Improving Music Emotion Labeling Using Human Computation

Brandon G. Morton, Jacquelin A. Speck, Erik M. Schmidt, and Youngmoo E. Kim
Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104
{bmorton, jspeck, eschmidt, ykim}@drexel.edu

ABSTRACT

In the field of Music Information Retrieval, there are many tasks that are not only difficult for machines to solve, but that also lack well-defined answers. In pursuing the automatic recognition of emotions within music, this lack of objectivity makes it difficult to train systems that rely on quantified labels for supervised machine learning. In recent years, researchers have begun to harness Human Computation for the collection of data spanning an excerpt of music. MoodSwings records dynamic (per-second) labels of players' mood ratings of music, in keeping with the unique time-varying nature of musical mood. Players collaborate to build consensus, ensuring the quality of data collected. We present an analysis of MoodSwings labels collected to date and propose several modifications for improving both the quality of the gameplay and the collected data as development moves forward.

Categories and Subject Descriptors
E.0 [Data]: General; H.1.2 [Information Systems]: Models and Principles

General Terms
Human Factors, Measurement

Keywords
human computation, music, mood prediction, moodswings, mood labels, emotion, valence, arousal

1. INTRODUCTION

The problem of automated mood detection of music is particularly challenging in that it attempts to quantify highly subjective human perceptions, requiring collection of ground truth data directly from humans. In recent years, Music Information Retrieval (Music-IR) researchers have attempted to harness human computation to collect ground truth data more efficiently by developing games similar to those designed for other purposes, such as computer vision [1]. Music labeling games, however, involve unique challenges due to the time-varying nature of audio.

In this paper we present an analysis of MoodSwings', our game designed to collect “ground truth” data labeling music by mood [2]. The game is targeted at collecting dynamic (per-second) labels of users’ mood ratings in real-time using a two-dimensional space of emotional components. We use the Arousal-Valence (A-V) space, a well-validated representation of mood from psychology research proposed by Thayer and Russell [3]. As in other collaborative games, players are partnered in order to verify each others’ results, providing a strong incentive for producing high-quality labels. Accordingly, scoring is designed to maximize consensus-building between partners.

We propose several improvements to MoodSwings suggested by an analysis of data collected thus far. These improvements could both produce more informative ground truth data and increase the amount of data collected by encouraging more people to play MoodSwings.

2. BACKGROUND

The success of data collection games like the ESP Game for image labeling has inspired Music-IR researchers to develop games for labeling audio. Note that the games described below are designed for collecting static labels, such as musical genre or artist, and do not account for time-varying labels. When working with audio data, which is by nature time-varying, this consideration can be crucial.

One of the earliest games developed for music label collection was MajorMiner [4]. While it does not emphasize collaboration as much as its successors, MajorMiner rewards consensus on the “tags” (i.e., labels) it collects by awarding points based on how often the tags players generate are used by others. Agreement between players can occur asynchronously, as players continue to earn points between games if others use their tags. An added benefit is preventing what is commonly referred to as the “cold start” problem, resulting from both an initial lack of data and lack of players to produce data. New song clips can be introduced systematically, reducing the number of clips in the corpus from which no useful inferences can be drawn due to lack of data.

Listen Game expands on MajorMiner’s model of entirely free-form label assignment, including both a free-form “freestyle” mode and a “normal mode” in which players are

1MoodSwings: http://schubert.ece.drexel.edu/moodswings
presented a predetermined tag vocabulary [5]. Players mark randomly chosen labels from this vocabulary as either “good” or “bad”, simplifying analysis by removing the need to consider redundant or nonsensical labels assigned by players. TagATune collects free-form tags, but employs a mechanism to discourage players from assigning nonsense labels. Partnered players describe the songs they hear to each other, then determine whether or not they are listening to the same song [6].

Herd It, the most recently developed Music-IR ground truth collection game, emphasizes agreements within a large group rather than between partnered pairs of players [7]. It consists of several mini-games in which groups of players select one of several labels provided for each song clip. Mini-games are designed to collect ground truth for a variety of Music-IR tasks, including genre and mood classification (e.g.: “which of the following colors best fits this song clip?”) Herd It is also unique in that it runs on the popular Facebook social networking website, which attracts user traffic. Facebook also facilitates the collection of statistics describing demographics of the Herd It user base, adding additional context to data collected based on agreements between players.

3. MOODSWINGS

MoodSwings is a collaborative game incorporating two players’ judgments of the moods of songs into gameplay [2]. The fundamental difference between MoodSwings and other Music-IR games is its collection of time-varying labels, which is particularly significant in the realm of audio because the signals themselves are time-varying. Players are partnered anonymously over the internet, with the goal of dynamically and continuously reaching agreement on the mood of five 30-second song clips drawn from a music database. The game board, shown in Figure 1, is a representation of the A-V space. It presents a continuum of mood ratings: valence, reflecting positive vs. negative emotion, is displayed on the vertical axis and arousal, reflecting emotional intensity, is on the horizontal axis.

Figure 1: Screenshot of MoodSwings gameplay. The red and yellow orbs represent each player.

3.1 Gameplay and Scoring

A MoodSwings game consists of 5 rounds, each with a different 30-second song clip. Partners are paired at the beginning of each game for its entire duration. Each player positions his or her cursor on the game board to indicate judgment of the mood of the song at each instant during the game. Scoring is based on the amount of congruency between the two cursor positions, with greater overlap resulting in more points earned. Partners’ cursors are visible only once every 3 seconds. However, cursor size decreases over time, increasing the difficulty of overlapping cursor positions. Players can also accumulate bonus points by “convincing” partners to agree with their mood assessments, providing incentive against capricious movements. If a player keeps his or her cursor position stationary for more than 1 second and their partner moves to the same position, 5 bonus points are awarded, indicated by a change in cursor color. The point system is designed so that a “good” score for a single round is about 100 points, or about 500 points for a game. High scores for the top five players and the top individual match scores are visible when a player first logs onto the MoodSwings website.

3.2 Music Corpus and Data Collected

MoodSwings draws from 7,135 songs from the uspop2002 database, developed by the the NEC Research Institute and the Laboratory for the Recognition and Organization of Speech and Audio (LabROSA) [8]. The database contains songs from 400 artists representative of best-selling music in the year 2002.

To date, a total of 5,177 MoodSwings games have been played. Over 150,000 mood labels have been collected for 1,158 songs, producing about 16% coverage of all available songs.

3.3 Application of Data

Some of our recent work uses MoodSwings data to predict a 2-D Gaussian distribution of emotional ratings within the A-V space for each 1-second time slice. We have explored several methods for multi-variate parameter regression as well as a variety of acoustic features. Shown in Figure 2 are the collected A-V labels and distribution projections resulting from regression analysis. Performance of the regression can be evaluated by the amount of overlap between a projection and its ground truth. This work has also been used to improve to the single player version of MoodSwings, which

Figure 2: Collected A-V labels and distribution projections from regression analysis.
is described in later sections of this paper.

4. ANALYSIS OF DATA

MoodSwings was designed to be played by two people simultaneously, which required accounting for the possibility of only one player being online at a given time. MoodSwings originally paired players with either a past recording of an earlier game or a simulated AI player, which randomly follows a real player’s movements. We hypothesized that enough data collected from human players would negate the effects of simulated players, but overcoming the cold start problem has proven unexpectedly difficult. It is necessary to attract many new players and encourage past users to play again, which we hope to accomplish through improvements to make the game more challenging and exciting. To better understand which aspects of the game should be improved, we examine data collected during each gameplay scenario.

4.1 Analyzing Labels vs. Time

Figure 3 shows the second-by-second distribution of all labels collected from MoodSwings. Most movement of players’ cursors occurs at the beginning of a song clip, decreasing over time. We see the distribution becomes less clustered around the origin between about 7-8 seconds. Figure 4 shows the mean distance from the origin of the A-V space to the collected labels, with a square marker indicating 85% of the total distance traveled at about 8 seconds. We infer that most players are unable to instantaneously judge the mood of a clip at its onset. This is consistent with other research by Bachorik, which concluded that most music listeners require about 8 seconds to judge the mood of a song [9].

We also see a difference in comparing games with real players vs. the random AI. More distance is traveled in games against the AI, whereas in head-to-head games players tend to compromise, pulling each others’ cursors more towards the origin.

4.2 Analyzing Live vs. Recorded Players

Figure 5 shows the mean distance between the two cursors for two possible gameplay configurations: head-to-head (two players online) and human-vs.-recording (one person plays against a recording of a past game). On average, head-to-head games have a smaller distance between the cursors, which indicates greater agreement. There is a 3-second periodic oscillation, particularly pronounced in the human-vs.-recording case. The amount of agreement here appears to be heavily dependent on whether or not the recorded cursor is visible to the human player (every 3 seconds). In contrast, the head-to-head case shows a less pronounced oscillation, more fluid agreement, and overall greater consensus between live players. All of these characteristics demonstrate agreement being reached faster and in a more consistent manner, further supporting preference for the head-to-head scenario.

4.3 Prior Distribution of Pop Music
The second-by-second distributions seen in Figure 3 also reflects a prior distribution of mood labels associated with popular music. Player responses tend to be clustered in the upper-right quadrant. While this prior distribution is not necessarily problematic when only considering one particular musical genre (it can be useful for optimizing classification), collecting ground truth data on pop music alone will not necessarily lead to inferences about the wide range of moods expressible by all musical genres.

5. PLANNED IMPROVEMENTS

We hope to use the presented analysis to supplement the use of recent, ongoing work on emotion prediction to improve the game itself. Our aim is to collect more useful data and produce a more enjoyable gameplay experience.

5.1 Encouraging Replay

The cold start problem is partially caused by the fact that many people play MoodSwings once, then never again. We hope to encourage replay by making the game more fun. Fun is difficult to quantify, but we hope that improving the AI and increasing the level of challenge will increase game traffic.

In single player games where players were paired with a simulated player, moves made by the original AI made it obvious to the human player that he or she was not partnered with another person. Recent work to predict time-varying distributions of MoodSwings labels has been used to implement a predictive AI in the single player version of the game, MoodSwings Single Player. The AI chooses its mood labels for a given song clip from a predicted mood distribution, making it behave more realistically. MoodSwings Single Player is designed as a feedback system to address the cold start issue: predicted mood distributions shape moves by the AI, which is used to obtain more data from human players to be fed back into the training of the predictive AI.

We are also testing a version of the game using shorter song clips with no breaks in between, for more fast-paced and exciting gameplay. We must, however, be careful not to shorten song clips to the point where players are unable to determine the mood of the song before the end of the clip (> 8 seconds).

5.2 Expanding the Music Corpus

As described earlier, the music comprising the MoodSwings corpus does not include the most current popular music (the most recent tracks are from 2002). We eventually hope to incorporate more current music into the corpus to make the game more entertaining, particularly for younger players.

Other interesting problems related to mood detection could be explored by incorporating other genres besides popular music into the corpus. Pop music is somewhat limited for two reasons: first, the moods of pop songs tend to change seldom or not at all over their durations. Second, pop songs almost always contain lyrics, which influence the perceived mood. We hope to use MoodSwings to collect ground truth for instrumental music, including jazz and classical pieces. We hypothesize that mood labels for instrumental music pieces would more evenly span the A-V space.

2 MoodSwings SP. http://music.ece.drexel.edu/mssp

6. DISCUSSION AND FUTURE WORK

We have proposed improvements to address the cold start problem and enhance the gameplay experience. In the future, the game could benefit from updated animations to produce a more modern-looking interface. Displaying players' statistics during gameplay would allow score comparisons, adding a competitive aspect to the game. One possibility for increasing the challenge level is to increase the number of players collaborating during gameplay. The level of challenge is increased due to multiple players striving for consensus. Finally, more players could be attracted to the game if it were available on a wider variety of platforms. Facebook could be an attractive platform for enhancing the social and collaborative aspects of MoodSwings. The game could also be successful as a mobile application on handheld devices.

7. ACKNOWLEDGMENTS

This work is supported by National Science Foundation award IIS-0644151.

8. REFERENCES