Ground truth for automatic music mood classification

Janto Skowronek, Martin F. McKinney, Steven van de Par
Philips Research Laboratories
Hightech Campus 36, 5656 AE Eindhoven, The Netherlands
{janto.skowronek,martin.mckinney,steven.van.de.par}@philips.com

Abstract

Automatic music classification based on audio signals provides a core technology for tools that help users to manage and browse their music collections. Since "mood" is also used as a browsing criterium, automatic mood classification could support the creation of the necessary metadata. We have developed a method to obtain a reliable "ground truth" database for automatic music mood classification. Our results confirm that excerpt selection is a non-trivial issue and that there are some mood labels that are relatively consistent across subjects.

Keywords: music mood classification, ground truth.

1. Introduction

Lu et al. [1] discussed that automatic mood classification can be criticized because the emotional meaning in music is highly subjective. However, they also stressed that there is a certain agreement on the music's mood and they showed that mood classification is possible. In addition, there is a strong application-oriented interest in mood classification: music download services [2] or audio players [3] allow music collection browsing using "mood" as one search criterium. Automatic mood classification could decrease the effort in providing the necessary metadata.

When developing a classification system, the definition and collection of a proper ground truth is crucial. From a classification point of view, the defined classes should be internally consistent. From an application point of view, users should have a clear and common understanding of the classes and the data in the classes should reflect the users' opinions. Both perspectives have two issues in common: class definition and material selection. In a study focussing on mood classification we addressed, therefore, two questions: How do we select music tracks that provoke or articulate emotions well enough emotions such listeners can assign a mood label to the track? How do we select labels that listeners can easily use for describing the music's mood and on which subjects agree when assessing individual music pieces?

2. Background

Lu et al. [1] set up a mood classification system which defined four mood categories, which were derived from a two-dimensional model of affect [4]. For the track selection, Lu et al. followed an expert-based approach: A music excerpt (western classical music) was appended to the ground truth only if three experts agreed on the mood.

In another study, Leman et al. [5] used 15 bipolar adjective pairs as mood descriptions selected by literature scan and trial experiments. Using a factor analysis on the gathered subjective data, they identified an underlying three dimensional space. Then they projected subjective mood assessments of music tracks onto that space and used linear regression in order to predict these projections with audio features computed from the corresponding music excerpts. Leman et al. used a larger set of mood labels than Lu et al. but they did not try to directly predict the mood labels, but rather their projections in the found three-dimensional space. With respect to track selection, Leman's approach was user based: 20 people were asked to propose music in which they recognize an emotional affect and to describe it, given no constraints about musical style.

Both studies used different mood labels as well as different methods for track selection. Regarding the final goal of developing a mood classification system and due to our experience from two pilot experiments, we saw advantages and disadvantages in both approaches; thus we decided to combine them into our own experimental method.

3. Experiment

In line with our research questions, the experiment design focussed on a careful selection of music excerpts and a broad search for proper mood labels.

Excerpt selection: Since complete music pieces can contain sections with different moods [1], we chose 20-second long excerpts from around the middle of 470 music tracks from 12 music genres. The selection was done manually such that the mood was likely to be constant within the excerpt by avoiding drastic changes in the musical characteristics (structure, loudness, instrumentation / timbre, tempo etc.) Then the 1^{st} author evaluated these excerpts using nine bipolar mood scales, which comprised the highest factor loadings from [5], the two axes of the affect model [4] and some additional considerations. For each of the 12 music genres we then chose the tracks with the five most extreme ratings, aiming to obtain a selection of excerpts with "easy to evaluate" moods.

Label definition: When defining the mood scales used for the experiment, we addressed two issues: **(1)** In a first pilot experiment subjects reported difficulties in assessing the mood with the two affect model axes used by Lu et al. [1]. Another pilot experiment showed low agreement across subjects for most of the nine bipolar mood scales mentioned in the previous subsection. Since we were interested in mood labels that are "easy to use" as well as consistent across subjects, we decided for a broad search for proper "direct" mood labels and not to restrict to the axes of an underlying mood space. Our collection of 33 adjectives covered the main directions of Leman's mood space (highest factor loadings in [5]) as well as the axes and the with 45 degrees rotated axes of the affect model [1, 4, 6], augmented by labels already used in applications ([2, 3] and some additional considerations. **(2)** In the pilot experiments we experienced that the exact interpretation of the labels influenced the judgments. In order to confine the meaning of the scales, we provided up to three synonyms of adjectives for each scale, which in several cases coincided with a merging of similar scales from different sources ([1] - [6]). Since most of the potential subjects were not native English speakers, four native-speakers of other languages (NL, D, F, I) provided translations of the labels, and the chosen subjects had to have one of these languages as mother tongue.

Practical experiment design: We collected data on 33 scales (7 points: strongly disagree to strongly agree) for 60 tracks with two repetitions (⇒ 3960 individual evaluations). To reduce experiment time, we distributed the tracks across 10 subjects such that a) each track is evaluated by four subjects, b) each subject has a different set of 24 tracks and c) the 12 music genres are equally represented in the set. In addition subjects were polled with questionnaires on their preferences with respect to music mood and on the ease-of-use of the provided labels.

4. Data analysis

We looked at various aspects such as influence of external factors and global mood, perceptual space spanned by the labels, similarity of labels, consistency within subjects etc. However, we focussed primarily on those results regarding the two main questions of this study: the proper selection of music tracks and the collection of useful mood labels.

Excerpt selection: In order to check how well the selected excerpts were "easy to judge", we averaged per track and label the ratings across subjects and sessions and computed a) the number of excerpts that got a *strong* judgment for at least one label, and b) the average number of *strong* judgments per track. Defining *strong* judgments as "agree" & "strongly agree", 31 tracks got at least one strong rating and on average every track got strong ratings in 1.47 labels.

Consistency across subjects: Due to the experiment design, blocks of 12 excerpts have been evaluated by the same four subjects. For each of these blocks (5 in total) we computed the Cronbach coefficient α [7] per label as a consistency measure across subjects. Then we averaged the α's across the blocks and identified 10 labels that had values between 0.7 and 0.84, which were in the range of the minimum value indicating "acceptable consistency" [7].

Importance and easiness of mood labels: Analyzing the experiment questionnaires (mean value across subjects, rounded to closest category), 20 mood labels were at least "important" for subjects; 16 of them were at least "easy to use".

Best labels: Interesting candidates for a mood classification system are those which belong to the "most consistent" group and which were assessed at least as "important" and at least as "easy to use". We found that such labels exist, they were: tender/soft, powerful/strong, loving/romantic, carefree/lighthearted, emotional/passionate, touching/moving, angry/furious/aggressive, sad.

5. Conclusions

We investigated how we can obtain a proper ground truth for a music mood classification system.

With respect to excerpt selection, only half of all excerpts received strong judgements. That means a proper choice of "easy to judge" excerpts is a non-trivial task, even for an experienced listener. Possibilities for optimization include increasing the number subjects in the selection process and extending the number of candidate tracks.

With respect to selection of mood labels, we were able to identify eight labels that were relatively consistent across subjects, perceived as easy to use and regarded as important. These are good candidates for a ground truth of an automatic mood classification system.

References

[1] L. Lu, D. Liu, H.-J. Zhang, *Automatic Mood Detection and Tracking of Music Audio Signals*, IEEE transactions on audio, speech, and language processing, Vol. 14(1), 5-18, 2006.

[2] http://www.allmusic.com

[3] http://www.moodlogic.com

[4] J.A. Russell, *A circumplex model of affect*, J. Personality & Social Psychology, Vol. 39, 1161-1178, 1980.

[5] M. Leman, V. Vermeulen, L. De Voogdt, D. Moelants, M. Lesaffre, *Prediction of Musical Affect Using a Combination of Acoustic Structural Cues*, J. of New Music Research, Vol. 34(1), 39-67, 2005.

[6] D.A. Ritossa, N.S. Rikkard, *The relative utility of 'pleasantness' and 'liking' dimensions in predicting the emotions expressed by music*, Psychology of Music, Vol. 31(1), 5-22, 2004.

[7] J.M. Bland, D.G. Altman, *Statistics notes: Cronbach's Alpha*, BMJ, Vol. 314, 572, 1997.