3) How can the output be interpreted for real-world applications?

To answer these questions, we first illustrate the data we used. Then, multiple state-of-the-art multi-label music recognition models and multi-view multi-label multi-instance models are introduced as baselines along with the corresponding feature extraction details. Next, we compare the music emotion recognition performance of our method to the baseline methods. We also study the influence of parameters. Furthermore, we study the effectiveness of our method by interpreting its outputs for music emotion clustering and music keywords extraction.

## 4.1 Datasets

We construct our dataset based on the All Music Guide [1], which lists thousands of English Pop songs labeled by experts with emotion words. We used these songs and labels as the ground truth. There are a total of 183 emotion labels and 5106 label-song entries. We then link the songs with the Million Song Dataset [3] using song titles and artists for acoustic feature extraction. For LRC lyrics matching, we searched the song titles and artists in Baidu[7] and downloaded all matched LRC files. The start and end time-stamps for music segments and sentences were retrieved from the Million Song Dataset and LRC files respectively. All in all, there were a total of 1493 songs and 122 labels. The average number of labels per song was 1.85, and the median number of songs per label was 10. The data is summarized in Table 2.

**Table 2: Data summary.**

| Statistics | Number |
|---|---|
| #Songs | 1493 |
| #Segments | 6055 |
| #Sentences | 31388 |
| #Labels | 122 |
| #Label-song pairs | 2767 |
| Average #labels per song | 1.85 |
| Median #songs per label | 10 |
| Standard Deviation of #songs per label | 6.6 |

## 4.2 Baseline Methods

We introduce several baseline methods for comparing the experimental results, including random guessing by top labels (Random), previous methods on multi-label music emotion classification such as binary relevance (BR), label powerset (LP), random k-labelsets (RAKEL) [21], and a state-of-the-art multi-view multi-label multi-instance method M3LDA [17]. BR, LP, and RAKEL are single-instance models while M3LDA and our proposed model HMER are multi-instance models. For M3LDA, either a segment or sentence can be regarded as an instance, leading to the two baselines M3LDA-seg and M3LDA-sen.

## 4.3 Feature Extraction

To describe the acoustic features from a music emotion perspective, we consider features such as timbre, pitch, rhythm, and loudness, which have been found to be effective in describing music emotion [28]. We extracted Mel-Frequency Cepstral Coefficient (MFCC)-like features with 12 bands, including timbre and pitch from the Million Song

[7]http://music.baidu.com

Dataset. Also, we use the maximum loudness of each segment to outline the overall loudness of the music.

For single-instance models (BR, LP, and RAKEL), we represent lyrics using term frequency-inverse document frequency (TF*IDF) [25]. Since the size of the lyrics dictionary is too large (more than 10,000), we adopted the most frequent 1000 words. For acoustic features, we take the mean and Standard Deviation (STD) every five seconds (i.e., the texture window), and take the mean and STD again for the complete song [23]. For rhythm, we use the overall tempo (beats per minute) and take the STD of beat intervals to represent the tempo regularity [11]. Finally, aggregating all four features, we have a 102-dimension acoustic feature vector. Finally, there are 1102 dimensions for feature vectors in single-instance models.

For multi-instance models (M3LDA and HMER), we simply count the word frequency for lyrics features. The size of lyrics dictionary is 17851. For acoustic features, we first extract timbre, pitch, loudness, and rhythm features every five seconds. Then, to construct the music token dictionary, we conducted a 1000-prototypes k-means clustering across the whole dataset. Then, each instance (segment for M3LDA-seg and sentence for M3LDA-sen and HMER) is represented by a 1000-dimension vector, where each dimension is the number of times the prototypes appear in the instance. For the sake of fairness, we also test single-instance models with prototype feature representation to be consistent with multi-instance methods, which are coined BR-Pro, LP-Pro, and RAKEL-Pro, respectively.

Lyrics features were used differently in different models because using all lyrics would result in too many dimensions for single-instance models; while in multi-instance models, a sentence is much shorter than a song, so using only 1000 lyrics would reduce most sentences to zero words.

## 4.4 Experimental Setup

Experiments on BR, LP, and RAKEL were conducted on MEKA[8] and MULAN [22], using random tree as the base classifier with recommended parameters.

Both M3LDA and HMER are iterative methods. The number of training and testing iterations were set as 500; the number of topics $K$ was set to 10 motivated by psychology studies such as Hevner's 8-cluster model [28]; $\eta$ was 250 [17]; $\gamma$ was set to 0.1; $\xi_y$ was set to 0.1 for each label $y$; Dirichlet priors of topic-label $\beta$, label-music $\beta_m$, and label-lyrics $\beta_l$ were all set to 0.01. For M3LDA, the weights of music and lyrics were both set to 0.5. For HMER, $\boldsymbol{\xi}$ was updated every 200 iterations; learning rate was set to 0.05; the threshold of convergence in updating $\boldsymbol{\xi}$ was 0.1.

We examine the performance of HMER both with and without updating $\boldsymbol{\xi}$, namely "HMER" and "HMER-$\xi$". For "HMER", we set $\xi_d^{(i)} = \xi_y$ and kept it stable during the whole training and testing process.

Note that all experiments were conducted using out-of-sample manner.

## 4.5 Music Emotion Recognition Results

We evaluated the performance using two standard metrics: $F_1$ score and mean average precision at ten (MAP@10). $F_1$ score is a binary based metric commonly used in evaluating a classifier, which is defined as follows:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{13}$$

In this paper, we return ten predicted emotions to calculate $F_1$. Mean average precision at 10 (MAP@10) is a ranking metric commonly used in evaluating recommender systems, which is defined as:

$$MAP@10 = \frac{\sum_{d=1}^{D} \left[ \frac{\sum_{k=1}^{10} P(k)}{10} \right]}{D} \tag{14}$$

where $P(k) = \frac{\#Correct}{k}$, and $D$ is the number of songs. In music emotion recognition, $F_1$ score shows how accurate songs are classified into each emotion label, while MAP@10 shows how well the top ten relevant songs are ranked for each emotion label.

The ten-run-average results of $F_1$ score and MAP@10 are listed in Table 3 along with standard deviations. The best performances are highlighted in bold-font. Statistical tests (Friedman test followed by Nemenyi post-hoc test at 5 percent significance level) are also listed in Table 3 to show the significance of performance difference, where $\{a, b\} \succ \{c, d\}$ indicates that both methods $a$ and $b$ perform significantly differently with methods $c$ and $d$.

Overall, single-instance methods such as BR, LP, and RAKEL did not perform well, with the best $F_1$ score at 0.11 by RAKEL and MAP@10 at 0.09 by LP. Multi-instance methods achieved significantly better performance, with at least 0.150 for $F_1$ score and 0.184 for MAP@10. The results show the benefits in assuming music emotions to be different for each instance, and the fitness of multi-instance models to music emotion recognition.

HMER-$\xi$ achieved the best performance, by 49% $F_1$ score and 326% MAP@10 improvement compared to the best-existing multi-label music emotion recognition baseline RAKEL, and 7.9% $F_1$ score and 8.0% MAP@10 improvement over M3LDA. HMER-$\xi$ also outperformed HMER by 5.1% on $F_1$ score and 3.6% on MAP@10, which shows that the optimization of $\xi$ captures the emotion correlation between music segments. Moreover, isolating the benefits of updating $\xi$, HMER still outperformed M3LDA by 2.6% and 4.2% on $F_1$ score and MAP@10 respectively, which clearly indicates the advantages of the *multi-layer multi-instance multi-view* formulation.

**Table 3: Performance comparison in terms of F1 Score and MAP@10 (higher is better).**

| Methods | $F_1$ Score | MAP@10 |
|---|---|---|
| Random | 0.002 | 0.006 |
| BR-Pro | 0.045±0.003 | 0.047±0.01 |
| LP-Pro | 0.03±0.003 | 0.054±0.006 |
| RAKEL-Pro | 0.035±0.004 | 0.05±0.001 |
| BR | 0.071±0.001 | 0.054±0.007 |
| LP | 0.094±0.001 | 0.09±0.001 |
| RAKEL | 0.110±0.008 | 0.087±0.002 |
| M3LDA-sen | 0.152±0.005 | 0.184±0.004 |
| M3LDA-seg | 0.150±0.002 | 0.213±0.004 |
| HMER | 0.156±0.004 | 0.222±0.001 |
| HMER-$\xi$ | **0.164±0.003** | **0.230±0.002** |

$\boldsymbol{F_1}$: {HMER-$\xi$} $\succ$ All others; {HMER} $\succ$ All others except for {M3LDA-sen}. **MAP@10**: {HMER-$\xi$} $\succ$ All others; {HMER} $\succ$ All others except for {M3LDA-seg}.

*Discussion.*

BR performed the worst because it simply transforms multi-label classification tasks into multiple single-label

[8]http://meka.sourceforge.net/

tasks, without considering label correlations. LP assumes the label powerset as a set of single labels, which considers label correlations and therefore performed better than BR. RAKEL ensembles multiple LP weak learners and further improved the performance. However, since these single-instance methods built models on song-level features, they failed to capture the music emotion dynamics structure. As for features representation, prototype representation was too sparse for single-instance models and led to the worst performance.

Both M3LDAs (i.e., M3LDA-seg and M3LDA sen) and HMERs (i.e., HMER-$\xi$ and HMER) can capture the multi-view multi-instance nature of music emotion dynamics. However, M3LDAs cannot deal with the *multi-layer multi-instance multi-view* property of music emotion and did not perform as well as HMERs due to two main reasons.

First, on one hand, M3LDAs are single-layer multi-instance models, which only consider either segments or sentences as instances (corresponding to M3LDA-seg and M3LDA-sen respectively). Also, M3LDAs only assume instances to follow the same emotion distribution, and emotion within each instance to be consistent; while HMERs assume segments to follow different yet correlated emotion distributions, sentences in the same segment to follow the same emotion distribution, and each sentence to have a consistent emotion. Apparently, compared to M3LDAs, HMERs highly match the observations that music emotions are usually different in different segments, mostly consistent within each segment, and almost always consistent within each sentence, which resulted in superior performance.

Second, since HMERs are multi-layer multi-view models, they can capture sentence-level music-lyrics correlations. Nevertheless, being bag-level multi-view models, M3LDAs can only capture song-level music-lyrics correlations, which are less accurate compared to HMERs. Such a difference also contributed to HMERs' better performance[9].

*Computational Time.*

The experiments were all conducted on a 64-bit computer with a 3.3GB CPU and a 4GB memory. Overall, single-instance methods such as BR, LP, and RAKEL finished in a very short time. For example, RAKEL spent 3 seconds for the whole training and testing process. Single-instance methods have lower computational complexity and smaller number of instances because they are song-level methods. The computational time of M3LDA-sen and HMERs are similar, with less than 1.5 hours for the whole process. M3LDA-seg spent half of the time compared to M3LDA-sen and HMERs because it is a segment-level method where the number of instances is much smaller. For prediction time, BR, LP, and RAKEL took 0.02, 4e-4, and 6e-4 seconds for testing a song, while M3LDA-sen and HMERs took 0.9 seconds, and M3LDA-seg took 0.5 seconds.

## 4.6 Parameter Study

In this section, we examine how parameters affect the performance of HMER. Three parameters ($\eta$ which decides $\alpha_d^{(i)}$, number of topics $K$, and the learning rate $\lambda$) were studied for both $F_1$ score and mean average precision at ten (MAP@10). The results are shown in Figure 3.

[9]Note that the performance may further improve using more advanced and sophisticated feature representation and selection as well as multi-modal fusion strategies.

According to Figure 3(a) and 3(b), when $\eta$ is between 200 to 300, both $F_1$ score and mean average precision achieve their best. When $\eta$ is greater than 400, the performance starts to decrease. The larger $\eta$ is, the larger $\alpha_d^{(i)}$ tends to be, and the more labels are assumed in each instance. Therefore, this result implies that the assumption of fewer emotions existing in a music segment is more reasonable.

Figure 3(c) and 3(d) show that when $K = 5$, both $F_1$ score and mean average precision achieve their best, which means that the best number of music emotion clusters is about five. For learning rate $\lambda$, the influence on both $F_1$ score and MAP@10 are smaller compared to $\eta$ and $K$ as shown in Figure 3(e) and 3(f), where $F_1$ score is in [0.165, 0.170] and the mean average precision is in [0.225, 0.235].

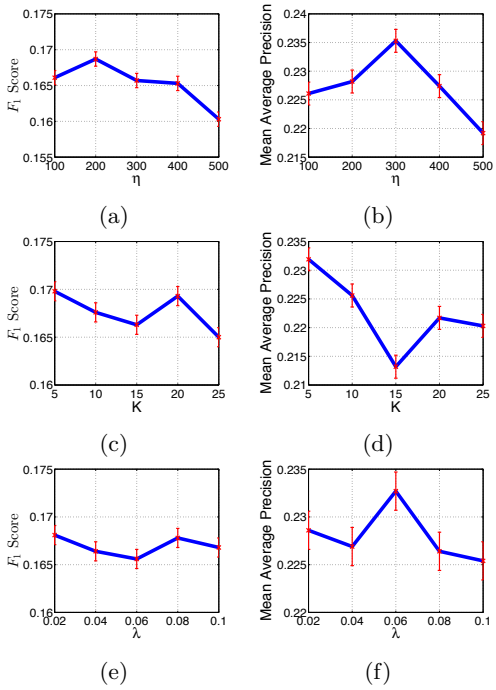


**Figure 3: Parameter study for HMER.**

## 4.7 $\varphi$: Music Emotion Clusters

Music emotion clustering is useful to enhance the user-experience in emotion-based music exploration because it avoids showing the large number of music emotions (e.g., there are 187 emotions defined by the All Music Guide). The topic-label distribution $\varphi$ trained by HMER can be interpreted as music emotion clusters. The top five topics/clusters for music emotions are listed in Table 4. The weight of each emotion represents its relevance to the topic/cluster. The representative emotions of the five topics/clusters are: happy, acerbic, sentimental, aggressive, and sad. The five clusters cover most clusters in the valence-arousal emotion plane [29].

## 4.8 $\varphi^{(l)}$: Correlation Between Music Emotion and Lyrics

HMER can also provide emotion-lyrics correlation using the label-lyrics distribution $\varphi^{(l)}$. Five emotions (delicate, cheerful, restrained, rollicking, and sweet) are presented and the corresponding major words are listed in Table 5. Personal pronouns and articles were filtered out. Overall, there

**Table 4: Top five topics of music emotion.**

| Topic | Major emotions(weight) |
|---|---|
| Topic 1 | happy(0.137), carefree(0.137), fun(0.091), joyous(0.091), innocent(0.091), cheerful(0.082) |
| Topic 2 | acerbic(0.308), sarcastic-cynical(0.264), fiery(0.089), snide(0.045), rollicking(0.045) |
| Topic 3 | sentimental(0.170), poignant(0.170), reserved(0.114), soothing(0.114), rousing(0.058), wistful(0.058) |
| Topic 4 | aggressive(0.184), menacing(0.153), rebellious(0.153), malevolent(0.123), angry(0.062) |
| Topic 5 | sad(0.295), somber(0.253), nocturnal(0.169), nihilistic(0.085), enigmatic(0.002) |

are no contradictory words for each emotion. Some words are frequently related to the reported emotions. For example, river and snow relate to delicate, dream and need relate to sweet. This result shows that our proposed model captures reasonable keywords describing emotions.

**Table 5: Top three lyrics words of music emotion.**

| Emotion | Major words(weight) |
|---|---|
| Delicate | river(0.016), snow(0.013), night(0.012) |
| Cheerful | forever(0.019), remember(0.018), happy(0.016) |
| Restrained | little(0.016), ocean(0.015), love(0.012) |
| Rollicking | Saturday(0.023), night(0.021), baby(0.018) |
| Sweet | dream(0.022), need(0.021), call(0.017) |

The label-lyrics distribution $\boldsymbol{\varphi}^{(l)}$ can also be useful in music exploration. For example, keywords/tags recommendation can be made for each song by the element-wise product of $\boldsymbol{\varphi}^{(l)}$ and the song-label distribution $\mathcal{L}_d$, which is useful in music previewing and indexing.

# 5. RELATED WORK

In this section, we give an overview of two aspects of related work: 1) music emotion recognition; 2) multi-view multi-label multi-instance models.

### *Music emotion recognition.*

Music emotion recognition is usually formulated as either a *classification* problem or a *regression* problem. For *classification* formulation, music is classified into multiple predefined emotion adjectives or clusters [28, 6], where Support Vector Machines [14] and Gaussian Mixture Models [15] were commonly used to solve it as a multi-class classification problem. However, these methods usually work on a small number of music emotion categories (e.g., four to eight), while a much larger number is needed (e.g., more than 50) for emotion based music recommender systems to provide fine-grained retrieval. When the number of emotions gets large, multiple similar yet different emotion labels can be used to describe music simultaneously. Therefore, music emotion recognition becomes a typical multi-label classification problem, where traditional multi-class methods fail since they assume each song has only one label.

To this end, multi-label classification methods have been proposed [21]. However, these methods assume that music emotions are consistent over each song, which ignores the fact that multiple labels may apply to different segments, but not necessarily the entire song. Thus, models trained for song-level labels and features may not be accurate.

Music emotion recognition is also commonly formulated as a *regression* problem to improve indexing accuracy. Emotion adjectives are translated into two dimensional real values, representing positiveness and strength of emotions respectively [5]. Then, regression models such as label propagation [26], Gaussian Mixture Models [24], and Support Vector Machine [11] are usually adopted. Specifically, LRC data was considered by Hu *et al.* in this line of research on song-level [10]. Although regression approaches provide accurate indexing, multiple labels are usually averaged as one single label [26] or fitted to a Gaussian distribution for each song [24], which does not consider the dynamics of music emotion. Schmidt *et al.* proposed to utilize music emotion dynamics to improve regression accuracy [20]. However, their method requires segment-level music emotion labels, and is therefore not scalable in real-world application.

### *Multi-view multi-label multi-instance learning.*

Multi-label multi-instance learning has been well studied in multi-label image classification [31, 30, 17]. In particular, M3LDA was proposed to address multi-view/modal multi-label image classification [17], where each image has two views: image and tags. Replacing image regions with music segments, and tags with lyrics, M3LDA can capture the dynamic nature of music emotion to some extent using the large number of song-level labels and lyrics. However, for music emotion recognition, lyrics and music are highly synchronized, which is different from the relation between image regions and tags. Also, the song-segment-sentence structure as well as music emotion correlation between both music segments and sentence cannot be handled by M3LDA due to single layer multi-instance structure and its exchangeability assumption for music segments.

# 6. DISCUSSION AND FUTURE WORK

### *Applicable scenarios.*

Our proposed model shows promising fitness to music emotion recognition. However, since HMER is in the multi-instance learning framework, the number of actual instances is much larger than single-instance learning methods, which is one major factor that slows down training and testing. Thus, HMER is mainly useful for off-line music emotion recognition tasks such as emotion based music indexing and context-free music recommendation. More efficient yet effective methods are still needed for real-time context-aware and personalized music recommendation.

In the future, we may consider parallel implementation of HMER by observing that each segment can be sampled independently given the prior obtained from the last iteration. Also, we may consider only segment-level music-lyrics synchronization to speed up HMER.

### *Extension to more views.*

Although we only consider music and lyrics views in this paper, HMER can be easily extended to more views given proper data. For example, in music videos, music, lyrics, and clips/images are highly synchronized and consistent in terms of expressed emotions. HMER can easily include clips/images to improve music emotion recognition by introducing a third view-variable on the sentence-layer.

# 7. CONCLUSION

In this paper, we investigated the dynamic nature of music emotion and modeled music emotion recognition as a novel *multi-label multi-layer multi-instance multi-view learning*

problem. We propose a novel hierarchical Bayesian model which uses sentence-level music and lyrics features, captures music emotion dynamics by a song-segment-sentence hierarchical structure. Our model also considers emotion correlations within a song by assuming its music segments follow correlated emotion distributions, and considers stricter emotion correlations within music segment by assuming its sentences follow the same emotion distributions. Experimental results have shown that our method outperforms several state-of-the-art models in terms of both $F_1$ score and mean average precision (MAP).

## 8. ACKNOWLEDGMENT

This work has been supported by Hong Kong Research Grants Council grants HKUST613112.

## 9. REFERENCES

[1] AllMusic moods. Online: http://www.allmusic.com/moods (9 Dec 2011).

[2] Bohemian rhapsody. Online (22 March 2014): http://www.queensongs.info/the-book/songwriting-analyses/no-synth-era/a-night-at-the-opera/bohemian-rhapsody.html.

[3] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 591–596, 2011.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. *Psychology*, (C-1):1–45, 1999.

[6] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma. Musicsense: contextual music recommendation using emotional allocation modeling. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 553–556, 2007.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[8] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu. Multi-layer multi-instance learning for video concept detection. *IEEE Transactions on Multimedia*, 10(8):1605–1616, 2008.

[9] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 159–168. ACM, 2010.

[10] Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 123–128, 2009.

[11] B. Jun Han, S. Rho, R. B. Dannenberg, and E. Hwang. Smers: Music emotion recognition using support vector regression. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 651–656, 2009.

[12] P. N. Juslin and J. A. Sloboda. *Music and emotion: Theory and research.* Oxford University Press, 2001.

[13] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *International Conference on Machine Learning and Applications*, pages 688–693. IEEE, 2008.

[14] T. Li and M. Ogihara. Content-based music similarity search and emotion detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 5, pages 705–708, 2004.

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

[16] M. I. Mandel and D. P. Ellis. Multiple-instance learning for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 577–582, 2008.

[17] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou. Multi-modal image annotation with multi-instance multi-label lda. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 1558–1564, 2013.

[18] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

[19] E. M. Schmidt and Y. E. Kim. Prediction of time-varying musical mood distributions using kalman filtering. In *IEEE International Conference on Machine Learning and Applications*, pages 655–660, 2010.

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

[21] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *Proceedings of the International Society for Music Information Retrieval Conference*, volume 8, pages 325–330, 2008.

[22] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2010.

[23] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 10(5):293–302, 2002.

[24] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 89–98, 2012.

[25] X. Wang, X. Chen, D. Yang, and Y. Wu. Music emotion classification of chinese songs based on lyrics using tf*idf and rhyme. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 765–770, 2011.

[26] B. Wu, E. Zhong, D. H. Hu, A. Horner, and Q. Yang. Smart: Semi-supervised music emotion recognition with social tagging. In *SIAM International Conference on Data Mining*, pages 279–287. SIAM, 2013.

[27] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[28] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):40, 2012.

[29] Y.-H. Yang and J.-Y. Liu. Quantitative study of music listening behavior in a social and affective context. *IEEE Transactions on Multimedia*, 15(6):1304–1315, 2013.

[30] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[31] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *Advances in Neural Information Processing Systems*, pages 1609–1616, 2006.