SVR BASED DOUBLE-SCALE REGRESSION FOR DYNAMIC EMOTION PREDICTION IN MUSIC

Haishu Xianyu, Xinxing Li, Wenxiao Chen, Fanhang Meng, Jiashen Tian, Mingxing Xu, Lianhong Cai

Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing, China
xumx@tsinghua.edu.cn, {lixinxin14,xyhs14}@mails.tsinghua.edu.cn

ABSTRACT

Dynamic music emotion prediction is to recognize the continuous emotion contained in music, and has various applications. In recent years, dynamic music emotion recognition is widely studied, while the inside structure of the emotion in music remains unclear. We conduct a data observation based on the database provided by Free Music Archive (FMA), and find that emotion dynamic shows different properties under different scales. According to the data observation, we propose a new method, Double-scale Support Vector Regression (DS-SVR), to dynamically recognize the music emotion. The new method decouples two scales of emotion dynamics apart, and recognizes them separately. We apply the DS-SVR to *MediaEval 2015, Emotion in Music* database, and achieve an outstanding performance, significantly better than the baseline provided by organizer.

Index Terms— Music, Emotion, Multi-scale, Double-scale Support Vector Regression

1. INTRODUCTION

As an important art form, music arouses real emotion effects in people. As such it has often been referred to as a language of emotion [1]. Music Emotion Recognition (MER) can be of help in understanding the content of the music, and has been widely used in music indexing, recommendations and in other application scenarios [2]. As an important part of MER, dynamic music emotion recognition has been widely investigated in recent years. In dynamic music emotion recognition, by following a particular labeling rate, time-continuous emotion labels can be marked, for the task of recognizing the emotion sequence as based on the music data [3].

One straightforward method is to predict the emotion labels based on short-term features [4]. The short-term features are extracted from frames, and the frame moves with a shift equal to that of the labeling period. A shortcoming in this method is that it lacks contextual information. In order to make up for this shortcoming, function smoothing can be added to the recognition result, which makes prediction result smoother and more accurate.

In recent years, Long Short-Term Memory (LSTM) has been applied also to the MER problem and has shown to have a good performance [5]. LSTM is a recurrent neural network (RNN) architecture published in 1997 [6]. Unlike the mapping of a single feature to a single label, LSTM maps the feature sequence to a label sequence, which can be used to more naturally combine local information and contextual information. However, LSTM does not take into account the inner structure of the music data, and so MER is considered as being only a sequence recognition problem.

Both the straightforward and LSTM methods detailed above are used to directly map feature into emotion. However, some researches have introduced a middle-layer space between the feature space and the emotion space. In 2014, Naveen Kumar and Rahul Gupta proposed a new method, which they proposed to predict dynamic labels from global features through using Haar transform [7]. In the same year, authors in [8] proposed a two-step method, by which the researcher would construct a dispersed middle-layer representation. These methods are attempts to make use of the distribution information of the music; however, they do not include consideration of the inner structure of music.

It has been proposed that music data takes a three-scale structure, and so presents different properties under different scales [9]. We propose that dynamic music emotion is affected by this multi-scale structure, and by making full use of scale information, recognition algorithm can be simpler and more effective.

In order to investigate the details of multi-scale structure in music, a data observation process was conducted. From the analysis, it was found that there were two principal scales of emotion dynamics in music: global-scale dynamics, which exists between different songs, and local-scale dynamics, with a period from 1 s to 3 s. The two scales of emotion

This work is supported by the 973 Program (2012CB316401) of China, and is partially supported by the National Natural Science Foundation of China (61171116, 61433018, 61375027) and 863 Program (2015AA016305).

dynamics presented different physical and static characteristics, and so should be treated separately.

Based on the analysis result, a new method is proposed to recognize time-continuous emotion in music: Double-scale Support Vector Regression (DS-SVR). The principle of the DS-SVR is simple: it treats the global-scale dynamics as the base platform, and the local-scale dynamics as the small changes on the platform. By decoupling the two scales of dynamics, recognizing them separately and then combining them the recognizing result is obtained. This method can make use of the global information within the music, while not losing the details relating to the emotion dynamic. The results were presented at *MediaEval 2015* from using methods based on DS-SVR, and the results produced were well received [10].

The rest of this paper is organized as follows. Section 2 shows the content and result of the data observation. Section 3 describes the DS-SVR algorithm used. The details of the experiment are presented in Section 4, and the results of the experiment are presented in Section 5. Conclusions are presented in Section 6.

2. DATA OBSERVATION

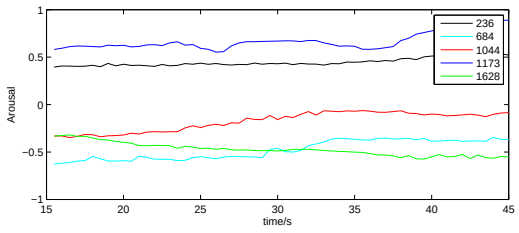
As was earlier proposed in [9], there exists three main dynamic scales in music. The lowest scale, *wave scale*, takes wavelength as an element, covering the frequency range from $20Hz$ to $2kHz$. The higher, *phrase scale*, corresponds to the music phrase, with periods ranging from $1s$ to $5s$. The highest is *movement scale*, and song may contains one or more movements, whereas the music style remains unchanged within the movement [11]. The data observation approach follows the assumption that since the music emotion is determined by the music data, a similar multi-scale structure should exist with respect to the music emotion sequence.

The database used for data observation was sourced from the "Emotion in Music" task in *MediaEval* [3]. The database contained 1000 songs, which were selected from *Free Music Archive(FMA)*. For each song, $30s$ emotion annotations containing Valence and Arousal values are provided with a 2Hz sampling rate. Thus, there were 60 Valence values and 60 Arousal values in each song [12].

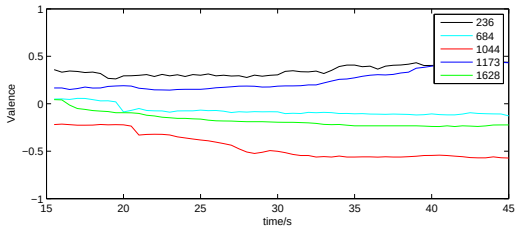
2.1. Multi-scale Structure of Emotion

The emotion changing curves of different songs is presented in Figure 1. Five songs were chosen at random from the database, and their Arousal curves can be seen in the figure. It can be seen that the emotion difference between songs was significant, while the emotion curve of the same song remained stable.

The variance of emotion labels inside the same song and between songs was calculated by:



(a) Arousal



(b) Valence

Fig. 1. Emotion sequence of song NO. 236, 684, 1044, 1173 and 1628

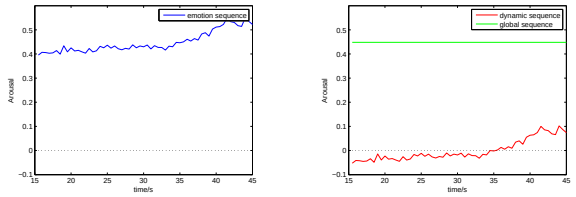
$$
\begin{aligned}
V_{in} &= E_i(Var_j(L_{i,j})) \\
V_{out} &= Var_i(E_j(L_{i,j})) \\
i &= 1 \ldots 1000, j = 1 \ldots 60
\end{aligned} \tag{1}
$$

where $E(\cdot)$ is the averaging function, $Var(\cdot)$ is the variance function, and $L_{i,j}$ is the Valence or Arousal value of the j-th label in the i-th songs. The result showed that $V_{in} = 0.0077, V_{out} = 0.1167$ on Valence, and $V_{in} = 0.0090, V_{out} = 0.1000$ on Arousal. It can be seen that V_{out} is 10 times greater than V_{in}.

Based on the analysis and calculation above, a two-scale structure in music emotion dynamic could be generalized, which corresponded to the three-scale structure in music data [9]:

- **Global-scale emotion dynamic**, corresponding to *movement scale*, represents the emotion variety between different songs, and determines the basic emotion of a song;
- **Local-scale emotion dynamic**, corresponding to *phrase scale*, represents the detailed emotion dynamic inside a song, with period larger than 1s, shorter than 3s.

As presented in Figure 2, global-scale dynamic and local-scale dynamic were coupled together. This can be considered as being similar to electrical direct current and alternating current: direct current determine the base current, and alternating current provides the dynamic. The two sorts of current can be coupled together to form the actual current, or handle separately by decomposition.



(a) Before Decoupling (b) After Decoupling

Fig. 2. Decoupling emotion dynamics of song NO. 236

2.2. Relationship Between Features and Emotion-Scales

The Person correlation coefficients were calculated with respect to features used in INTERSPEECH 2013 ComParE Challenge [13] and global-scale / local-scale emotion dynamic. The most relevant features are listed in Table 1.

feature	r	
	global	local
mfcc[1]-maxPos	**0.576**	0.377
spectralHarmonicity-centroid	**0.569**	N/A
mfcc[1]-minPos	**0.564**	0.221
spectralRollOff90.0-risetime	**0.554**	0.155
peakRangeRel	**0.551**	0.016
spectralHarmonicity	0.228	**0.526**
lengthL1norm	0.162	**0.430**
lengthL1norm-dif	0.148	**0.427**
spectralEntropy	0.207	**0.393**
spectralFlux	0.233	**0.386**

Table 1. Top-5 features relevant to global-scale and local-scale emotion. (r represents Pearson correlation coefficient)

Features which were relevant to global-scale dynamic included MFCC, spectral RollOff and peakRange. These features were related with music genre, or music style, which corresponded to the *movement scale*. Features relevant to local-scale dynamic were each short-term features. These features were related to the local timbre in music, which corresponded to the *phrase scale* in music data [9].

From the table 1, it can be seen that global-scale and local-scale dynamic are related but with different features. Thus, by using different features when recognizing global-scale and local-scale dynamic this could promote the accuracy of prediction.

3. ALGORITHM

From the above analysis, a new recognition method named DS-SVR was proposed, which consists of two independent SVRs on two different scales. SVR [14] has advantages for high-dimensional regressions since the SVR optimization is independent with the dimension of the input.

3.1. Notation

Let $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s\}, Y = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t\}$ be two vector sequences. Using the following notations: $\overline{X}$ is the average value of X; $\{\overline{X}\}$ is a sequence consisting of s elements, all elements equal to $\overline{X}$; $\langle X, Y\rangle$ by combining X and Y together; when $s = t$, $X \pm Y = \{\boldsymbol{x}_1 \pm \boldsymbol{y}_1, \cdots, \boldsymbol{x}_s \pm \boldsymbol{y}_s\}$.

$S_t = \{M_1^t, \cdots, M_m^t\}$ and $S_e = \{M_1^e, \cdots, M_n^e\}$ represent the training set and the evaluation set, respectively. Here the m, n are the sizes of the two sets. M_i^t is the i-th song in the train set, and M_j^e is the j-th song in the evaluation set. Correspondingly, $L_t = \{L_1^t, \cdots, L_m^t\}$ and $L_e = \{L_1^e, \cdots, L_n^e\}$ were taken to represent the emotion annotations of training set and evaluation set, respectively. Here L_i^t, L_j^e are each label sequences, where each label in the sequence contains an Arousal value and a Valence value.

3.2. Double-scale SVR

For each song M_i^t in S_t, a global feature vector $\boldsymbol{x}_i^t$ was extracted with a local feature sequence Y_i^t. As for L_i^t, the average label $\overline{L_i^t}$ was calculated and also the sequence $D_i = L_i^t - \{\overline{L_i^t}\}$. Two models were trained with SVR:

$$\begin{aligned} mod1 &: \{\boldsymbol{x}_1^t, \cdots, \boldsymbol{x}_m^t\} \rightarrow \{\overline{L_1^t}, \cdots, \overline{L_m^t}\} \\ mod2 &: \langle Y_1^t, \cdots, Y_m^t\rangle \rightarrow \langle D_1, \cdots, D_m\rangle \end{aligned} \tag{2}$$

For song M_j^e in S_e, the feature vector $\boldsymbol{x}_j^e$ was extracted, and also the feature sequence Y_j^e. These features were input into $mod1$ and $mod2$:

$$\begin{aligned} \{\boldsymbol{x}_1^e, \cdots, \boldsymbol{x}_n^e\} &\xrightarrow{mod1} \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n\} \\ \langle Y_1^e, \cdots, Y_n^e\rangle &\xrightarrow{mod2} \langle Z_1, \cdots, Z_n\rangle \end{aligned} \tag{3}$$

Finally, the emotion label sequence of j-th music in the evaluation set was calculated from $\boldsymbol{w}_j$ and Z_j:

$$P_j = Z_j + \{\boldsymbol{w}_j\}, j = 1 \ldots n \tag{4}$$

Thus, $\{P_1, \cdots, P_n\}$ is the prediction result.

3.3. Dealing with Variable-length of Song

The lengths of songs were not fixed in S_e and S_t(In Section 4, but the lengths of songs were fixed in S_t). Thus, when conducting DS-SVR on S_e and S_t, there were three options:

op1 Extracting only one global feature for each song in S_e and S_t, and obtaining one global emotion value.

op2 Dividing each song in S_t and S_e into several segments, each segment l seconds, with no overlap between segments.

op3 Dividing each song in S_t and S_e into several segments, each segment l seconds, with $p\%$ overlap.

4. EXPERIMENT

4.1. Feature

The feature was extracted by using *openSMILE*. The feature set that was used in global-SVR was INTERSPEECH 2013 ComParE Challenge feature set (*IS13-ComParE*), and it contained 6373 features [13].

The feature set used for local-SVR was a subset of IS13-ComParE set, *IS13-ComParE-LLD*, with only the low-level descriptors included. It contained 260 features [15].

4.2. Experiment Configuration

The data set was divided up as per *MediaEval 2015*. The training set contained 431 clips, and each song had dynamic annotations for 30 seconds. The evaluation set contained 58 complete songs, and the lengths of dynamic annotations were not fixed, ranging from 49 seconds to 627 seconds [3].

The experiment reified each of the three options mentioned in Section 3.3, with $l = 30, p = 50$. The three experiments for op1, op2 and op3 were named Expe1, Expe2 and Expe3 respectively.

4.3. Contrast Experiments

Three contrast experiments were chosen from MediaEval 2015, which were constructed around DRNN, LSTM-RNN and SVR. The DRNN and LSTM-RNN have shown good performance with respect to MER problem in recent years, and SVR is the algorithm on which DS-SVR is based.

1. DRNN: Deep RNN (DRNN) was used to predict and when performing back-propagation. For this project the team used Limited Memory Broyden Fletcher Goldfarb Shanno algorithm (LBFGS) to update the weights [16].

2. SVR: UNIZA system based on SVR with Radial Basis kernel function was used [17].

3. LSTM-RNN: The model was by Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) for dynamic Arousal and Valence regression [18].

5. RESULT AND DISCUSSION

Table 2 presented the experimental results of DS-SVR and the contrasting methods. From these results it can be seen that DS-SVR performs best for Valence, and second-best for Arousal. The three experiments had similar performance in terms of RMSE, whereas Expe3 performed best in terms of Pearson correlation coefficient.

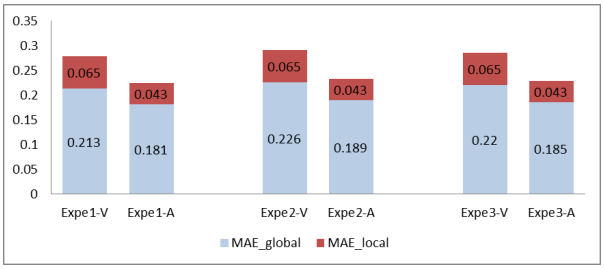


Fig. 3. MAE of Expe1, Expe2 and Expe3

Method	Valence		Arousal	
	RMSE	r	RMSE	r
Expe1 (op1)	**0.303**	0.01	0.250	0.56
Expe2 (op2)	0.310	0.01	0.245	0.58
Expe3 (op3)	0.307	**0.03**	0.248	0.60
Baseline [3]	0.366	0.010	0.270	0.360
DRNN [16]	0.336	-0.01	0.342	0.26
SVR [17]	0.366	-0.02	0.255	0.51
LSTM-RNN [18]	0.366	0.02	**0.234**	**0.61**

Table 2. Experiment results. (r represents Pearson correlation coefficient and RMSE represents Root Mean Squared Error. Baseline was provided by organizers.)

The mean average error (MAE) of the three experiments were also calculated. Figure 3 shows the composition of MAE; it can be seen that the error mainly arises in global-scale. When the global-scale error was large, the effect of the local-scale prediction results in large error. Thus, there is a need to accurately recognize the global-scale emotion as the basis for the dynamic music emotion recognition. DS-SVR considers the global-scale emotion independently, which can promote the global-scale recognition effect.

6. CONCLUSION

In this paper, in order to recognize dynamic emotion in music, a data observation was conducted and a DS-SVR method was proposed. This method benefited from the multi-scale structure in music, and showed a notably better performance when compared with the other methods considered.

As mentioned above in Section 2, there may be one or more music movements in a song. As a future work, we propose to extend the DS-SVR method to adaptively learn the boundaries of movements and adopt a dynamic model (like RNN, LSTM) for the local emotion, which may then help the method performs better in relation to longer musical pieces.

7. REFERENCES

[1] Carroll C Pratt, "Music as the language of emotion.," The Library of Congress, 1952.

[2] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*. Citeseer, 2010, pp. 255–266.

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

[4] Mathieu Barthet, György Fazekas, and Mark Sandler, "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models," in *Proc. CMMR*, 2012, pp. 492–507.

[5] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus R Scherer, "The munich lstm-rnn approach to the mediaeval 2014 emotion in music task," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, October 2014.

[6] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," in *Neural computation*. 1997, vol. 9, pp. 1735–1780, MIT Press.

[7] Naveen Kumar, Rahul Gupta, Tanaya Guha, Colin Vaz, Maarten Van Segbroeck, Jangwon Kim, and Shrikanth S Narayanan, "Affective feature design and predicting continuous affective dimensions from music," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, October 2014.

[8] Yuchao Fan and Mingxing Xu, "Mediaeval 2014: Thuhcsil approach to emotion in music task using multi-level regression," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, October 2014.

[9] Aniruddh D. Patel, "Music, language, and the brain," 2008, Oxford University Press.

[10] Mingxing Xu, Xinxing Li, Haishu Xianyu, Jiashen Tian, Fanhang Meng, and Wenxiao Chen, "Multi-scale approaches to the mediaeval 2015 "emotion in music" task," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[11] J. Kent Williams, "Engaging music: Essays in music analysis.: Engaging music: Essays in music analysis," *Music Theory Spectrum*, vol. 29, no. 2, pp. 263–275, 2007.

[12] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang, "1000 songs for emotional analysis of music," in *Proc. of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 2013, pp. 1–6.

[13] Mihye Lim and Ilsun Ko, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *Proceedings of Interspeech*, pp. 148–152, 2013.

[14] Alex Smola and Vladimir Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[15] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, 2013.

[16] 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*, September 2015.

[17] Chmulik Michal, Guoth Igor, Malik Miroslav, and Jarina Roman, "Uniza system for the "emotion in music" task at mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[18] Coutinho Eduardo, Trigeorgis George, Zafeiriou Stefanos, and Schuller Bjrn, "Automatically estimating emotion in music with deep long-short term memory recurrent neural networks," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.