Emotion in Music Task at MediaEval 2015

Anna Aljanaki
Information and Computing Sciences
Utrecht University
the Netherlands
a.aljanaki@uu.nl

Yi-Hsuan Yang
Academia Sinica
Taipei
Taiwan
yang@citi.sinica.edu.tw

Mohammad Soleymani
Computer Science Dept.
University of Geneva
Switzerland
mohammad.soleymani@unige.ch*

ABSTRACT

The *Emotion in Music* task is held for the third consecutive year at the MediaEval benchmarking campaign. The unceasing interest towards the task shows that the music emotion recognition (MER) problem is truly important to the community, and there is a lot remaining to be discovered about it. Automatic MER methods could greatly improve the accessibility of music collections by providing quick and standardized means of music categorization and indexing. In the *Emotion in Music* task we provide a benchmark for automatic MER methods. This year, we concentrated on a single task, which proved to be the most challenging in the previous years: dynamic emotion characterization. We put special emphasis on providing high-quality ground truth data and maximizing inter-annotator agreement. As a consequence of meeting a higher quality demand, the dataset both for training and evaluation is smaller than in the previous years. The dataset consists of music licensed under Creative Commons from the Free Music Archive, medleyDB dataset and Jamendo. This paper describes the dataset collection, annotations, and evaluation criteria of the task.

1. INTRODUCTION

Contemporary music listeners rely on online music services such as Spotify, iTunes or Soundcloud to access their favorite music. In order to make their collections accessible, music libraries need to classify music by genre, instrumentation, tempo and mood. Automatic solutions to auto-tagging problem are invaluable because they make annotation fast, cheap and standardized. Emotion is one of the most important search criteria for music. Automatic MER (music emotion recognition) algorithms rely on ground truth for training. There are many ways in which such a ground truth can be generated [6]; using different affective representations or different temporal granularity. Depending on the affective model or temporal resolution, the evaluation criteria can vary. These discrepancies make it very difficult to compare different methods. The *Emotion in Music* task is designed to develop a benchmark and an evaluation framework for such a comparison.

The task is held for the third year in the MediaEval benchmarking campaign for multimedia evaluation[1] [1,14]. Building on our experience in the last two years, we concentrate on a single dynamic emotion characterization task and on offering high quality ground truth.

The only other current evaluation task for MER is the audio mood classification (AMC) task of the annual music information retrieval evaluation exchange[2] (MIREX) [8]. In this task, 600 audio files are provided to the participants of the task, who have agreed not to distribute the files for commercial purposes. However, AMC has been criticized for using an emotional model that is not based on psychological research. Namely, this benchmark uses five discrete emotion clusters, derived from cluster analysis of online tags, instead of more widely accepted dimensional or categorical models of emotion. It was noted that there exists semantic or acoustic overlap between clusters [12]. Furthermore, the dataset only applies a singular static rating per audio clip, which belies the time-varying nature of music. Since 2013, another set of 1,438 segments of 30 seconds clipped from Korean pop songs has been used in MIREX as well. However, the same five-class taxonomy is adopted for this Korean set.

Since the first edition of the *Emotion in Music* task in 2013 we have opted for characterizing the per-second emotion of music as numerical values in two dimensions — valence (positive or negative emotions expressed in music) and arousal (energy of the music) (VA) [13, 17], making it easier to depict the temporal dynamics of emotion variation. The VA model has been widely adopted in affective research [2, 6, 9–11, 15, 18–20]. However, the model is not free of criticisms and some other alternatives may be considered in the future. For example, the VA model has been criticized for being too reductionist and that other dimensions such as dominance should be added [5]. Moreover, the terms 'valence' and 'arousal' may be sometimes too abstract for people to have a common understanding of its meaning. Such drawbacks of the VA model can further harm the inter-annotator agreement of the annotations for an annotation task which is already inherently fairly subjective.

2. TASK DESCRIPTION

This year we offer only one task - *dynamic emotion characterization.* However, in order to permit a thorough com-

*This research was supported in part by Ambizione program of the Swiss National Science Foundation and the FES project COMMIT/. We thank Alexander Lansky, from Queens University, Canada and Yu-Hao Chin from National Central University, Taiwan for assistance with the song selection and annotations.

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

[1] http://www.multimediaeval.org
[2] http://www.music-ir.org/mirex/wiki/

parison between different methods, this year we require the participants to submit two different runs.

- In one run, the participants are required to submit their features and we use a baseline regression method (linear regression) to estimate dynamic affect. Any features automatically extracted from the audio or the metadata provided by the organizers are allowed.
- In the second required run, all the participants are required to use the baseline features that we provided (see Section 3 for details) to compare their machine learning methods. Participants are also free to submit any combination of the features and machine learning methods up to the total of five runs.

The participants will estimate the valence and arousal scores continuously in time for every segment (half a second long) on a scale from –1 to 1. The participants have to submit both predictions of valence and arousal, their feature set, if different from the basic provided one, and their predictions when using a universal feature set. We will use the Root-Mean-Square Error (RMSE) as the primary evaluation measure. We will also report the Pearson correlation (r) of the prediction and the ground truth. We will rank the submissions based on the averaged RMSE. Whenever the difference based on the one sided Wilcoxon test is not significant (p>0.05), we will use the averaged correlation coefficient to break the tie.

3. DATASETS AND GROUND TRUTH

Our datasets consist of royalty-free music from several sources: `freemusicarchive.org` (FMA), `jamendo.com`, and the medleyDB dataset [3]. The development set consists of 431 clips of 45 seconds, which were selected from last year's data based on inter-annotator agreement criteria. The test set comprises 58 complete music pieces with an average duration of 234 ± 105.7 seconds.

The development set is a subset of clips from last years [1, 14], all of which are from FMA. The subset was selected according to the procedure described below:

1. We deleted the annotations which Pearson's correlation with the averaged annotations for the same song is below 0.1. If less than 5 annotators remain after the deletion, we discarded the song.
2. For the remaining songs and remaining annotations, we calculated the Cronbach's α. If it was bigger than 0.6, the song was retained.
3. The mean (bias) of every dynamic annotation was changed to match the averaged static annotation for the same song.

This procedure resulted in a reduction from 1,744 songs to 431 songs (the rest did not have consistent enough annotations), each of which was annotated by 5–7 workers from the Amazon Mechanical Turk (MTurk). The Cronbach's α is 0.76 ± 0.12 for arousal, and 0.73 ± 0.12 for valence.

The evaluation set consists of 58 complete songs, one half from medleyDB dataset [3] of royalty-free multitrack recordings and another half from the `jamendo.com` music website, which provides music under Creative Commons license. We selected songs with some emotional variation in them from genres corresponding to the ones in the development set. We used the same annotation interface as the previous two years: a slider that is continuously moved by an annotator while listening to music. The position of the slider indicates the magnitude of valence or arousal.

	Arousal		Valence	
	RMSE	r	RMSE	r
openSMILE + MLR	0.27±0.11	0.36±0.26	0.37±0.18	0.01±0.38
Average baseline	0.28±0.13	–	0.29±0.14	–

Table 1: Baseline results.

The evaluation data we collected this year is different in several respects. First, we opted for full-length songs to cover the whole affective variation. Second, we partially annotated the data in the laboratory. The evaluation set is annotated by 6 people; two onsite and 4 conscientious MTurk workers, where 29% of the annotations was done in the lab. This way, we can compare the agreement between the onsite workers and the crowdworkers. The annotators listened to the entire song before starting with the annotation, to get familiar with the music and to reduce the reaction time lag. The workers were only payed the full fee after their work was reviewed and appeared to be of high quality. The Cronbach's α this year is 0.65 ± 0.28 for arousal, and 0.29 ± 0.94 for valence. In comparison, the Cronbach's α for another two existing datasets MoodSwing [16] and AMG1608 [4], is 0.41, 0.46 for arousal, and 0.25, 0.31 for valence, respectively. As compared to our dataset, the consistency of annotations has improved for arousal, but not for valence.

It can be found that, there is a mismatch between the training and test sets in terms of the duration of the clips (45-second segments versus full songs) and the data sources (FMA versus medleyDB and jamendo). In contrast, in either 2013 or 2014 the training and test sets were of the same length and both were from FMA [1, 14].

3.1 Baseline features

In order to enable comparison between different machine learning algorithms, we provide a baseline universal feature set, extracted with openSMILE [7], consisting of 260 low-level features (mean and standard deviation of 65 low-level acoustic descriptors, and their first-order derivatives). In addition to the audio features, we also provide meta-data covering the genre labels obtained from FMA, and, for some of the songs, folksonomy tags crawled from `last.fm`.

4. BASELINE RESULTS

For the baseline, we used the openSMILE toolbox [7] to extract 260 feature from nonoverlapping segments of 500ms, with frame size of 60ms with a 10ms step. We used multiple linear regression (MLR), following last years. The results are shown in the first row of Table 1. Compared to the last year (for arousal, $r = 0.27 \pm 0.12$, for valence, $r = 0.19 \pm 0.11$), the baseline is worse. We also calculated an average baseline by using the average of all the development set ground truth as the prediction result for all the songs. In terms of RMSE, this average baseline performs better for valence and at the same level for arousal.

5. CONCLUSIONS

A task has been developed to analyze emotion in music. Annotations were collected using both onsite annotator and crowdsourcing workers. The quest for higher quality labels has led to a lower number of training and evaluation samples.

6. REFERENCES

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

[2] M. Barthet, G. Fazekas, and M. Sandler. Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. In *Int'l Symp. Computer Music Modelling & Retrieval*, pages 492–507, 2012.

[3] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. MedleyDB: A multitrack dataset for annotation-intensive mir research. In *Proc. ISMIR*, 2014.

[4] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen. The AMG1608 dataset for music emotion recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 693–697, 2015.

[5] G. Collier. Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, 35(1):110–131, 2007.

[6] T. Eerola. Modelling emotions in music: Advances in conceptual, contextual and validity issues. In *AES International Confernece*, 2014.

[7] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of ACM MM*, pages 835–838, 2013.

[8] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, pages 462–467, 2008.

[9] A. Huq, J. P. Bello, and R. Rowe. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, 39(3):227–244, 2010.

[10] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2010.

[11] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis; using physiological signals. *IEEE Trans. Affective Computing*, 3(1):18–31, 2012.

[12] C. Laurier and P. Herrera. Audio music mood classification using support vector machine. In *MIREX Task on Audio Mood Classification*, 2007.

[13] J. A. Russell. A circumplex model of affect. *J. Personality & Social Science*, 39(6):1161–1178, 1980.

[14] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pages 1–6, 2013.

[15] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *IEEE Trans. Multimedia*, 16(4):1075–1089, 2014.

[16] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2011.

[17] R. E. Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, 1989.

[18] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. Modeling the affective content of music with a Gaussian mixture model. *IEEE Transactions on Affective Computing*, 6(1):56–68, 2015.

[19] S. Wang and Q. Ji. Video affective content analysis: a survey of state of the art methods. *IEEE Trans. Affective Computing*, PP(99):1, 2015.

[20] Y.-H. Yang and H.-H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intel. Systems & Technology*, 3(4), 2012.