Emotion in Music task: lessons learned

Anna Aljanaki
University of Geneva
Geneva, Switzerland
aljanaki@gmail.com

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

Mohammad Soleymani
University of Geneva
Geneva, Switzerland
mohammad.soleymani@unige.ch

ABSTRACT

The Emotion in Music task was organized within MediaEval benchmarking campaign during three consecutive years, from 2013 to 2015. In this paper we describe the challenges we faced and the solutions we found. We used crowdsourcing on Amazon Mechanical Turk to annotate a corpus of music pieces with continuous (per-second) emotion annotations. To assure sufficient quality of the data, the annotation process on Mechanical Turk requires sufficient attention. Labeling music with emotion continuously proved to be a very difficult task for listeners, where both time delay and demand for absolute ratings degraded the quality of the data. We suggest certain transformations to alleviate the problems. Finally, the length of the annotated segments (0.5-1s) led to task participants classifying music on the equally small time scale, which only allowed them to capture changes in dynamics and timbre, but not musically meaningful harmonic, melodic and other changes, which occur on a larger time scale. LSTM-RNN based methods, which allow to incorporate larger context, gave better results than other methods, but still the proposed methods did not show significant improvement over the years and the task was concluded.

1. INTRODUCTION

The Emotion in Music task was first proposed by M. Soleymani, M.N. Caro, E.M. Schmidt and Y.-H. Yang in 2013 [9]. The task was targeting music emotion recognition algorithms for music indexing and recommendation, predominantly for popular music. The most common paradigm for music retrieval by emotion is the one when emotion is assigned to an entire piece of music. However, a piece of music exists in time and assigning just one emotion to a piece of music is most of the time incorrect. Therefore, music excerpts were annotated continuously using a paradigm that was first suggested for studying dynamics and general emotionality in music — Continouos Response Digital Interface (CRDI) [3]. The CRDI was later adapted by E. Schubert to record emotional response on valence and arousal scale [7]. In the first year of the task, static and dynamic tasks existed side by side. However, later the static task was dropped as less challenging. The decision to focus on continuous tracking on emotion had both pros and cons. On the bright side, it made the Emotion in Music benchmark very distinct from the existent Mood Classification task at MIREX[1] benchmark. The continuous emotion recognition is also arguably less researched than static emotion recognition. However, the pragmatic utilitarian aspect of the task valued in the MediaEval community became less prominent. There are much less applications and much less interest (at least currently) for automatic recognition of emotion varying over time.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands

2. CROWDSOURCING THE DATA

Music annotation with emotion is a time-consuming task, which often generates very inconsistent responses even with conscientious annotators. Therefore, it is difficult to verify the responses, because many inconsistencies can be attributed to individual perception. We used crowdsourcing (on Amazon Mechanical Turk (AMT)) to annotate the data, we paid the workers to annotate the music, and the workers had to pass a test before being admitted to the task. In the first 2 years, we did not monitor the quality of the work after the test was passed. We tried to estimate the lower bound of the number low-quality workers by only counting the people who did not move their mouse at all when annotating. Some of the songs may not have any emotional change in them, but at least some initial movement from the start position is necessary before stabilizing. In year 2014, 99 annotators annotated 1000 pieces of music, of them only 2 people did not move their mouse at all, and they annotated only a small amount of songs.

However, the agreement between the annotators was not very good both in 2013 and 2014 (less than 0.3 in Cronbach's α). In 2015, we changed the procedure to a more lab-like setting by hiring 5 annotators to annotate all the dataset, half of them in the lab and half on the AMT [1]. The quality was much better. This could also be attributed to the other changes, such as choosing full length songs, choosing the best annotators of the previous years, negotiating a fare compensation in advance on a Turker forum[2] and introducing a preliminary listening stage.

Despite having highly qualified annotators, the following problems were not resolved:

1. **Absolute scale ratings**. The ratings had to be given on an absolute scale while estimating changes in arousal and valence. Though the annotators often agreed on the direction of change, the magnitude of change was often different. We suggest shifting the annotation so

[1] http://www.music-ir.org/mirex
[2] http://www.mturkgrind.com/

that its mean is where the mean was indicated by the annotator (beforehand).

2. Human annotators have a **reaction time**. The biggest time lag is observed in the beginning of the annotation (around 13 seconds), but after every musical change a small time lag is also present. The beginnings of the annotations had to be deleted as unreliable.

3. **The time scale.** Some of the annotators would react to every beat and every note, and some annotators would only consider changing their arousal or valence at section bounds.

3. SUGGESTED SOLUTIONS

Participants received the data as a sequence of features and annotations with either half a second or one second frame rate. Many participants extracted their own features, but almost always the windows for feature extraction were smaller than the 0.5-1s, and the features were very low-level, mostly relating to timbral properties of the sound (energy distribution across spectrum) and loudness.

The best solutions in all the years were obtained using Long-Short Term Memory Recurrent Neural Networks. Although the arousal prediction performance improved over the three years (see Table 1), the accuracy obtained when predicting valence did not improve much (Table 2). It is a known issue that modeling valence in music is more challenging both due to the higher subjectivity associated with valence perception and in part due to the absence of salient valence-related audio features that can be reliably computed by state-of-the-art music signal processing algorithms [12, 5, 4]. The almost twofold improvement in arousal can also be attributed to the improvement in the quality and consistency of the data. In year 2015, the situation with valence became even worse, because we invested extra effort to assemble the data set in such a way, that valence and arousal would not be correlated (by picking more songs from the upper left ("angry") and lower right ("serene") quadrants). Because of the difference in the development and evaluation sets' distributions, the evaluation results were inaccurate in 2015. We trained an LSTM-RNN with the features supplied by the participants and evaluation set data. Using 20-fold cross-validation, we obtained more accurate estimation of the state-of-the-art performance on valence. The best result for valence detection on the test-set of 2015 was achieved using JUNLP team's features ($\rho = .27 \pm .13$ and $RMSE = .19 \pm .35$) [6]. JUNLP team used feature reduction to find the features which were most important for valence. However, the result is still much worse than the one obtained for arousal. A very interesting finding was that even though some sophisticated procedures for feature dimensionality reduction and BLSTM-RNNs were suggested by the participants, an almost equally good result could be obtained for arousal by using just eight low-level timbral features, and linear regression with smoothing.

4. FUTURE OF THE TASK

The major problem that we encountered when organizing the task was assembling good quality data. The improvement in performance over the years was partly dependent on that. The problems arising when using the continuous response annotation interface seem to be unsolvable unless either the task or the interface change.

Method	ρ	RMSE
2013, BLSTM-RNN [10]	$.31 \pm .37$	$.08 \pm .05$
2014, LSTM [2]	$.35 \pm .45$	$.10 \pm .05$
2015, BLSTM-RNN [11]	$.66 \pm .25$	$.12 \pm .06$

Table 1: Winning algorithms on arousal, ordered by Spearman's ρ. BLSTM-RNN – Bi-directional Long-Short Term Memory Recurrent Neural Networks.

Method	ρ	RMSE
2013, BLSTM-RNN [10]	$.19 \pm .43$	$.08 \pm .04$
2014, LSTM [2]	$.20 \pm .49$	$.08 \pm .05$
2015, BLSTM-RNN [11]	$.17 \pm .09$	$.12 \pm .54$

Table 2: Winning algorithms on valence, ordered by Spearman's ρ.

One of the possible solutions is to change the underlying task. It seems that the algorithms developed by the teams can track musical dynamics rather well. Many expressive means in music are characterized by gradual changes (e.g., *diminuendo, crescendo, rallentando*). Tracking these changes in tempo and dynamics could be useful as a preliminary step to tracking emotional changes. Changes in timbre can also be tracked in a similar way on a very short time scale.

Another possibility is changing the interface. One of the alternative continuous annotation interfaces suggested by E. Schubert uses categorical model instead of a dimensional one [8]. Using categorical model would eliminate the problem with absolute scaling.

A more sophisticated interface could also allow to modify the annotation afterwards by changing the scale (squeezing or expanding), removing the time lags.

At last, one of the major questions with the continuous emotion tracking task is its practical applicability. In most cases, the estimation of the overall emotion of the song, or the most representative part of a song, is most useful to users. Retrieval by continuous emotion tracking could be useful when a song with a certain emotional trajectory is necessary, for instance, for production music or soundtracks. Another possible application would be finding the most dramatic (emotionally charged) moment in a song to be used as a snippet. Moreover, as music is often used as a stimulus in the affective computing community to study emotion prediction by brain waves or physiological signals, a model to predict dynamic emotion in music can be helpful in this research. Departing from such bottom-up needs and requirements, hopefully the problem could be reformulated in a better way.

5. ACKNOWLEDGMENTS

We would like to thank Erik M. Schmidt, Mike N. Caro, Cheng-Ya Sha, Alexander Lansky, Sung-Yen Liu and Eduardo Countinho for their contributions in the development of this task. We also thank all the task participants and anonymous turkers for their invaluable contributions.

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*, 2015.

[2] E. Coutinho, F. Weninger, B. Schuller, and K. R. Scherer. The munich lstm-rnn approach to the mediaeval 2014 emotion in music task. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, 2014.

[3] D. Gregory. Using computers to measure continuous music responses. *Psychomusicology: A Journal of Research in Music Cognition*, 8(2):127–134, 1989.

[4] D. Guan, X. Chen, and D. Yang. Music Emotion Regression Based on Multi-modal Features. In *Symposium on Computer Music Multidisciplinary Research*, pages 70–77, 2012.

[5] C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen. Exploring Relationships between Audio Features and Emotion in Music. In *Proceedings of the 7th Triennal Conference of European Society for Cognitive Sciences of Music*, pages 260–264, 2009.

[6] B. G. Patra, P. Maitra, D. Das, and S. Bandyopadhyay. Mediaeval 2015: Music emotion recognition based on feed-forward neural network. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[7] E. Schubert. Continuous response to music using a two dimensional emotion space. In *Proceedings of International Conference of Music Perception and Cognition*, pages 263–268, 1996.

[8] E. Schubert, S. Ferguson, N. Farrar, D. Taylor, and G. E. McPherson. Continuous Response to Music using Discrete Emotion Faces. In *Proceedings of the 9th international symposium on computer music modeling and retrieval*, 2012.

[9] M. Soleymani, M. N. Caro, E. M. Schmidt, and Y.-H. Yang. The mediaeval 2013 brave new task: Emotion in music. In *Working Notes Proceedings of the MediaEval 2013 Workshop*, 2013.

[10] F. Weninger, F. Eyben, and B. Schuller. The TUM approach to the mediaeval music emotion task using generic affective audio features. In *Working Notes Proceedings of the MediaEval 2013 Workshop*, 2013.

[11] M. Xu, X. Li, H. Xianyu, J. Tian, F. Meng, and W. Chen. Multi-scale approaches to the mediaeval 2015 "emotion in music" task. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[12] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, 2008.