PKU-AIPL' Solution for MediaEval 2015 Emotion in Music Task*

Kang Cai, Wanyi Yang, Yao Cheng, Deshun Yang, Xiaoou Chen
Institute of Computer Science and Technology,
Peking University, Beijing, China
{caikang, yangwanyi, chengyao, yangdeshun, chenxiaoou}@pku.edu.cn

ABSTRACT

In this paper, we describe the PKU-AIPL Team solution of *Emotion in Music* task in MediaEval benchmarking campaign 2015. We extracted and designed several sets of features and used continuous conditional random field(CCRF) for dynamic emotion characterization task.

1. INTRODUCTION

In *Emotion in Music* task, labelers provided v-a labels using a sliding bar while they listened to the music, which made the labels of the music segments strongly dependant on their previous segments. In our solution, we first estimate each segment's label based on the audio features, assuming music segments are independent instances. Then, we break the independence assumption and further optimize the labels by modeling music emotion labeling as a continuous conditional random field process.

The rest of this paper is organized as follows. Section 2 describes our system in detail. Section 3 presents the performance of our solution and analyze it.

2. SYSTEM DESCRIPTION

In this section, we introduce our system in detail. The predicting procedure contains the following three steps: First, select a set of features that represent music audio signal adequately. Second, apply a regression model that performs well in the range of ten thousand items and optimize the predicting results according to the relationship of continuous clips in a piece of music. Finally, considering the delayed reaction when people tag the label of music emotion, we investigate the proper time of delayed reaction. The three steps of our solution are shown as follows.

2.1 Feature Extraction

We preprocess the original audio files of the development data as follows: First, we transformed the music from *mp3* format to *wav* format. Second, segmented the music (15s to 45s period) into 60 clips, each with 500ms duration. Then we extracted features of each 500ms-clip.

*This work has been supported by the Natural Science Foundation of China(Multimodal Music Emotion Recognition technology research No.61170167)

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany

2.1.1 Mel-Frequency Cepstrum Coefficients

We divide the signals of songs into 50%-overlapping frames of 1024 samples length (about 23ms). We compute 13 Mel-Frequency Cepstrum Coefficients (MFCCs) with the 0th component included on each frame as a 13-D feature vector, as well as the delta-MFCCs .

2.1.2 Some General Short-term Features

Like MFCCs, we divide the signals of songs into 50%-overlapping frames of 1024 samples length (about 23ms). Then we compute Short Time Energy, Spectral Centroid, Spectral Entropy, Spectral Flux, Spectral Roll Off and Zero Cross Rate on each frame as a 6-D feature vector.

2.1.3 Edge orientation histogram on Mel Spectrogram

The spectrogram is a nearly complete representation of music, and furthermore, it provides a way for us to investigate the relationship between audio signal and emotion from a visual angle [7]. We find there exists strong relationship between the edge orientations in spectrograms and music emotions. We put forward our method by extracting EOH feature on audio spectrogram [8].

The procedure of our proposed algorithm can be described as follows: Convert the audio signal to the spectrogram with Mel time-frequency representations. The gradients at the point(x,y) in the Mel Spectrogram S can be found by convolving Sobel masks with S. Then we get edge orientation of each point on spectrogram by dividing the strength of Y dimension by that of X dimension. Finally, we index the edge orientations to a certain number of bins, which form edge orientation histogram on Mel Spectrogram.

Table 1: Development data results on various features, Fusion stands for the fusion of MFCCs, Short-term Features, EOH-MEL, OPENSMILE

	V		A	
Features	R^2	MSE	R^2	MSE
MFCCs+DMFCCs	0.4719	0.0662	0.4682	0.0621
Short-term Features	0.3828	0.0770	0.3787	0.0718
EOH-MEL	0.2705	0.0916	0.2088	0.0917
OPENSMILE	0.4873	0.0639	0.4514	0.0642
Fusion	**0.5159**	**0.0606**	**0.4803**	**0.0608**

2.1.4 Feature processing

An efficient and effective method of statistics for features of all the windows in a piece of music is to calculate the

Table 3: Official results on the test data

	V		A	
Run	RMSE	ρ	RMSE	ρ
1	0.3433± 0.1940	0.0016±0.4319	0.2410±0.1066	0.5243±0.3034
2	0.3669±0.1664	0.0086±0.3693	0.2567±0.0997	0.5025±0.2206
3	**0.3348±0.1868**	**0.0181±0.4350**	**0.2382±0.1052**	**0.5403±0.2694**

Table 2: The performance of predicting model by applying various lagging time

	V		A	
Lagging time	R^2	MSE	R^2	MSE
0ms	0.4867	0.0644	0.4507	0.0641
500ms	**0.4873**	**0.0639**	0.4514	0.0642
1000ms	0.4853	0.0642	0.4585	0.0633
1500ms	0.4801	0.0648	**0.4625**	**0.0629**
2000ms	0.4689	0.0662	0.4587	0.0633

means and variances. However, the windows of a piece of music construct a time series and the inner-connection between those windows cannot be revealed simply through means and variances. We, therefore, seek a proper way to reflect this connection in terms of time.

In this system, we build an Auto-Regressive (AR) and Moving Average (MA) Model to sort out the relationship between windows in terms of time. First of all, we analyze the features of all windows and sequence them in the light of time. Each dimension of the features forms an independent time series. Then, we gain new parameters by modeling those time series using the AR and MA Model. These parameters, together with means and variances, form the new features, among the 121 dimensions of which, means amount to 32 $(19 + 13)$ dimensions, variances 32 $(19 + 13)$ dimensions, AR model 19 dimensions and MA model 38 dimensions. Then we combine above features with EOH-MEL and OPENSMILE features to form the total features of 393 dimensions.

We evaluate the these features on the development set by splitting it into development and test set, while making sure that no samples from the same song are both in the development and test set. The following experiment conducted on the development set also takes the above method.

2.2 CCRF for dynamic task

Considering the emotion labels of adjacent scores in the same piece of music are time-continuous, we try to model them as an interrelated sequence. The model we employ is continuous conditional random field (CCRF). Conditional random field is used as a probabilistic graph model, which has the ability to express the long-range dependence and overlapping features, and can better solve the problem of the bias of the label, and all the features can be globally normalized, and the global optimal solution can be obtained. Notably in contrast to hidden Markov models (HMMs), CRFs do not need the independence assumption and Markov assumption, which is necessary for HMMs.

We adopted the CCRF model with SVR as the base classifier to model continuous emotions in dimensional space. We denote $\{x_1, x_2, \cdots, x_n\}$ as a set of labels predicted by SVR, and $\{y_1, y_2, \cdots, y_n\}$ as a set of final labels that we want to predict, $x \in R^m$ and $y \in R$. CCRF is defined as a conditional probability distribution over all emotion values. It can represent both the content information and the relation information between emotion values,which is useful for dynamic emotion evaluation[2].

2.3 Lagging time

When people tag the emotion scores for music, especially for the time-continuous clips of the music, they need the response time for receiving and processing sound, then tagging by hand. So we make an assumption that music clips do not correspond to the scores directly, but with a certain lag. Based on this assumption, we test on development set by varying the lagging time to find the best one. The experimental results are shown in Table 2 and we find that the lagging time for tagging V scores is about 500ms and for tagging A scores is about 1500ms, which is, however, inferred under the experimental conditions with the certain features and regression model of our choice, and needs more experiments to prove.

3. RESULTS AND CONLUSION

For CCRF, we set $n = 61$ for the training of the five runs, which means the number of the clips in one song, $q = 431$, i.e., the number of songs in development set.

Run 1 uses the given features extracted by OPENSMILE and the regression model of our choice, SVR+CCRF. Run 2 uses the features of our choice, the fusion of various features, and the given regression model Multiple Linear Regression (MLR). Run 3 uses both the features and the regression model of choice. We submitted these three runs and the results obtained by test dataset are shown in Table 3. We report the official challenge metrics, Pearson correlation (ρ) and Root-Means-Squared error (RMSE) for dynamic regression.

The results show that Run 3, which uses both the features and the regression model of choice, performs best. It means that our features and regression model performs better than features extracted by OPENSMILE and MLR. The RMSE of valence (V) and arousal (A) predicting are both in an acceptable range. However, we notice that the V predicting results gets a low ρ even close to 0, which looks strange compared with the high ρ of A predicting results. A possible reason is that V predicting is harder than A predicting. The fact that RMSE of V predicting results is lower than that of A predicting results also proves it.

4. REFERENCES

[1] Aljanaki, A., Yang, Y., Soleymani, M.: Emotion in Music Task at MediaEval 2014. In: MediaEval 2014 Workshop (2014)

[2] Baltrusaitis, T., Banda, N., Robinson, P.: Dimensional affect recognition using continuous conditional random fields. In: IEEE International Conference and Workshops, 1–8 (2013)

[3] Juslin, P.N., Sloboda, J.A.: Music and emotion: Theory and research. Oxford University Press (2001)
[4] Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. In: IEEE Transactions on Audio, Speech, and Language Processing, 14(1), 5–18 (2006)
[5] Fornari, J., Eerola, T.: The pursuit of happiness in music: Retrieving valence with high-level musical descriptors. In: the Computer Music Modeling and Retrieval (2008)
[6] Korhonen, M.D., Clausi, D., Jernigan, M.: Modeling emotional content of music using system identification. In: IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36(3), 588–599. (2005)
[7] Dennis, J., Tran, H.D., Li, H.: Spectrogram image feature for sound event classification in mismatched conditions. In: Signal Processing Letters, IEEE, 18(2), 130–133 (2011)
[8] Canny, J.: A computational approach to edge detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 679–698 (1986)
[9] Thayer, R.E.: The biopsychology of mood and arousal. Oxford University Press (1989)