

Addressing Problems in the First Adaptive System Evaluation Challenge

David N. Chin

University of Hawaii
Department of Information & Computer Sciences
1680 East West Rd, POST 317
Honolulu, HI 96822 USA
chin@hawaii.edu

Abstract. The adaptive evaluation challenge system has several problems with assumptions including ignoring other emotions, ignoring other influences on happiness and assuming ratings of 5.5 have neutral happiness impact. The recommended evaluation combines established affective response to music surveys and heart rate monitoring to measure happiness. First, ratings are recalibrated by asking users whether music clips affect their emotions positively, neutrally, or negatively. Clips that affect emotions other than happiness are filtered out and a pilot study is used to determine how many positive/negative clips are needed to maximize/minimize Happiness. Finally a custom series of clips are designed for each user in the study following standard DJ music selection practices. The measured/reported happiness of the users is compared to the proposed Happiness models to determine the best model. Ethics of the study are also discussed.

1 Introduction

The First Adaptive System Evaluation Challenge presents an evaluation challenge adaptive system (ECAS for short) that adaptively selects a sequence of music clips to play to a group of users to maximize the overall Happiness of the group. It uses precompiled ratings for each music clip by each user for selection and hypothesizes that any particular user's Happiness can be modeled as positively impacted by higher rated music clips and negatively by lower rated music clips. The goals of the adaptive system evaluation challenge are to determine which of three models best predict Happiness with the models varying in assumptions including whether Happiness is bounded and whether current mood affects the user's reaction to a music clip.

2 Problems

There are several problems with the assumptions of the ECAS. First, the ECAS assumes that users will feel happy when listening to music that they rate highly. In

fact, music engenders a wide range of emotional responses in listeners other than happiness, include opposing emotions such as sadness, dislike, anger, hopelessness, fear, remorse, fears confirmed, shame, and reproach [1]. Also, emotional response to music is very individualistic and can be colored by the person's idiosyncratic associations with the music. For example, I feel very sad whenever I hear what most people would consider a happy song, Israel "Iz" Kamakawiwo'ole's very beautiful "Somewhere over the Rainbow, What a Wonderful World" because of the singer's unfortunate death from overweight shortly after the song's debut (a tragic waste of a wonderful talent from a treatable eating disorder). Others might rate a happy music clip highly, yet feel sad/angry because the song is associated with a former girl/boyfriend, deceased loved one or a negative event.

Another problem with the ECAS is the assumption that Happiness is based purely on current Happiness and the Impact of the music clips. In fact, response depends on many other factors. For example, if a music clip is associated with the user's spouse, then response to the clip may depend on the current status of the relationship. Even the quality of the music playback system may annoy or even enrage some users who consider the poor sound quality an insult to their favorite music. For the ECAS, group dynamics, which change constantly, can influence the response of users to the music clips. Also, previous music clips influence response. For example, a well-liked slow-paced clip may engender negative responses if users are bored by a long series of slow clips. Likewise, a fast-paced clip may be viewed as irritating if users are over simulated by a series of previous fast-paced clips. This is why classical symphonies typically have movements that vary in tempo and mood and why DJs work in sets of 2-6 fast songs of similar genre and vary the music type to suit the audience [2].

Also, the ECAS assumes that a 5.5 music clip rating (halfway in the 1-10 ratings) is neutral in happiness for all users. In fact users are known to vary in how they bunch ratings: some tend to rate very few items low in the scale whereas others tend to rate very few items high in the scale *even when told where neutral is in the scale*. As a result, clips with a 6 or 7 rating for one user may actually have negative impacts on Happiness while clips with a 4 or 5 rating may have positive impacts on another user.

3 Measurement

A major problem is how to measure happiness. Standard questionnaires can be used. It is best to use previously validated questionnaires specifically designed for music response such as Asmus' 9-Affective Dimensions (9-AD) [3] test or Bartel's CART-M test (Cognitive-Affective Response Test) [4]. It may be possible to use shortened forms of these questionnaires containing only those questions related to happiness. Another possibility is to use physiological measurements such as heart rate, which has been found in some studies to increase due to emotional response to music [5]. Of course there is still the matter of whether the emotional response is happiness or some other emotion and whether the emotional response is positive or negative. Nevertheless, a physiological measure provides a different perspective than

questionnaires, which suffer from unreliable introspection and differences in scale among users (i.e., is one user's happiness +2 equivalent to another user's +2?). Physiological measures do not suffer from introspection problems and are easily normalized and many such as heart rate are easily measured with minimal and inexpensive equipment. Using only physiological measures would *not* be recommended. However when all measures agree, then that gives extra confidence in the results and when measures disagree, the results can be labeled as suspect. The advantage of using both types measures is that each covers the weaknesses of the other.

4 Evaluation

This evaluation addresses the Adaptive Challenge evaluation goals, namely determining which of three different Happiness models as represented by three parameterized equations best predicts Happiness for a variety of users.

4.1 Recalibration

First recalibrate the ratings to users' actual neutral points by asking users to re-rate the music clips as to whether hearing the clip would *usually* make the user more happy, less happy, or neutral. The questionnaire should also ask about other emotions. The results should be used to filter out those clips that engender emotions other than happiness in the user, because such music clips will confound the Happiness measurement. It may be useful to add a multi-point scale to happiness (and the other emotions) that can be used to infer the relationship between ECAS ratings and happiness (is it linear?). However this would lengthen the recalibration considerably and it is unclear whether users' self-ratings of happiness correlate with their actual responses. By eliminating music clips that affect other emotions, any changes in heart rate can be assumed to be due to positive or negative Happiness. The direction can be assumed to be the same as the user originally rated the music clip. For example a larger increase in heart rate to a positively rated music clip is assumed to have a higher Happiness rating and a larger increase in heart rate to a negatively rated music clip is assumed to have a lower Happiness rating.

The recalibration should follow standard experimental practices with a rest period between music clips that is long enough for the user's emotions to settle back to a neutral base. A pilot study should be done to determine the optimal length of the rest period, which should be long enough to allow settling for the strongest music clips. Here, the physiological measures are very helpful. Since heart rate lags the emotional stimulation, once heart rate settles down, then one can safely assume that the user's emotional state has also settled.

4.2 Pilot Study

Now we are ready to test the Happiness equations. Rather than attempt to study multiple users in a group where individuals can influence each other with respect to their enjoyment of a music clip, the study should evaluate a single individual at a time. To maintain emotional detachment, the study should be carried out in a neutral room with no distractions (e.g. posters on the wall, windows, window savers on nearby computers, reading material on a desk, experimenters walking around, etc.). Because the purpose of the ECAS is to select a sequence of music clips that are played continuously one after another, the same format should be used for the experiment. The user should fill out an emotion questionnaire that is shorted to only ask about Happiness at the very start and end of each music sequence and periodically during the middle of the music sequence. The optimal period between questionnaires can be determined in the pilot study by giving very frequent questionnaires and noting how long it takes between significant changes in the questionnaire answers. Heart rate should be measured continuously.

The order of music clips must be carefully designed for each user. It is probably best to follow standard DJ practices [2] to group related music clips and vary the tempos. A pilot study should be performed to determine how many positive music clips of what Ratings in a row tend to max out Happiness (assumption A5) and likewise for negative clips. The pilot study should have at least several participants so as to determine the variability among users. A random selection of music clips probably will not work well to move Happiness to the extremes possible with just music, which will best distinguish Happiness equation 3 from 1 and 2, so a series of clips should be selected to get to those extremes for each user.

4.3 Full study

The full study will present a small number of users (one at a time) with personalized sequences of music clips designed to maximize and minimize their Happiness as well as randomly selected sequences. Users should be presented with long enough sequences to achieve the max/min as determined by the pilot study. The selection of music clips should be based on the individual's recalibrated ratings and follow standard DJ practices [2] even though the actual ECAS may not be able to do so because the purpose of this study is not to test the ECAS in situ, but to gather data to determine which (if any) of the Happiness equations is correct. Users should be continuously monitored for heart rate and should be asked to fill out questionnaires about their Happiness periodically throughout the session.

4.4 Analysis

The first step in analysis is to verify that the change in Happiness before and after hearing a music clip is correlated with the Rating of the clip. The Pearson chi-square test can be used to compute the correlation between the categorical variables of delta Happiness (as derived from the surveys and heart rate) and Impact. Multiple users

should be tested to determine the variability among users. Next the three different Happiness equations can be tested to see which equation fits best with the data after optimal selection of the equation parameters based on the data. A simple F test can be used to determine fit. It may be that none of the equations really fit the data, in which case the assumptions may need to be reconsidered.

4.5 Caveats

Note that this experiment does not really test whether the ECAS will work well as advertised. First, the experiment only looks at one person at a time, thus ignoring group dynamics, which adds another layer of confusion to the Happiness equations. Second, this experiment only deals with the single emotion of Happiness and music that engenders other emotions are not modeled at all. Third, the experiment does not actually use the adaptive music-clip selection algorithm. Further experiments would be needed to determine whether the ECAS will work as advertised as opposed to the stated goals of the Adaptive System Evaluation Challenge, which all relate to evaluating the Happiness models and not to evaluating the actual effectiveness of the adaptation. Fourth, there are critical problems in the database of the ECAS that really should be addressed before trying to evaluate it. For example, the tempo of the music clips, an essential element for DJ selection [2] is missing from the database. Also, the context of the group is completely ignored. For example, if the group is at a dance party versus a funeral, a completely different selection of music would make the group happy. These problems should probably be addressed before trying to deploy or evaluate the full ECAS.

5 Ethics

Because asking users to listen to music that is suspected to cause negative happiness in the user is akin to torture (e.g., General Manuel Noriega was blasted with rock music to persuade him to leave his sanctuary), there are sensitive ethical questions about this study. At worst, this study will have to avoid negative Impact clips and use only positive and neutral clips.

References

1. Vink, A.: Music and Emotion. Living apart together: a relationship between music psychology and music therapy. In: *Nordic Journal of Music Therapy* 10(2) (2001) 144-158
2. DJU: Music Selection 101. At: <http://dju.prodj.com/courses/music/c2.shtml>
3. Asmus, E. P.: The development of a multidimensional instrument for the measurement of affective responses to music. In: *Psychology of Music* 13 (1985). 19-30
4. Bartel, L. R.: The Development of the Cognitive Affective Response Test. In: *Psychomusicology* 11 (1992) 15-26
5. Dainow, E.: Physical Effects and Motor Responses to Music. In: *Journal of Research in Music Education* 25(3) (1977) 211-221