The MediaEval 2013 Brave New Task: Emotion in Music

M. Soleymani
Department of Computing
Imperial College London, UK
m.soleymani@imperial.ac.uk

M.N. Caro and E.M. Schmidt
Drexel University, USA
{mc947,eschmidt}@drexel.edu

Y.-H. Yang
Academia Sinica
Taiwan
yang@citi.sinica.edu.tw

ABSTRACT

Music is composed to be emotionally expressive. Emotional associations of music thus provide an especially natural feature for music indexing and recommendation. Emotion in Music Task is a brave new task addressing emotional characterization of music. In addressing the difficulties of emotion annotation we have turned to crowdsourcing, using Amazon Mechanical Turk. The dataset consists entirely of Creative Commons music from the Free Music Archive, which as the name suggests, can be shared freely without restrictions. In this paper, the dataset collection, annotations, and evaluation criteria as well as the two required and optional runs are described.

1. INTRODUCTION

The *Emotion in Music* task is a brave new task in the MediaEval 2013 benchmarking initiative for multimedia evaluation[1]. In seeking to develop tools for navigating today's vast digital music libraries, emotional associations provide an especially natural domain for indexing and recommendation. Because there are a myriad of challenges to such a task, powerful tools are required for the development of systems that automate the prediction of emotion in music. As such, a considerable amount of work has been dedicated to the development of automatic music emotion recognition (MER) systems [6]. Given the perceptual nature of human emotion, most existing work on MER has pursued supervised machine learning approaches, training MER systems using emotion labels or ratings entered by human subjects for a number of training clips.

The only current evaluation task for MER is the audio mood classification (AMC) task of the annual music information retrieval evaluation exchange[2] (MIREX) [1]. The audio files (totaling 600 clips) are available to the participants of the task, who have agreed not to distribute the files for commercial purposes. Being the only benchmark in the field of MER so far, this contest draws many participants every year. However, AMC describes emotions using five discrete emotion clusters instead of affect dimensions (e.g., valence and arousal). The clusters do not have origins in psychology literature, and some have noted semantic or acoustic overlap between clusters [3]. Furthermore, the dataset only applies a singular static rating per audio clip, which belies the time-varying nature of music.

Our new benchmarking corpus employs Creative Commons[3] (CC) licensed music from the Free Music Archive[4] (FMA), which enables us to redistribute the content. For annotations we have turned to crowdsourcing using Amazon Mechanical Turk (MTurk)[5], as others have found success using these tools to label large libraries [2,5]. In addition we have developed a two-stage procedure for filtering out poor quality workers, where workers must first pass a test demonstrating a thorough understanding of the task, and an ability to produce good quality work. The final dataset spans 1000, 45-second clips, and each clip is annotated by a minimum of 10 workers, which is substantially larger than any existing music emotion dataset.

2. TASK DESCRIPTION

This task comprises of two subtasks. In the first task, the dynamic emotion characterization task, the emotional dimensions, arousal and valence, should be determined for the given song continuously in time; the temporal resolution is one second. The second task, the static emotion characterization task, requires participants to deploy multimodal features to automatically detect arousal and valence for each song. We developed a dataset of 1000 songs which are split into the development set (700 songs) and the test set (300 songs). These affective features can be used in recommendation and retrieval platforms. There are already examples of mood based or emotion based online radios, e.g., Stereomood [6].

2.1 Run description

Our task comprises two tasks: *Subtask 1*, dynamic estimation: In this task, the participants will estimate the valence and arousal scores continuously in time. For every segment, which is 1 second long, valence and arousal scores between -1 and 1 should be estimated. Each team can submit up to 3 runs for this task. *Subtask 2*, static estimation: In this task, the participants will estimate the valence and arousal scores of the whole 45 seconds excerpt extracted from a song. Each team can submit 3 runs for this task

For both subtasks, and for the main run, any features automatically extracted from the audio or the metadata provided

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

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

[3] http://creativecommons.org/
[4] http://freemusicarchive.org/
[5] http://mturk.com
[6] www.stereomood.com

by the organizers are allowed. This is the required run. Optional runs, or general runs, include the possibility for the participants to use additional external data.

3. DATASET AND GROUND TRUTH

We downloaded the top 300 songs (ranked according to #listens) in MP3 format for each of the following eight genres: Blues, Electronic, Rock, Classical, Folk, Jazz, Country, and Pop. We did not consider other genres such as International, Novelty, Old-times, and Spoken because they are either ambiguous or contain non-music. We then excluded overly long (>10 minutes) and overly short (<1 minutes) songs, and picked the top 125 songs from each genre, leading to a dataset of 1,000 songs. We did not explicitly limit the number of songs contributed by each artist, but found 53-100 unique artists for each genre, providing a very good distribution across numerous recording artists.

Quality control is a key issue in crowdsourcing, and our strategy was designed following many current state-of-the-art crowdsourcing approaches [4]. A two-step approach was taken for worker recruitment. The first step was publishing the qualification task that consisted of a single micro-task or Human Intelligence Task (HIT) involving two songs. Participants were provided with the definitions of arousal and valence and they were asked to give their demography information, including, gender, age, location. Next, they were asked to play two short music audio clips which contained highly dynamic emotion shifts; they then indicated whether arousal and valence were increasing or decreasing, ideally demonstrating an understanding of the dimensional model. In addition, they were also asked to indicate the genre of the song using multiple choice check boxes. Finally, we asked the workers to write two to three sentences describing the clips they listened to, ideally demonstrating a willingness to put reasonable effort into a task.

Workers were chosen and qualifications were granted for the main task by considering the quality of their description and the correctness of their answers in the qualification task, i.e., first step. The second step, main task, involved annotating the songs continuously over time once for arousal and another time for valence; the main task involved a series of 334 micro-tasks. Each micro-task involved annotating 3 audio clips of 45 seconds on arousal and valence scales dynamically and static, as a whole. Workers were paid \$0.25 USD for the qualification HITs and \$0.40 USD for each main HIT that they successfully completed.

The static ratings given to the whole clips by the workers on both arousal and valence were averaged to serve as the ground truth for subtask 2. The dynamic annotation of the first 5 seconds of 45 seconds clips were discarded due to instability of their values. The arousal and valence dynamic annotation including 40 values corresponding to the last 40 seconds of the clips were averaged to generate the ground truth for the dynamic emotion estimation for subtask 1.

In order to measure the inter-annotation agreement, we calculated Krippendorff's alpha on an ordinal scale for the static annotations. The Krippendorff's alpha for the static annotations on the whole clips were 0.32 for valence and 0.35 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 5 seconds. The average W is 0.23 ± 0.16 for arousal and 0.28 ± 0.21 for valence. The observed agreement was statistically significant for arousal in 60.0% of songs and for valence in 65.8% of songs.

4. BASELINE RESULTS

The following features were extracted from audio signals: Mel-Frequency Cepstrum Coefficients (MFCC), octave-based spectral contrast, Statistical Spectrum Descriptors (SSDs) which is composed of spectral centroid, spectral flux, spectral rolloff, and spectral flatness in that order, Chromagram. The following features were extracted using Echonest[7] API: timbre, pitch, and loudness features.

A Multivariate Linear Regression (MLR) was selected for the baseline system because it is a simple and generalizable prediction method. The MLR was trained on the development set and evaluated on the test set. All the annotations including for the static and dynamic ones were scaled between $[-0.5, 0.5]$. The Euclidean distance between the estimated arousal and valence points as well as R^2 were calculated for the evaluation of the static results. To evaluate the dynamic results, mean distance and Kendall's Tau ranking correlation were used. The average values of arousal and valence on the training set was chosen as the random level baseline to be compared with our results. To evaluate the estimation models from content features R^2 and mean absolute error (distances) are reported for static estimation and Kendall Tau (τ) is reported with distance for dynamic estimation. The reported measures on dynamic annotated data are averaged for all the clips. Random level results are calculated by setting the target to the average score in the training set. The results that are significantly better (Wilcoxon test $p < 0.01$) than the random level were the arousal static estimation, $Distance = 0.10 \pm 0.07$, $R^2 = 0.07$, and arousal dynamic estimation, $Distance = 0.08 \pm 0.05$, $\tau = 0.15 \pm 0.22$. On the estimation of static ratings, the arousal estimations are far better than valence estimations which are in the order of chance level. Consistently, arousal estimation results are superior to valence estimation on the continuous, dynamic affect estimation task.

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

[4] 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.

[5] 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.

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

[7] http://www.echonest.com/