Emotion Recognition From Music Enhanced By domain Knowledge

Abstract

The well-established music grammar is often used to change audio elements of music to invoke audiences' emotional experience. Such music grammar, referred to as domain knowledge, is crucial for affective music content analyses, but has not been thoroughly explored. In this paper, we propose a novel method to analyze music emotion through exploring domain knowledge. Specifically, we first infer probabilistic dependencies between different main musical elements and emotions from the summarized music theory. Then, we transfer the domain knowledge to constraints, and formulate affective music content analysis as a constrained optimization problem. Experiments on the Music in 2015 database and the AMG1608 database demonstrate that the proposed affective content analyses method can successfully leverage well-established music grammar for better emotion regression from music content.

1 Introduction

We are surrounded by digital music collections due to the popularity of the Internet and the proliferation of user friendly MP3 players. Since almost every piece of music is created to convey emotion, naturally, music emotion recognition has attracted increasing attention in recent years. Automatic emotion recognition from music pieces has wide potential application in both music creation and music distribution.

The framework of current research into music emotion recognition mainly consists of feature extraction and classification. First, various features, including timbre, rhythm and harmony, are extracted from music pieces. Then, a classifier, such as support vector machine, is used to classify music pieces into several discrete emotion categories, or a regressor, such as support vector regression, is adopted to predict continuous emotional dimensions, such as valence and arousal. An extensive review of emotion recognition from music can be found in [Yang and Chen, 2012].

Although various discriminative features and classifiers have been developed, automatic emotion recognition from music pieces is still a very challenging task due to the complexity and subjectivity of human emotions, and the rich variety of music content.

Almost all the current work on music emotion recognition focuses on developing discriminative features and classifiers. This kind of data-driven approach does not successfully exploit the domain knowledge of emotion and music, i.e. the inherent psychological relationship between human emotion and music, which carries crucial information for music emotion recognition.

Specifically, main musical dimensions, i.e., rhythm, tonality, timbre, dynamics and pitch are often used to affect users' emotional experience. The tempo, mode, brightness, loudness and pitch can represent the five main musical dimensions respectively[Lartillot and Toiviainen, 2007]. Fig. 1[Sloboda, 2011]summarized the relations between music elements and emotions. From fig. 1, we can find that fast tempo is often used to generate the arousal atmosphere, while the slow tempo is used to create quiet environment [Fernández-Sotos *et al.*, 2016; Gabrielsson and Lindström, 2010; Gomez and Danuser, 2007; Husain *et al.*, 2002]. Major mode can be used to induce happiness and excitement, and minor mode can create a more tense and sad music[Miller, 2005]. Brightness is related to arousal[Gabrielsson and Lindström, 2010]. Higher brightness can be used to induce excitement and astonishment, while lower brightness can be used to induce sadness and softness. As for loudness, higher loudness can be used to induce anger, fear and excitement, and lower loudness can create a more relaxed and quiet music[Gabrielsson and Lindström, 2010]. High pitch may lead to happiness, anger and fear, while low pitch may induce sadness [Gabrielsson and Lindström, 2010].Such inherent dependencies between music elements and emotions can be leveraged for emotion recognition from music, but have not been explored yet.

Therefore, in this paper, we propose a novel method to analyze musical emotion through exploring domain knowledge. As a primary study to explore music theory for music emotion analysis, this paper takes main musical dimensional elements as an example to demonstrate the feasibility of the proposed music emotion analyses method enhanced through exploring domain knowledge. Specifically, we first infer probabilistic dependencies between main musical dimensional elements and emotions from the summarized music theory. Then we transfer the domain knowledge to constraints and formulate music emotion analyses as a constrained optimization prob-

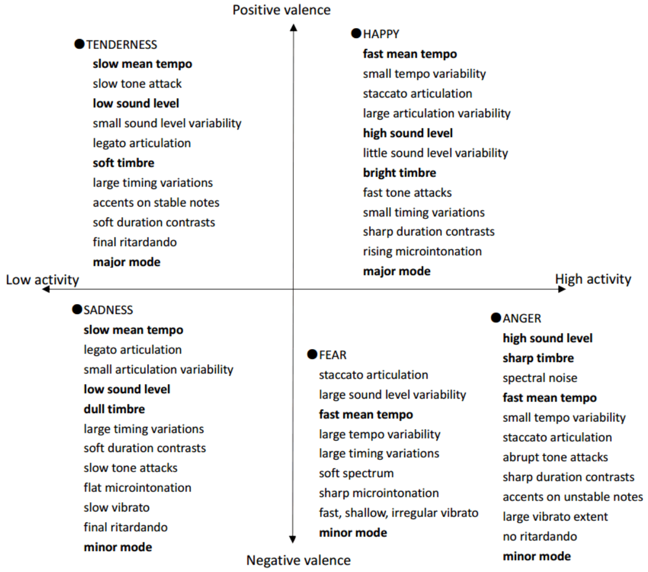

Figure 1: Music elements used by composers to communicate emotions to audiences

lem. Experiments on two benchmark databases demonstrate the superiority of the proposed method.

2 Domain Knowledge

Musical dimensions including rhythm, tonality, timbre, dynamics, and pitch, are used by composers to communicate emotions to audiences[Lartillot, 2011]. In this section, we introduce the dependencies between musical elements and emotions from the summarized music theory.

2.1 Rhythm

Tempo has great power to establish the mood of music and can greatly affect the emotions of the audiences [Fernández-Sotos *et al.*, 2016; Gabrielsson and Lindström, 2010; Gomez and Danuser, 2007; Husain *et al.*, 2002]. Generally, tempo represents the rhythm of the music[Lartillot, 2011]. By adjusting the tempo, the composers can fully design the emotions which are contained in their music. Specifically, as mentioned in [Gabrielsson and Lindström, 2010], *fast tempo* is designed to create the exaggerated atmosphere, which invokes high arousal mood from the audiences. On the other hand, the *slow tempo* is adopted to create quiet atmosphere, which invokes low arousal mood from the audiences.

Perceived music tempo can influence audiences' feelings. Specifically, when listening to music with fast tempo signals, people intuitively associate it with activity/excitement, happiness/joy/pleasantness, potency, surprise, flippancy, anger, uneasiness, and fear. Slower tempos are associated with calmness/serenity, peace, sadness, dignity/solemnity, tenderness, longing, boredom and disgust[Gabrielsson and Lindström, 2010].

Tempo is typically categorized into Largo (40-60bpm), Adagio (66-76bpm), Andante (76-108bpm), Moderato (108-120bpm) and Allegro(120-168) [Miller, 2005]. Largo, Larghetto, and Adagio are in slow pace, Andante is in a intermediate pace and typically considered as walking pace, and Moderato and Allegro are in fast pace. Since it is not easy to distinguish slow pace from intermediate pace, we treat them as an group. There are obvious difference between fast pace and intermediate pace, and we adopt the 108bpm as the threshold for categorizing the tempo as fast or slow rhythm.

Table 1: The dependencies between four music elements (tempo, mode, brightness and loudness) and emotions. Note that the √ demonstrates great dependencies between emotion and the music elements. Details are discussed in Sec. 2

	high A	low A	high V	low V
fast tempo	√			
slow tempo		√		
major mode			√	
minor mode				√
high brightness	√			
low brightness		√		
high loudness	√			
low loudness		√		

2.2 Tonality

Tonality is one of the most important musical elements for music presentation. Since mode is a system of musical tonality involving a type of scale coupled with a set of characteristic melodic behaviors[Miller, 2005], composers control the musical tonality by adjusting the mode. Generally, the mode is scaled into a heptatonic scale in which the first-, third-, and fifth-scale degrees play important roles. Composers control the musical tonality by adjusting the mode. Generally, the mode is scaled into a heptationic scale, in which the first, third, and fifth scale degrees play important roles. As stated in [Miller, 2005], the mode is categorized into two groups: major mode and minor mode. Specifically, the major mode is constructed by adjusting the first-, third-, and fifth-scale degrees with a major triad, while the minor mode is constructed by adjusting the first-, third-, and fifth-scale degrees with a minor triad. Major modes tend to convey a sense of grace, while minor modes imply anxiety and sadness. Studies show that major mode is strongly correlated to grace, serene, and solemn, while minor mode is strongly correlated to dreamy, dignified, tension, disgust, and anger[Husain *et al.*, 2002]. Thus, the major mode is used to invoke positive valence from audiences, and the minor mode is used to induce negative valence.

In this paper, we extract the mode features which ranges between -1 and +1 with the MIR toolbox. After obtaining the mode features, we adopt the median mode value as the threshold and categorize the clips into major or minor tonality. Specifically, clips with mode values above the median are considered major tonality, while clips with mode values below the median are assigned minor tonality.

2.3 Timbre

Musical timbre is an important components constructing the music piece. By playing music with different musical instruments and equalizer, the composer can communicate the emotion of the music with the audiences. Composers typically compose the music in a bright tone for constructing joyful, angry or thrill atmosphere, while compose the music in dull sound for delivering the mood of tender or depression [Gabrielsson and Lindström, 2010]. Thus, the bright sound is used for invoking high valence from the audiences, while the dull sound is used to induce low valence.

Audiences are more likely to feel high arousal emotions[Trochidis and Lui, 2015], e.g., excitement and astonishment while listening to bright sounds. Dull sounds induce feelings of sadness or tenderness

In this paper, we adopt the brightness [Wessel, 1979] as the feature of musical timbre. Specifically, brightness measures the proportion of high frequency components (above 1500HZ) in the music piece. The formulation is shown as below:

$$Brightness = X_{above}/X_{total} * 100\% \tag{1}$$

where X_{above} represents component whose the energy above 1500Hz, and X_{total} represents the total energy of the music. We adopt the median brightness value as the threshold and categorize the clips into high timbre or low timbre. Clips are labelled as high timbre if their brightness values are above the median, while clips are assigned to the low timbre category if their brightness values are below the median.

2.4 Dynamics

In music, the dynamics of a piece is the variation in loudness between notes or phrases. Musicians often make use of loudness to create dynamic of a music piece. When playing a song, singers and instrumentalists express emotions through the loudness of the song. Specifically, they often refrain or chorus the song in louder in order to induce the high arousal from the audience. The volume of the music can strongly influence arousal[Gabrielsson and Lindström, 2010]. Loud sound usually conveys the emotion of anger, excitement, surprise and great joy. On the contrary, soft sound expresses peaceful mood, tender and sadness [Gabrielsson and Lindström, 2010]. Audience tend to feel a high arousal mood while listening to high dynamic songs, and tend to feel a low arousal mood while listening to low dynamic songs.

Loudness is strongly associated with arousal of human' emotion[Trochidis and Lui, 2015]. Specifically, loud sound can raise intension or excitement from the audiences, while low amplitude often soothes and pacify or deliver a sense of melancholy.

In this paper, root-mean-square amplitude (rms) is used to represent loudness. We extract the rms with the MIR toolbox, which uses values between 0 and 1. To categorize the rms, we adopt the median value as the threshold and separate the dataset into high loudness and low loudness. Songs with loudness larger than the median are assigned as *high loudness*, while the rests are labelled as *low loudness*.

2.5 Pitch

Pitch is another fundamental feature of music, which represents the judgement of frequency of notes. However, the observations and conclusions of the current studies regarding valence and arousal vary widely.

Specifically, in the valence space, some works [Collier and Hubbard, 1998; Hevner, 1937] states that high pitch can result in high valence emotions such as joy, happiness, and glad, while low pitch tends to provoke low valence feelings of sadness, agitation. However, there also exists works [Ilie and Thompson, 2006; Scherer and Oshinsky, 1977] noting that high pitch can lead to low valence emotion, e.g. anger and fear, while low pitch can induce high valence emotions such as pleasure from audiences.

The relation between arousal and pitch is also debated. Some works [Coutinho and Cangelosi, 2009; Scherer and Oshinsky, 1977] point out that high pitch level can provoke high arousal emotions from audiences, e.g. tension, excitement and anger. However, other work [Hevner, 1937] states that high pitch level may also lead to some low arousal emotions from the audiences like serene and grateful. Statements in [Hevner, 1937; Scherer and Oshinsky, 1977; Rigg, 1940] also demonstrates the contradictory in relations between low pitch level and emotions.

Although several researches note that pitch level has an effect on an audience' mood, the association between pitch and emotions is still obscure.

In conclusion, the dependencies between emotions and main musical dimensions including rhythm, tonality, timbre and dynamics discussed above are shown in Table 1.

3 Proposed Method

3.1 Problem Statement

Denote three tuple $S = \{(x_i, h_i, y_i)|i = 1, ..., N\}$, where x_i represents D-dimensional features, $h_i = (h_i^t, h_i^m, h_i^b, h_i^l) \in \{0, 1\}$ represents the binarized tempo values, mode values, brightness values and loudness values respectively, $y_i \in \{y_i^v, y_i^a | -1 \leq y_i^v, y_i^a \leq 1\}$ represents continuous valence and arousal values, and N is the number of training samples. The goal is to learn a classifier f(x, w) as follows:

$$\min_w \sum_{i=1}^{N} \alpha \ell(f_\theta(x_i), y_i) + \sum_{i=1}^{N} \beta L(x_i, h_i, y_i) \tag{2}$$

where α and β are the coefficients, $\ell(f_\theta(x_i), y_i)$ represents the basic loss function, and $L(x_i, h_i, y_i)$ captures the domain knowledge between music elements h and the emotion values y. The first term represents the loss function over training samples. The second term represents the regularization term reflecting domain knowledge.

For the first term, any loss function can be used. In this paper, we adopt the support vector regression as the basic loss function:

$$\ell(f_\theta(x_i), y_i) = \frac{1}{2}||\mathbf{w}||^2 + \alpha \sum_{i=1}^{N} \ell_\epsilon(f(x_i, w) - y_i) \tag{3}$$

where the function $\ell_\epsilon(z)$ satisfy the below:

$$\ell_\epsilon(z) = \begin{cases} 0, & if|z| \leq \epsilon \\ |z| - \epsilon, & otherwise. \end{cases} \tag{4}$$

where ϵ is a constant which defines the maximum deviation allowed for a prediction to be considered as correct; α is used as a trade-off between the model complexity and regression loss.

For the second term, any domain knowledge, i.e, the relations between music elements and emotions, can be exploited to build better emotion classifiers from music. In this paper, domain knowledge of four music elements, i.e., tempo, mode brightness and loudness are discussed, with respect to dynamic, rhythm, timbre and tonality of the music dimension.

3.2 Representation of Domain Knowledge

Domain knowledge in arousal space From Table 1, tempo, brightness and loudness have the strong relationship with musical emotion in the arousal space. Fast tempo features, high brightness and high loudness are more possible to invoke high arousal mood of audiences, while the slow tempo features, low brightness and low loudness are more likely to invoke the low arousal of the audiences. Thus we can infer the probabilistic dependencies between tempo and arousal emotion as:

$$\begin{aligned} p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 1) > p(\hat{y^a} < 0|h^{\{t,b,l\}} = 1) \\ p(\hat{y^a} < 0|h^{\{t,b,l\}} = 0) > p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 0) \end{aligned} \tag{5}$$

where $p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 1)$ and $p(\hat{y^a} < 0|h^{\{t,b,l\}} = 1)$ indicate the probabilities of high arousal and low arousal respectively, when observing fast tempo, high brightness and loudness. $p(\hat{y^a} < 0|h^{\{t,b,l\}} = 0)$ and $p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 0)$ indicate the probabilities of low arousal and high arousal respectively, when given slow tempo, low brightness and low loudness.

In this paper, we adopt ReLU function to penalize the samples violating the domain knowledge. The corresponding penalty $l_i^{\{ta,ba,la\}}(x_i, h_i, \hat{y_i})$ from the domain knowledge according to Eq. 5 is encoded as below :

$$\begin{aligned} &\ell_i^{\{ta,ba,la\}}(x_i, h_i, \hat{y_i}) \\ &= h_i^{\{t,b,l\}} * [p(\hat{y^a} < 0|h^{\{t,b,l\}} = 1) - p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 1)]_+ + \\ &(1 - h_i^{\{t,b,l\}}) * [p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 0) - p(\hat{y^a} < 0|h^{\{t,b,l\}} = 0)]_+ \\ &= h_i^{\{t,b,l\}} * [1 - 2 * p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 1)]_+ \\ &+ (1 - h_i^{\{t,b,l\}}) * [2 * p(\hat{y^a} \geq 0|h^{\{t,b,l\}} = 0) - 1]_+ \end{aligned} \tag{6}$$

where $[\cdot] = max(\cdot, 0)$.

Since there is no obvious relationship between mode and arousal, we treat the major mode and minor equal important. In other words, major mode and minor mode have equal chances to invoke low arousal mood or high arousal mood from audiences. Hence, mode information is not used in arousal space.

Domain knowledge in valence space From Table 1, major mode(high-value mode) features are more possible to invoke high valence mood from audiences, while the minor mode(low-value mode) features are more likely to invoke the low valence of the audiences in the valence space. Thus we can infer the probabilistic dependencies between mode and valence emotion as:

$$\begin{aligned} p(\hat{y^v} \geq 0|h^m = 1) > p(\hat{y^v} < 0|h^m = 1) \\ p(\hat{y^v} < 0|h^m = 0) > p(\hat{y^v} \geq 0|h^m = 0) \end{aligned} \tag{7}$$

Thus the corresponding constraint $l_i^{mv}(x_i, h_i, \hat{y_i})$ for valence according to Eq. 7 is encoded as below:

$$\begin{aligned} &\ell_i^{mv}(x_i, h_i, \hat{y_i}) = h_i^m * [p(\hat{y^v} < 0|h^m = 1) - p(\hat{y^v} \geq 0|h^m = 1)]_+ \\ &+ (1 - h_i^m) * [p(\hat{y^v} \geq 0|h^m = 0) - p(\hat{y^v} < 0|h^m = 0)]_+ \\ &= h_i^m * [1 - 2 * p(\hat{y^v} \geq 0|h^m = 1)]_+ \\ &+ (1 - h_i^m) * [2 * p(\hat{y^v} \geq 0|h^m = 0) - 1]_+ \end{aligned} \tag{8}$$

Since there is no obvious relationship between valence and another elements, e.g. tempo, brightness, loudness, the information of tempo, brightness and loudness is not used in valence space.

3.3 Proposed Model

We propose to learn classifier with the objectives as below:

$$\begin{aligned} F^{\{a,v\}} = \frac{1}{2}w^T w + \alpha \sum_{i=1}^{N} \ell_\epsilon(f(x_i, w) - y_i) + \\ \beta^t \sum_{i=1}^{N} \ell_i^{\{ta\}}(x_i, h_i^t, \hat{y_i}) + \beta^m \sum_{i=1}^{N} \ell_i^{\{mv\}}(x_i, h_i^m, \hat{y_i}) + \\ \beta^b \sum_{i=1}^{N} \ell_i^{\{ba\}}(x_i, h_i^b, \hat{y_i}) + \beta^l \sum_{i=1}^{N} \ell_i^{\{la\}}(x_i, h_i^l, \hat{y_i}) \end{aligned} \tag{9}$$

where w is the parameter of the classifier, α, β^t, β^m, β^b and β^l are coefficients. In this model, we use $f(x, w) = w \cdot \phi(x)$ as our score function where $\phi(x)$ maps the features space into the kernel space. According to the property of logistic regression, we apply sigmoid function to replace the probalilistic dependencies between audio elements and emotion labels as follow:

$$\begin{aligned} p(\hat{y} > 0|h) = \sigma(f(x, w)) \\ p(\hat{y} \leq 0|h) = 1 - \sigma(f(x, w)) \end{aligned} \tag{10}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$.

In order to solve the optimization we adopt the stochastic gradient descent(SGD) to solve the problem. The updating rule is shown as follows:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \frac{\partial F^{\{a,v\}}}{\partial w} \tag{11}$$

where t and η indicate the number of iterations and the learning rate respectively.

The gradient of loss function to the weight can be computed as below:

$$\begin{aligned} &\frac{\partial F^{\{a,v\}}}{\partial w} = w + \alpha \sum_{i=1}^{N} \frac{\partial \ell_i(f(x_i, w) - y_i)}{\partial w} \\ &+ \beta^t \sum_{i=1}^{N} \frac{\partial \ell_i^{\{ta\}}(f(x_i, h_i^t, \hat{y_i})}{\partial w} + \beta^m \sum_{i=1}^{N} \frac{\partial \ell_i^{\{mv\}}(f(x_i, h_i^m, \hat{y_i}))}{\partial w} + \\ &\beta^b \sum_{i=1}^{N} \frac{\partial \ell_i^{\{ba\}}(f(x_i, h_i^b, \hat{y_i}))}{\partial w} + \beta^l \sum_{i=1}^{N} \frac{\partial \ell_i^{\{la\}}(f(x_i, h_i^l, \hat{y_i}))}{\partial w} \end{aligned} \tag{12}$$

where the specific gradient of loss function to the weight is computed as:

$$\frac{\partial \ell_i(f(x_i,w)-y_i)}{\partial w} = \begin{cases} 0, & if\ \ |f(x_i)-y_i| \leq \epsilon \\ \phi(x), & otherwise. \end{cases} \qquad (13)$$

$$\frac{\partial \ell_i^{ta}(f(x_i,h_i^t,\hat{y_i}))}{\partial w} = \begin{cases} -2\sigma(f(x_i,w))[1-\sigma(f(x_i,w))]\phi(x_i), \\ if\ h_i^t = 1\ and\ 1-2\sigma(f(x_i,w)) \geq 0 \\ 2\sigma(f(x_i,w))[1-\sigma(f(x_i,w))]\phi(x_i), \\ if\ h_i^t = 0\ and\ 2\sigma(f(x_i,w))-1 \geq 0 \\ 0, \qquad otherwise. \end{cases} \qquad (14)$$

Gradients of ℓ_i^{ta}, ℓ_i^{mv}, ℓ_i^{ba} and ℓ_i^{la} can be computed as Eq. 14 similarly.

The learning algorithm is shown in Algorithm 1.

Algorithm 1 Training algorithm of the proposed model

```
Input:
  training samples$(x_i, h_i, y_i)$,
  coefficient $\alpha$, $\beta^t$, $\beta^m$, $\beta^b$ and $\beta^l$ learning rate $\eta$
Output: Model parameters $w$
  Randomly initialize w;
  repeat
    for each training sample $(x_i, h_i, y_i)$ do
      Calculate the probabilistic dependencies $p(\hat{y} > 0|h)$
      and $p(\hat{y} \leq 0|h)$ as Eq.10;
      Calculate the specific gradient as Eq.13 and Eq.14;
    end for
    Calculate $\frac{\partial F^{\{a,v\}}}{\partial w}$ as Eq.12
    $w \leftarrow w - \eta(\frac{\partial F^{\{a,v\}}}{\partial w})$
  until
  Converges
  return w
```

After learning, the proposed approach can infer the affective value for testing samples according to function $f(x, w)$.

4 Experiments

4.1 Experimental conditions

To demonstrate the effectiveness of the proposed method, we conduct experiments on two benchmark databases: the Music Emotion in 2015 database [Aljanaki *et al.*, 2015] and the All Music Guide 1608 database(AMG1608) [Chen *et al.*, 2015].

The Music Emotion in 2015 database consists of royalty-free music, with diverse genres of rock, classical, pop, jazz, country, folk, rap etc.[Bittner *et al.*, 2014]. The database is divided into two subsets: the development set and the test set. Specifically, the development set consists of 430 clips of 45 seconds, and the test set is comprised of 58 complete music pieces with an average duration of 234 ± 105.7 seconds. We use 260 low-level feature set provided by [Aljanaki *et al.*, 2015], which are extracted using openSMILE features. The 260 dimensional feature set represent the music from 65 dimensional mean deviation, 65 dimensional standard deviation, and their first-order derivatives from acoustic descriptors. We also extract tempo, mode, brightness, loudness with MIR toolbox.

The AMG1608 database consists of 1608 preview clips of Western songs, collected from a popular music stream service named 7digit. Each preview clips is 30-second long. For experiments, we adopt the four-fold cross-validation on the database. We use the public feature set provided by [Chen *et al.*, 2015], including MFCC, Tonal, Spectral and Temporal. We also extract tempo, mode, brightness, loudness with MIR toolbox.

To further demonstrate the effectiveness of domain knowledge, we conduct the following experiments in the arousal space: music audio emotion analysis ignoring all domain knowledge (**none**), music audio emotion analysis only exploiting single domain knowledge (**tempo**, **brightness**, **loudness**), music audio emotion analysis exploiting two of domain knowledge(**tempo+brightness**, **tempo+loudness**, **brightness+loudness**) and music audio emotion analysis exploiting all domain knowledge (**tempo+brightness+loudness**). In the valence space, since mode is the only musical elements that affects the valence, we conduct experiments as: music audio emotion analysis ignoring all domain knowledge (**none**), and music audio emotion analysis exploiting mode (**mode**). We also conduct experiments using music audio emotion analysis fusing the musical elements as features (**fusion**).

Root-Mean-Square Error(RMSE) and Pearson Correlation(R) is adopted to evaluate the effectiveness of the proposed method.

During model training, we first initialize the weights to small random number, then we conduct model selection with grid search, by choosing the hyper parameter α, β^t, β^m, β^b and β^l ranging from {0.1, 1, 10, 20, 50} for simplicity. For each method, we monitor the objective cost on the training set and choose the hyper parameters with the smallest objective cost. On the Music Emotion in 2015 database, a fixed split of training/validation/testing 400/30/58 is adapted. On the AMG1608 database, we adopt 4-fold cross-validation.

Table 2: Music emotion analyses results on the music in 2015 database and the AMG1608 database in valence space

	Music in 2015 database		AMG1608 database	
	RMSE	R	RMSE	R^2
none	0.357	0.012	0.275	0.064
fusion	0.351	0.019	0.272	0.063
mode	**0.318**	**0.044**	**0.254**	**0.140**

Table 3: Music emotion analyses results on the music in 2015 database and the AMG1608 database in arousal space

	Music in 2015 database		AMG1608 database	
	RMSE	R	RMSE	R^2
none	0.270	0.3740	0.2670	0.5680
fusion	0.270	0.377	0.262	0.589
tempo	0.2626	0.4649	0.265	0.5975
brightness	0.2650	0.4887	0.266	0.6257
loudness	0.2618	0.4759	0.252	0.6068
tempo+brightness	0.2454	0.5185	0.264	0.6395
tempo+loudness	0.2550	0.5417	0.246	0.6162
brightness+loudness	0.2566	0.5782	0.244	0.6461
tempo+brightness+loudness	**0.2340**	**0.5970**	**0.240**	**0.669**

4.2 Experimental results and analysis

Table 2 and Table 3 show the music audio analyses results on the Music Emotion in 2015 database and the AMG1608 database in the valence space and arousal space. From Table 2 and Table 3, we observe as follows:

First, the proposed method exploiting all domain knowledge has the best performance among all methods with the lowest RMSE and highest Pearson correlation. Specifically, compared with music audio analyses ignoring all domain knowledge, the proposed method achieves 0.039 and 0.021 decrement of RMSE, and 0.032 and 0.076 increment of Pearson correlation, with respect to the Music Emotion in 2015 database and the AMG1608 database in the valence space. In the arousal space, the proposed method decrease the RMSE of 0.036 and 0.027, and increase the Pearson correlation of 0.223 and 0.101 on the Music Emotion in 2015 database and the AMG1608 database respectively. The method ignoring domain knowledge is totally data-driven method, which only learns the mapping from the extracted features to the predictions and it ignores the well-established music knowledge. On the contrary, the proposed method leverages both domain knowledge and training data, and thus achieves better performance.

Second, the methods leveraging more domain knowledge have better performance than that exploiting less domain knowledge. Specifically, in the arousal space, the methods leveraging one domain knowledge is inferior to the methods leveraging two domain knowledge. Since temp, brightness, and loudness describes the music from different aspects, the effects of these musical elements on the music emotion analyses are complementary. Thus, the methods leveraging more domain knowledge can capture more relations between music elements and emotion, and achieves better performance.

4.3 Comparison with related work

To further demonstrate the effectiveness of the proposed method, we compared the proposed method with the state of art.

On the Music Emotion in 2015 database, we compare the proposed method with Aljanaki's [Aljanaki *et al.*, 2015], Liu's [Liu *et al.*, 2015], Chin's [Chin and Wang, 2015], Markor's [Markov and Matsui, 2015], and Patra's[Patra *et al.*, 2015]. Specifically, Aljanaki *et al.* provided the baseline for MediaEval 2015. Liu *et al.* proposed Arousal-Valence Similarity Preserving Embedding (AV-SPE) to extract the intrinsic features embedded in music signal, and train the SVR which takes the extracted features as the input and the emotion values as labels; Chin *et al.* adopted deep recurrent neural network to predict the valence and arousal for each moment of a song; Markor *et al.* used Kernel Bayes Filter (KBF) for predicting the valence and arousal. Patra *et al.* proposed the music emotion recognition system consisting of feed-forward neural networks, which predicts the dynamic valence and arousal values continuously. The comparisons are shown in Table 4. From the table, we observe as follows:

Compared with the others' works, the proposed method has best performance in most cases. The state of the art only learns the maps from the features, and makes prediction of the music emotion. On the contrary, the proposed method not only learns the mapping from the features, but also captures the dependencies between musical elements and emotions through domain knowledge. Thus the proposed capture more information, and achieves better performance.

Table 4: Comparison with related works on the Music Emotion in 2015 database

Database	Music emotion in 2015			
Models	Arousal		Valence	
	RMSE	R	RMSE	R
Our Model	**0.234**	**0.597**	**0.318**	**0.044**
Baseline	0.27	0.36	0.37	0.01
Liu *et al.*'s	0.2377	0.5610	0.3834	-0.0217
Chin *et al.*'s	0.2555	0.3417	0.3359	-0.0103
Markov *et al.*'s	0.419	0.498	0.620	-0.035
Patra *et al.*'s	0.2689	0.4678	0.3538	-0.0082

Table 5: Comparison with related work on the AMG1608 database

Database	AMG1608			
Models	Arousal		Valence	
	AED	R^2	AED	R^2
Our Model	**0.240**	**0.669**	**0.254**	**0.140**
Baseline	0.288	0.651	0.288	0.120

Rare work is conducted on the AMG1608 database. Thus, we only compare the proposed method with the baseline methods provided in [Chen *et al.*, 2015]. In [Chen *et al.*, 2015], Chen *et al.* adopted the Music emotion recognition (MER) system to recognize music emotion on the AMG1608 database. We adapted the Average Euclidean Distance (AED) and Pearson correlation as evaluation. The comparison is shown in Table. 5. From the table, we observe as follows:

Compared with baseline method, the proposed method achieve better performance of AED and Pearson correlation. Since the proposed method captures the more information by constraints of domain knowledge, it is reasonable that the proposed method achieves better performance.

Taking the comparisons above into consideration, the proposed method has an excellent generalization ability with respect to affective audio music analysis. This demonstrates the effectiveness of the proposed method.

5 Conclusion

In this paper, we propose a novel method to analyze music emotion recognition through exploring domain knowledge. We first investigate the probabilistic dependencies between emotions and music elements, i.e., tempo, mode, brightness and loudness. Then we transfer such probabilistic dependencies to the domain knowledge constraints for music emotion recognition. The experimental results on the Music emotion in 2015 database and the AMG1608 database demonstrate the importance of the domain knowledge. This further demonstrates the superiority of the proposed method to music emotion recognition.

References

[Aljanaki *et al.*, 2015] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[Bittner *et al.*, 2014] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, pages 155–160, 2014.

[Chen *et al.*, 2015] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. The amg1608 dataset for music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 693–697. IEEE, 2015.

[Chin and Wang, 2015] Yu-Hao Chin and Jia-Ching Wang. Mediaeval 2015: Recurrent neural network approach to emotion in music tack. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[Collier and Hubbard, 1998] William G Collier and Timothy L Hubbard. Judgments of happiness, brightness, speed, and tempo change of auditory stimuli varying in pitch and tempo. *Psychomusicology*, 17(1/2):36–55, 1998.

[Coutinho and Cangelosi, 2009] Eduardo Coutinho and Angelo Cangelosi. The use of spatio-temporal connectionist models in psychological studies of musical emotions. *Music Perception: An Interdisciplinary Journal*, 27(1):1–15, 2009.

[Fernández-Sotos *et al.*, 2016] Alicia Fernández-Sotos, Antonio Fernández-Caballero, and José M Latorre. Influence of tempo and rhythmic unit in musical emotion regulation. *Frontiers in Computational Neuroscience*, 10, 2016.

[Gabrielsson and Lindström, 2010] Alf Gabrielsson and Erik Lindström. The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications*, pages 367–400, 2010.

[Gomez and Danuser, 2007] Patrick Gomez and Brigitta Danuser. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7(2):377–387, 2007.

[Hevner, 1937] Kate Hevner. The affective value of pitch and tempo in music. *The American Journal of Psychology*, 49(4):621–630, 1937.

[Husain *et al.*, 2002] Gabriela Husain, William Forde Thompson, and E Glenn Schellenberg. Effects of musical tempo and mode on arousal, mood, and spatial abilities. *Music Perception: An Interdisciplinary Journal*, 20(2):151–171, 2002.

[Ilie and Thompson, 2006] Gabriella Ilie and William Forde Thompson. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception: An Interdisciplinary Journal*, 23(4):319–330, 2006.

[Lartillot and Toiviainen, 2007] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[Lartillot, 2011] Olivier Lartillot. Mirtoolbox 1.3. 4 user's manual. *Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland*, 2011.

[Liu *et al.*, 2015] Yang Liu, Yan Liu, and Zhonglei Gu. Affective feature extraction for music emotion prediction. 2015.

[Markov and Matsui, 2015] Konstantin Markov and Tomoko Matsui. Dynamic music emotion recognition using kernel bayes' filter. 2015.

[Miller, 2005] Michael Miller. *The complete idiot's guide to music theory*. Penguin, 2005.

[Patra *et al.*, 2015] Braja Gopal Patra, Promita Maitra, Dipankar Das, and Sivaji Bandyopadhyay. Mediaeval 2015: Music emotion recognition based on feed-forward neural network. In *MediaEval*, 2015.

[Rigg, 1940] Melvin G Rigg. The effect of register and tonality upon musical mood. *Journal of Musicology*, 1940.

[Scherer and Oshinsky, 1977] Klaus R Scherer and James S Oshinsky. Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, 1(4):331–346, 1977.

[Sloboda, 2011] John Sloboda. *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2011.

[Trochidis and Lui, 2015] Konstantinos Trochidis and Simon Lui. Modeling affective responses to music using audio signal analysis and physiology. In *International Symposium on Computer Music Multidisciplinary Research*, pages 346–357. Springer, 2015.

[Wessel, 1979] David L Wessel. Timbre space as a musical control structure. *Computer music journal*, pages 45–52, 1979.

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