Cross-Cultural Mood Regression for Music Digital Libraries

Xiao Hu
University of Hong Kong  
Pokfulam Road  
Hong Kong S.A.R.  
xiaoxhu@hku.hk

Yi-Hsuan Yang  
Academia Sinica  
128 Academia Road, Section 2,  
Nankang, Taipei 11529, Taiwan  
yang@citi.sinica.edu.tw

ABSTRACT
Mood is a popular access point in music digital libraries and online music repositories, and is often represented as numerical values in a small number of emotion-related dimensions (e.g., valence and arousal). As music mood is recognized as culturally dependent, this study investigates whether regression models built with music data in one culture can be applied to music in another culture. Results indicate that cross-cultural predictions of both valence and arousal values are feasible.

Categories and Subject Descriptors

General Terms
Measurement, Performance, Experimentation.

Keywords
Music mood, regression, cross-cultural, music digital libraries

1. INTRODUCTION
Nowadays as more and more music has been disseminated beyond the country and cultural boundaries and listened by people from all over the world [6]. Music digital libraries (MDL) and online music repositories are striving to meet the needs of users from a global population. Mood as a metadata type of music information has been used for organizing and searching for music in many MDL, but it is recognized that music from different cultural background may have different mood profiles [6] and listeners in different cultural groups may perceive the same pieces of music as in different moods [1]. This study aims to investigate whether and to what extent prediction models of music mood built on music in one culture can be applied to music in another culture. Specifically, three music datasets consisting of Chinese and Western songs as well as mood annotations from Chinese and Western listeners are evaluated. The results will shed light on the feasibility of applying mood regression in cross-cultural MDL.

2. MUSIC MOOD REPRESENTATION
There are two major kinds of music mood representations: categorical and dimensional models. The former represents music mood with a set of discrete categories such as “happy” and “calm.” Researchers have developed classification models to categorize music into different mood categories, and a recent study reported that classification models based on the categorical mood representation can be applied cross-culturally [6]. Dimensional mood representations use numerical values in a low dimensional emotion space to indicate the mood of music. The most widely accepted dimensional model is the Valence-Arousal model [4] where the mood of a piece of music is represented as a pair of values indicating the degrees of valence (i.e., level of pleasure) and arousal (i.e., level of energy). A number of studies have developed regression models to automatically predict the valence and arousal values for music pieces [2]. However, there has been little research investigating the cross-cultural generalizability of those regression models and this study aims to bridge this research gap.

3. EXPERIMENT DESIGN
Three datasets consisting music in different cultures and mood annotations by listeners in different cultural groups are used to evaluate the cross-cultural generalizability of music mood regression models. All of the three datasets were annotated in the valence and arousal dimensions and each song has one pair of values representing the overall mood expressed by the song. The first dataset is CH496 which contains 30-second clips extracted from 496 Chinese Pop songs released in Taiwan, Hong Kong and Mainland China. The clips were annotated by three music experts born and raised up in Mainland China. Therefore this dataset represents a Chinese cultural background in terms of both music and annotators. The second dataset is MER60 [5] which consists of 60 pieces of 30-second clips of English Pop songs and was annotated by 40 non-expert listeners born and raised up in Taiwan (where people generally identify themselves as having a Chinese cultural background). The third dataset is DEAP120 which contains 120 pieces of 1-minute video clips of Europe and North America music collected from YouTube [3]. The clips were annotated by 14-16 European students and thus this dataset represents a Western cultural background in terms of both music and annotators. Table 1 summarizes characteristics of the datasets.

<table>
<thead>
<tr>
<th>Table 1. Characteristics of the three datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music culture</td>
</tr>
<tr>
<td>Annotator culture</td>
</tr>
<tr>
<td>Size</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Type</td>
</tr>
</tbody>
</table>

As in previous studies, regression models were built for valence and arousal dimensions separately [2]. All combinations of the three datasets were evaluated with one dataset for training and the other for testing. When the same dataset was used for both
training and testing, it was within-dataset evaluation. When different datasets were used for training and testing, it was cross-dataset evaluation and the two datasets were balanced by random sampling from the larger dataset.

Psychoacoustic features were extracted from each of the audio clips in the three datasets. These audio-based features have been used and reported as effective in previous studies on music mood regression [2] and classification [6]. The regression model used in this study was Support Vector Regression (SVR) with the RBF kernel, which has been shown as highly effective and robust in music mood regression [2][5]. The parameters of SVR were determined by grid searches on the training data. The performance measure reported in this study is squared correlation coefficient ($R^2$) which measures the level of the agreement between the predicted values and the annotated values. All experiments were run in 10-fold cross-validation and were repeated 20 times. Pairwise student t-test was used in comparing differences of performances.

4. RESULTS

Tables 2 and 3 present the $R^2$ values of the regression models of the arousal and valence dimensions built on psychoacoustic features across all the nice combinations of the three datasets.

Table 2. Regression performances (in $R^2$) on arousal

<table>
<thead>
<tr>
<th>Arousal</th>
<th>CH496 [train]</th>
<th>MER60 [train]</th>
<th>DEAP120 [train]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH496</td>
<td>0.80</td>
<td>0.73</td>
<td>0.42</td>
</tr>
<tr>
<td>MER60</td>
<td>0.77</td>
<td>0.77</td>
<td>0.47</td>
</tr>
<tr>
<td>DEAP120</td>
<td>0.67</td>
<td>0.70</td>
<td>0.44</td>
</tr>
</tbody>
</table>

The results of within-dataset regressions on arousal are shown in the diagonal cells in Table 2. For CH496 and MER60 datasets, the within-dataset regression achieved better performances ($R^2 = 0.80$ for CH496 and $R^2 = 0.77$ for MER60) than cross-dataset regression between the two datasets ($R^2 = 0.77$ for testing on CH496 and $R^2 = 0.73$ for testing on MER60), but the differences were not significant ($p = 0.103$ for CH496; $p = 0.052$ for MER60). In addition, the $R^2$ values were comparable to other studies on predicting arousal values for music [2][5]. Therefore, cross-dataset prediction between CH496 and MER60 is feasible. The fact that the two datasets contain music from different cultures indicates regression models on arousal can be generalized across the cultural boundary given both datasets are annotated by listeners from the same cultural background.

When using DEAP120 as training data (the third row), performances on CH496 and MER60 further reduced to $R^2 = 0.67$ and $R^2 = 0.70$. Although the differences between these performances and within-dataset performances were significant ($p < 0.001$ for CH496; $p = 0.003$ for MER60), the performance values were still comparable to the literature [2]. However, when using DEAP120 as test data (the third column), the performances were not good no matter which dataset was used as training data. The observation that arousal prediction on DEAP120 is generally difficult may be because arousal perception of music video is also influenced by the visual channel while in this study only audio features were used in the regression model.

For results on valence (Table 3), it is not surprising that the performances of the valence dimension were much worse than those of the arousal dimension. This is consistent with previous research which has found that valence values are much harder to predict than arousal values [2], partially because the subjectivity in annotating valence values.

Table 3. Regression performances (in $R^2$) on valence

<table>
<thead>
<tr>
<th>Valence</th>
<th>CH496 [test]</th>
<th>MER60 [test]</th>
<th>DEAP120 [test]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH496</td>
<td>0.25</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>MER60</td>
<td>0.26</td>
<td>0.11</td>
<td>0.22</td>
</tr>
<tr>
<td>DEAP120</td>
<td>0.14</td>
<td>0.22</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Similar to arousal prediction, cross-dataset predictions between CH496 and MER60 seem feasible as the performances ($R^2 = 0.26$ for testing on CH496 and $R^2 = 0.15$ for testing on MER60) were comparable to or better than those of within-dataset predictions ($R^2 = 0.25$ for CH496 and $R^2 = 0.11$ for MER60) and were similar to other related studies [2]. Unlike the results in arousal prediction, the within-dataset prediction on DEAP120 achieved fairly good performance ($R^2 = 0.21$) compared to the literature [2]. This seems to suggest that the visual and audio channels in DEAP120 affected valence perception in a consistent manner and thus using only audio features could predict valence values annotated based on both video and audio cues. The cross-dataset predictions between MER60 and DEAP120 achieved better or similar performances ($R^2 = 0.22$) to within-dataset predictions of both datasets ($R^2 = 0.11$ and 0.21). These observations seem to suggest that cross-dataset regression on valence is feasible when the datasets are either 1) composed of music in different cultures but annotated by listeners in the same cultural group (CH496 and MER60) or 2) annotated by listeners in different cultural groups but composed of music in the same culture (MER60 and DEAP120). The worst performance of cross-dataset prediction happened between CH496 and DEAP120 ($R^2 = 0.08$), possibly due to the fact that these two datasets differ in cultural background of both the music and the annotators.

5. CONCLUSIONS AND FUTURE WORK

In this study, we have investigated cross-cultural and cross-dataset generalizability of psychoacoustic feature based regression models in predicting valence and arousal values of music pieces. Experiments on three datasets revealed that generalizability is supported for arousal and valence predictions when the annotators were from the same cultural background. For arousal prediction the annotations need to be based on music audio only. These findings have practical values for providing cross-cultural access to music. Future work will investigate the problem on music in other cultures beyond Chinese and Western cultures.

6. REFERENCES


