Multi-scale Approaches to the MediaEval 2015 “Emotion in Music" Task

Mingxing Xu, Xinxing Li, Haishu Xianyu, Jiashen Tian, Fanhang Meng, Wenxiao Chen
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, {lixinxing1991, xyhs2010}@126.com

ABSTRACT

The goal of the “Emotion in Music” task in MediaEval 2015 is to automatically estimate the emotions expressed by music (in terms of Arousal and Valence) in a time-continuous fashion. In this paper, considering the high context correlation among the music feature sequence, we study several multi-scale approaches at different levels, including acoustic feature learning with Deep Brief Networks (DBNs) followed a modified Autoencoder (AE), bi-directional Long-Short Term Memory Recurrent Neural Networks (BLSTM-RNNs) based multi-scale regression fusion with Extreme Learning Machine (ELM), and hierarchical prediction with Support Vector Regression (SVR). The evaluation performances of all runs submitted are significantly better than the baseline provided by the organizers, illustrating the effectiveness of the proposed approaches.

1. INTRODUCTION

The MediaEval 2015 “Emotion in Music” has only one task dynamic emotion characterization, including two required runs (one for feature extraction with linear regression, another for regression model with the baseline feature set provided by the organizers) and up to other three runs (any combination of the features and machine learning techniques) to permit a thorough comparison between different methods. In the task this year, the development data contains 431 clips with the best annotation agreement selected from the last year data and the evaluation data consists of 58 full-length songs. For more details, please refer to [1].

In order to predict and trace the evolution of music emotion more precisely, we investigated several multi-scale methods implemented at three different levels, including acoustic feature level, regression model level and emotion annotation level. For acoustic feature level, features were organized in groups according to their time scales and fundamentals. Deep learning algorithm was used to learn new features integrated multi-scale information about music emotion. Inspired by BLSTM-RNNs' capability of mapping sequence to sequence [3], we trained some BLSTM-RNNs with different length sequences and fused them using the extreme learning machine to produce the final prediction. In addition, we proposed a hierarchical regression to predict the global trend and local fluctuation of music dynamic emotion separately.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2. METHODOLOGY

2.1 Feature Learning

We used openSMILE toolbox to extract 65 Low-Level Descriptors (LLDs) with configuration `IS13_ComParE_lld` (see [9] for details) and divided them into 3 groups as follows: A) 26 LLDs related to `audSpec`; B) 29 LLDs related with `pcm-fftMag` and Mel-Frequency Cepstral Coefficient (MFCC); C) 10 LLDs related to voice. In addition, we adopted the idea proposed in [5] to extract Compressibility (comp), Spectral Centre of MASS (SCOM) and Median Spectral Band Energy (MSBE) at the local scale, and used MIR Toolbox [6] to extract 20 other features related to music attributes, including dynamic RMS energy, Tempo, Event Density, Spectrum centroid, Flatness, Irregularity, Skewness, Kurtosis, Rolloff85, Rolloff95, Spread, Brightness, Roughness, Entropy, Spectral Flux, Zero crossing rate, HCDF, Key mode, Key clarity and Chromagram centroid, and then assembled them as group D. The frame size was 60 ms for group C and 25 ms for other groups. In all groups, overlapping windows were used with a 10 ms step.

For features of each group in 1 s window with 0.5 s overlap, we calculated the mean, STD, slope and Shannon entropy functionals, delta coefficients together with the STD and slope functionals, and acceleration coefficients together with the STD functionals. This resulted in 4 feature sets with dimension 182, 203, 70 and 161, respectively.

Four different Deep Belief Networks (DBNs) were used to learn the higher representation for each group features independently, which were then fused by a special Autoencoder with a modified cost function considering sparse and heterogeneous entropy (details described in [11]), to produce the final features at a rate of 2 Hz for the succeeding regression.

2.2 Multi-scale BLSTM-RNNs Fusion

2.2.1 *Models training*

Considering the high context correlation among the music emotion feature sequence, we used bi-directional Long Short-Term Memory recurrent neural networks (BLSTM-RNNs) which worked quite well on numerous tasks involving sequence modeling in recent years [10, 7, 2, 8], to predict dynamic music emotion.

Separate BSLTM-RNNs were trained for arousal and valence regression. BLSTM-RNNs with 5 hidden layers (250 units per layer and direction) were used. The first two layers were pre-trained with whole development set (431 clips) and test set (58 songs). Training with learning rate 5E-6 was stopped after a maximum of 100 iterations or after 20 iterations without improving the validation set error. To alleviate over-fitting, Gaussian noise with zero mean and standard deviation 0.6 was added to the input activations, and sequences were presented in random order during training. All BLSTM-RNNs were trained with CURRENNT [1].

We trained 4 kinds of BLSTM-RNNs with different time scale (i.e. sequence length) of 60, 30, 20 and 10, respectively, on a training set containing 411 clips, and validated them on the remained 20 clips selected randomly according to the genre distribution of the test data (i.e. 58 complete songs). We totally made 5 different data partitions (411 clips for training, 20 clips for validation) and computed 3 trials of the same model each with randomized initial weights, among which the best one was selected. Hence, there were 5 different BLSTM-RNNs for each time scale.

2.2.2 *Model selection and fusion*

In order to select 4 models with different time scales for fusion, we applied two different criteria separately to compose two groups of 4 models. The first criterion was RMSE-first which just selected the model with the best RMSE for each time scale, while the second criterion was considering both the RMSE and the data partition to guarantee the training sets of the selected models for fusion be different from each other. In our experiments, there were 2 models shared by two groups; in other words, there were 6 unique models for fusion.

At the fusion step, we averaged the predictions produced by all 6 models as the final result. In addition to this simple fusion policy, we trained an Extreme Learning Machine (ELM) [4] for fusion. The input feature vector of ELM consisted of the original predictions of 4 different time scale BLSTM-RNNs, their delta derivatives and the smoothed values generated through a triangle-filter with length of 50. Two separate ELMs were constructed to fuse the corresponding predictions of the two model groups mentioned above. Finally, the outputs of two ELMs were averaged to produce the final emotion prediction.

2.3 Hierarchical Regression

The aim of hierarchical regression is to predict the global trend and local fluctuation of music dynamic emotion separately. Firstly, a global Support Vector Regression (SVR) was built to predict the mean of dynamic emotion attributes of whole song with 6373 song-level global features extracted using OpenSMILE toolbox with `IS13_ComParE` configuration (see [9] for details). Then, OpenSMILE with configuration `IS13_ComParE_lld` was used to extract 130 segment-level features whose means and standard deviations were calculated with 1s window and 0.5s shift to form local features to predict the fluctuation of dynamic emotion attributes for each 0.5s clip by a local SVR. Finally, each fluctuation value predicted by the local SVR and the mean value predicted by the global SVR were added to form the final emotion prediction for the corresponding 0.5s clip.

3. RUNS AND EVALUATION RESULTS

We submitted four runs for the task this year. The specifics of each run are as follows: Run 1) Multi-scale BLSTM-RNNs based Fusion with the simple average policy was performed with the baseline feature set. Run 2) Same as Run 1, but using ELMs for fusion. Run 3) Same as Run 1, but using new features learnt with the method described in Section 2.1. In all above runs, test data was segmented into fixed-length clips with 50% overlap according to the time-scale of BLSTM-RNNs specified. Run 4) The SVR based hierarchical regression described in Section 2.3.

Table 1: Official evaluation results on test data.

Dimension	System	RMSE	r
Valence	Baseline	0.366 ± 0.18	0.01 ± 0.38
	Run 1	0.331 ± 0.18	0.12 ± 0.54
	Run 2	0.308 ± 0.17	$\mathbf{0.15 \pm 0.47}$
	Run 3	0.349 ± 0.19	0.02 ± 0.51
	Run 4	$\mathbf{0.303 \pm 0.19}$	0.01 ± 0.40
Arousal	Baseline	0.270 ± 0.11	0.36 ± 0.26
	Run 1	$\mathbf{0.230 \pm 0.11}$	$\mathbf{0.66 \pm 0.25}$
	Run 2	0.234 ± 0.11	0.63 ± 0.27
	Run 3	0.240 ± 0.12	0.52 ± 0.37
	Run 4	0.250 ± 0.15	0.56 ± 0.24

In Table 1, we report the official evaluation metrics (r - Pearson correlation coefficient; and RMSE - Root Mean Squared Error). The results showed that all runs were significantly better than the baseline result. Considering the comprehensive performance, we observed that Run 2 was the best one. However, Run 2 was not better than Run 1 consistently, which indicated that ELMs might be trained insufficiently. Both Runs 1 and 2 worked particularly well in r, which was attributed to the BLSTM-RNNs' capability of mapping sequence to sequence. The reason why the new features in Run 3 did not make an expected improvement might be the low level features were not appropriate to represent different time-scales. Although the method in Run 4 was simple, it delivered comparable RMSE and r for Arousal among all runs, and performed quite well for Valence, but only in RMSE not in r, which may be related to the decomposition of global trend and local fluctuation. We believe it is a promising algorithm.

4. CONCLUSIONS

We describe THU-HCSIL teams approaches to the Emotion in Music task at MediaEval 2015. Several multi-scale approaches at three levels have been compared with the baseline system, including acoustic feature learning, multi-regression fusion and hierarchical prediction of emotion features. The results show that the proposed methods are significantly better than the baseline system, illustrating the effectiveness of the multi-scale approaches. In future work, we plan to investigate how to select the time scale automatically and systematically. In addition, the audio files of the test data in the pre-training stage of submitted Runs 1–3 may limit the generalizability of the trained model and some more evaluations are needed.

5. ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (No.61171116, 61433018), the National High Technology Research and Development Program of China (863 Program) (No. 2015AA016305).

[1] https://sourceforge.net/p/currennt

6. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[2] Y. C. Fan, Y. Qian, F. L. Xie, and F. K. Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *The 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[3] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5–6):602–610, June 2005.

[4] G. Huang, Q. Zhu, and C. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

[5] N. Kumar, R. Gupta, T. Guha, and C. Vaz. Affective feature design and predicting continuous affective dimensions from music. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, October 2014.

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

[7] H. Sak, A. Senior, and F. Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

[8] L. Sun, S. Kang, K. Li, and H. Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4869–4873, April 2015.

[9] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. On the acoustics of emotion in audio: What speech, music and sound have in common. *Frontiers in Psychology*, 4 (Article ID 292):1–12, May 2013.

[10] M. Wollmer, Z. X. Zhang, F. Weninger, B. Schuller, and G. Rigoll. Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[11] M. Xu and H. Xianyu. Heterogeneity-entropy based unsupervised feature learning for personality prediction with cross-media data. *submitted to The Thirtieth AAAI Conference on Artificial Intelligence*, 2016.