Emotion in Music Task at MediaEval 2014

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

Emotional expression is an important property of music. Its emotional characteristics are thus especially natural for music indexing and recommendation. The *Emotion in Music* task addresses the task of automatic music emotion prediction and is held for the second year in 2014. As compared to previous year, we modified the task by offering a new feature development subtask, and releasing a new evaluation set. We employed a crowdsourcing approach to collect the data, using Amazon Mechanical Turk. The dataset consists of music licensed under Creative Commons from the Free Music Archive, which can be shared freely without restrictions. In this paper we describe the dataset collection, annotations, and evaluation criteria, as well as the two required and optional runs.

1. INTRODUCTION

Huge music libraries create a demand for tools providing automatic music classification by various parameters, such as genre, instrumentation, emotion. Among these, emotion is one of the most important classification criteria. This task presents many challenges, starting from its internal ambiguity and ending with audio processing difficulties [8]. As musical emotion is subjective, most existing work on MER relies on supervised machine learning approaches, training MER systems with emotion labels provided by human annotators. Currently, many researchers collect their own ground-truth data, which makes direct comparison between their approaches impossible. A benchmark is necessary to facilitate the cross-site comparison. The *Emotion in Music* task appears for the second time in the MediaEval benchmarking campaign for multimedia evaluation[1] and is designed to serve this purpose.

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) [1]. 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 [4]. Furthermore, the dataset only applies a singular static rating per audio clip, which belies the time-varying nature of music.

In our corpus we employ the music licensed under Creative Commons[3] (CC) from the Free Music Archive[4] (FMA), which enables us to redistribute the content. We do not use volunteers or online tag mining to collect the annotations, but pay the annotators to perform the task via Amazon Mechanical Turk (MTurk)[5], in a similar way as [2, 7]. We filter poor quality workers by making them first pass a test demonstrating a thorough understanding of the task, and an ability to produce good quality work. The final dataset spans 1744 clips of 45 seconds, and each clip is annotated by a minimum of 10 workers, which is substantially larger than any existing music emotion dataset with continuous annotations.

2. TASK DESCRIPTION

This year, similar to last year, the task comprises two subtasks. The first task is *dynamic emotion characterization* (main task). The second task, *feature design*, is introduced for the first time this year. New features, which have either not been developed before, or have not been applied to MER, should be proposed and applied to automatically detect arousal and valence for the whole song. The tasks will be trained on a development set of 744 songs and evaluated on a evaluation set of 1000 songs.

2.1 Run description

In *Subtask 1*, dynamic estimation, 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. In *Subtask 2*, feature design, the participants will develop new features and predict the valence and arousal scores of whole 45 second excerpts (on average, i.e. statically). Only one new feature will be evaluated in each run. For both tasks, together, each team can submit up to 5 runs, totally.

For the main (dynamic subtask) run, any features automatically extracted from the audio or the metadata provided by the organizers are allowed. For the dynamic emotional analysis we will use the Pearson correlation calculated per

[1] http://www.multimediaeval.org
[2] http://www.music-ir.org/mirex/wiki/
[3] http://creativecommons.org/
[4] http://freemusicarchive.org/
[5] http://mturk.com

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

song and averaged for the final value. We will also report the Root-Means-Squared Error (RMSE). We will rank the submissions based on the averaged correlations. Whenever the difference based on the one sided Wilcoxon test is not significant (p>0.05), we will use the RMSE to break the tie. The feature design task will be also evaluated based on the averaged across songs Pearson correlation and three runs. The participants can apply any non-linear transformation to their designed features to maximize the correlation.

3. DATASET AND GROUND TRUTH

For the description of the development set we refer to [5]. This year we collected more data in a similar way, but included external sources for metadata. We used the last.fm API to collect tags for matching songs from FMA. The songs that are already in last year's corpus were excluded. Then, we chose 1000 songs with the largest number of tags. Each song is from one or several genres from the following list: *Soul, Blues, Electronic, Rock, Classical, Hip-Hop, International, Experimental, Folk, Jazz, Country,* and *Pop*. We excluded songs from these genres: *Spoken, Old-time historic, Experimental* (in case the latter was the only genre that song belonged to). We also manually checked the music and excluded the files with bad recording quality or those not containing music, but speech or noise. For each artist, we selected maximally 5 songs to be included in the dataset.

To assure the adequate quality of the ground-truth, we created a procedure to select only the workers who are motivated and qualified to do the task, following current state-of-the-art crowdsourcing approaches [6]. All the workers had to pass a qualification test that was later evaluated manually. It consisted of three stages. Prior to the test, participants were provided with the definitions of arousal and valence, and could watch an instruction video. In the first stage, they listened to two short music audio clips, which contained distinctive emotional shift, and annotated arousal and valence continuously. In the second stage, workers described the emotional shift, and in the third stage, they described the song and indicated its genre. We also collected anonymized personal information from the workers, including, gender, age, and location, and asked them to take a short personality test.

Based on the quality of musical descriptions, and the correctness of their answers in the qualification task, we granted qualifications to the workers, after which they could proceed to the second step (the main task). The main task involved annotating the songs continuously over time once for arousal and once for valence, which in total constituted 334 micro-tasks. Each micro-task involved annotating 3 audio clips of 45 seconds on arousal and valence scales dynamically and statically, as a whole. The workers also characterized the song in emotional terms, and reported confidence of their answers, as well as familiarity and liking of the music. Workers were paid $0.25 USD for the qualification HITs and $0.40 USD for each main HIT that they successfully completed. On average, each HIT took 10 minutes.

To measure the inter-annotation agreement for the static annotations, we calculated Krippendorff's alpha on an ordinal scale. The values were 0.22 for valence and 0.37 for arousal, which are in the range of fair agreement. For the dynamic annotations, we used Kendall's coefficient of concordance (Kendall's W) with corrected tied ranks. Kendall's W was calculated for each song separately after discarding the annotations of the first 15 seconds. The average W is 0.2 ± 0.13 for arousal and 0.16 ± 0.11 for valence, which indicate weak agreement.

4. BASELINE RESULTS

For the baseline, we used MIRToolbox [3] to extract 5 features (spectral flux, harmonic change detection function, loudness, roughness and zero crossing rate) from non-overlapping segments of 500ms, with frame size of 50ms. We used multilinear regression as we did last year. For valence, the correlation averaged across songs was 0.11 ± 0.34 and RMSE: 0.19 ± 0.11. For arousal, the correlation was 0.18 ± 0.36 and RMSE was 0.27 ± 0.12. As compared to last year (for arousal, $r = 0.16 \pm 0.35$, for valence, $r = 0.06 \pm 0.3$), the baseline is higher. We also calculated the random baseline by averaging all the predictions. The RMSE for random average baseline is 0.18 ± 0.11 for valence and 0.21 ± 0.12 for arousal, which means that in terms of RMSE random baseline performs better.

5. ACKNOWLEDGMENTS

We are grateful to Sung-Yen Liu from Academia Sinica for helping with the task organization. This research was supported in part by European Research Area, the CVML Lab.[6], University of Geneva, and by the FES project COMMIT/.

6. REFERENCES

[1] 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.

[2] Y. E. Kim, E. Schmidt, and L. Emelle. Moodswings: A collaborative game for music mood label collection. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, pages 231–236, 2008.

[3] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects, Bordeaux*, 2007.

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

[5] 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*, CrowdMM '13, pages 1–6, New York, NY, USA, 2013. ACM.

[6] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010*, Geneva, Switzerland, 2010.

[7] 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.

[8] Y.-H. Yang and H. H. Chen. *Music Emotion Recognition*. CRC Press, Boca Raton, Florida, 2011.

[6] http://cvml.unige.ch