Beatsens' Solution for MediaEval 2014 Emotion in Music Task*

Wanyi Yang[1], Kang Cai[1], Bin Wu[2], Ying Wang[2],
Xiaoou Chen[1], Deshun Yang[1], Andrew Horner[2]
[1]Institute of Computer Science and Technology,
Peking University, Beijing, China
[2]Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong, China
{yangwanyi, caikang}@pku.edu.cn, {bwuaa, ywangbf}@cse.ust.hk,
{chenxiaoou, yangdeshun}@pku.edu.cn, horner@cse.ust.hk

ABSTRACT

In this paper, we describe the Beatsens Team solution of *Emotion in Music* task in MediaEval benchmarking campaign 2014. We extracted and designed several sets of features and used continuous conditional random field(CCRF) for dynamic emotion characterization task. The best runs for Pearson correlation are 0.23 ± 0.56 and 0.12 ± 0.55 of valence and arousal respectively, for RMSE are 0.12 ± 0.06 and 0.09 ± 0.05.

1. INTRODUCTION

The *Emotion in Music* task aims to estimate valence and arousal values for 500ms music segments. In this 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 dependent on their previous segments. More details concerning the dataset collection can be found in [1]. Therefore, in our solution, we consider the labeling process as a continuous conditional random field (CCRF) process, where the valence-arousal(v-a) values not only depend on the music segments' acoustic contents, but also their preceding segments. The final results have also shown the advantages of CCRF modeling.

In this paper, we first introduce our solution in feature extraction and modeling. Then, we present the results in terms of both various feature combinations and model parameters.

2. SYSTEM DESCRIPTION

In this section, we introduce the feature design and model of our system. The basic logic of our system is that 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. We describe our solution in details as follows.

Table 1: Features extracted by MIRToolBox

Parts	Features	Dim.
Dynamics	RMS energy, Slope, Attack, Low energy	7
Rhythm	Tempo, Fluctuation peak, Fluctuation centroid	5
Spectral	Spectrum centroid, Brightness, Spread, Skewness, Kurtosis, Rolloff95, Rolloff85, Spectral Entrophy, Flatness, Roughness, Irregularity, Zero crossing rate,Spectral flux, MFCC, DMFCC	78
Harmony	Chromagram peak, Chromagram centroid, Key clarity, Key mode, HCDF	10

2.1 Feature Extraction

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. Features were extracted from the audio signal by MIRToolbox[1]. Both mean and standard deviations of the features were calculated. There were 54 features in total. Table 1 shows the features in detail.

2.2 CCRF for dynamic task

As labelers used a slide bar when labeling, emotion values change continuously but not mutationally, it is better to define the labeling model as a function on all the emotions in one song. We adopted the CCRF model with SVR as the base classifier to model continuous emotions in dimensional space.

In CCRF, 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].

*This work has been supported by the Natural Science Foundation of China(Multimodal Music Emotion Recognition technology research No.61170167) and Hong Kong Research Grants Council grants(HKUST613112).

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

[1] Version 1.5: https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

Table 4: Official results on the test data

Run	A		V	
	ρ	RMSE	ρ	RMSE
1	0.220±0.571	0.117±0.056	**0.124±0.546**	0.089±0.054
2	0.178±0.562	**0.107±0.055**	0.098±0.516	0.092±0.055
3	0.224±0.552	0.122±0.058	0.110±0.543	**0.086±0.055**
4	**0.231±0.564**	0.122±0.057	0.113±0.551	0.088±0.056
5	0.230±0.548	0.121±0.057	0.112±0.540	0.088±0.054

Table 2: Development data results on various clip length of MFCC, ALL stands for the feature consisting of 0.5s, 1s, 2s, 4s, 8s, COMB stands for combining the above six features' regression results as input of CCRF

	A		V	
Clip Length	R^2	MSE	R^2	MSE
0.5s	0.630	0.034	0.330	0.040
1s	0.618	0.034	0.317	0.041
2s	0.603	0.035	0.298	0.042
4s	0.585	0.037	0.283	0.044
8s	0.576	0.038	0.264	0.046
ALL	0.610	0.034	0.306	0.042
COMB	**0.638**	**0.032**	**0.346**	**0.039**

Table 3: Various frame length of MFCC, ALL stands for the feature consisting of 11.6ms, 23.2ms, 46.4ms, COMB stands for combining the above four features' regression results as input of CCRF

	A		V	
Frame Length	R^2	MSE	R^2	MSE
11.6ms	0.627	0.034	0.318	0.041
23.2ms	0.630	0.034	0.330	0.040
46.4ms	0.626	0.034	0.321	0.041
ALL	0.641	0.032	0.363	0.039
COMB	**0.646**	**0.031**	**0.371**	**0.038**

3. EXPERIMENTS AND RESULTS

With the selected attributes, we modeled the data using Support Vector Regression(SVR), K-Nearest Neighbor(KNN) and evaluated them on the training set with 4-fold cross validation. All of the results show that SVR outperforms KNN, so SVR is adopted in our runs.

For CCRF, we set $n = 61$ for the training of the five runs, which means the number of the clips in one song, $q = 744$, i.e., the number of songs in development set.

3.1 Experiments of Run1 and Run2

The 54 features are divided into four parts: dynamics, spectrum, rhythm, and harmony [3]. We compared the four perceptual dimensions and the combination of them, results showed that Spectral+Dynamic+Rhythm performs the best. This method is adopted in Run1.

With the features of Run1, we evaluated an SVR associated with three kernels: radial basis functions, linear and polynomial, and a series of $C(cost)$. Results showed that *Linear kernel* gives better result and $C = 2^{-3}$ performs best.

Because 500ms is too short for information extracting, some features failed to be extracted. Thus, we further extend the clip length to 1s and extract the features again. Finally we concatenate the new 1s-clip feature with original 500ms-clip feature to get the feature of Run2.

3.2 Experiments of Run3, Run4 and Run5

In addition, we found that Mel-frequency cepstral coefficient(MFCC) is one of the most important spectral features. As 0.5s is too short to convey the emotion completely, we made considerable experiments with MFCC by choosing various clip lengths and frame lengths.

Experiment a: We separately extracted MFCC of 0.5s, 1s, 2s, 4s, 8s clips to convey more information than a single 0.5s clip. The results are shown in Table 2. Comparing the six single features, the 0.5s clip performs best and this method is adopted in Run3.

For the combination, take six features' regression labels as input of CCRF and the final result outperforms the single 0.5s clip slightly, this method is adopted in Run4.

Experiment b: Considering frame length being an important parameter, we set different frame lengths (11.6ms, 23.2ms, 46.4ms), and extracted MFCC respectively. Table 3 shows that the results of different frame lengths remain basically unchanged, COMB performs the best. This method is adopted in Run5.

The results obtained by test dataset are shown in Table 4. We report the official challenge metrics, Pearson correlation(ρ) and Root-Means-Squared error (RMSE) for dynamic regression. We can conclude that such a simple set of feature as MFCC, performs even much better than more features. The combination of various clip lengths of MFCC perform the best, achieving a sufficiently good performance on a new dataset.

4. CONCLUSION

We have presented the Beatsens Team solution to the 2014 MediaEval Emotion in Music task. Best result on valence estimation was obtained by Run4, and best result on arousal estimation was obtained by Run1, they both used CCRF modeling. Further work will be conducted on feature selection and optimization of CCRF.

5. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2014. In *MediaEval 2014 Workshop, Barcelona, Spain*, October 16-17 2014.

[2] T. Baltrusaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[3] Y. Song, S. Dixon, and M. Pearce. Evaluation of musical features for emotion classification. In *ISMIR*, pages 523–528, 2012.