# A Study on Cross-cultural and Cross-dataset Generalizability of Music Mood Regression Models

**Xiao Hu**
University of Hong Kong
xiaoxhu@hku.hk

**Yi-Hsuan Yang**
Academia Sinica
yang@citi.sinica.edu.tw

## ABSTRACT

The goal of music mood regression is to represent the emotional expression of music pieces as numerical values in a low-dimensional mood space and automatically predict those values for unseen music pieces. Existing studies on this topic usually train and test regression models using music datasets sampled from the same culture source, annotated by people with the same cultural background, or otherwise constructed by the same method. In this study, we explore whether and to what extent regression models trained with samples in one dataset can be applied to predicting valence and arousal values of samples in another dataset. Specifically, three datasets that differ in factors such as cultural backgrounds of stimuli (music) and subjects (annotators), stimulus types and annotation methods are evaluated and the results suggested that cross-cultural and cross-dataset predictions of both valence and arousal values could achieve comparable performance to within-dataset predictions. We also discuss how the generalizability of regression models can be affected by dataset characteristics. Findings of this study may provide valuable insights into music mood regression for non-Western and other music where training data are scarce.

## 1. INTRODUCTION

Music from different cultural backgrounds may have different mood profiles. For example, a recent study on cross-cultural music mood classification [1] found that fewer Chinese songs are associated with radical moods such as 'aggressive' and 'fiery,' compared to Western songs. It has also been reported that people from different cultural backgrounds often label music mood differently [2]. It is thus interesting to investigate whether and to what extent automatic music mood recognition models can be applied cross-culturally. This is particularly relevant as more and more non-Western music is gaining researcher's attention [3] while Music Information Retrieval (MIR) techniques are still predominately developed and tested using Western music.



It has been found that music mood classification models trained on English songs can be applied to Chinese songs and vice versa, although the performances were significantly degraded from those in within-cultural experiments [1]. As music mood can be represented not only by discrete categories but also in dimensional spaces [4], it is of research and practical interests to investigate whether mood regression models built with dimensional mood spaces can be generalized cross cultural boundaries. More generally, in this paper we investigate whether mood regression models can be generalized cross different datasets with distinct characteristics.

To explore the cross-cultural and cross-dataset generalizability of regression models, we apply two analysis strategies: 1) to train and evaluate regression models using three datasets that differ in music (stimulus) cultural background, annotator (subject) cultural background, stimulus type, and annotation method; 2) to use different sets of audio features in building regression models. The first analysis will provide empirical evidences on whether and under which circumstances mood regression models can be generalizable cross-culturally and cross-datasets. The second analysis will help identify a possible set of audio features that can be effective across datasets. Such knowledge is insightful for building mood recognition systems applicable to situations where training data are expensive or otherwise difficult to obtain.

## 2. RELATED RESEARCH

### 2.1 Categorical and Dimensional Representations of Music Mood

Mood as an essential aspect of music appreciation has long been studied in music psychology [5] where numerous mood models have been developed.[1] These models can be grouped into two major categories. The first is categorical models where mood is represented as a set of discrete classes such as 'happy,' 'sad,' and 'angry,' among others. Many studies on music mood in MIR are

[1] We use the terms mood and emotion interchangeably in this paper, although they bear different meanings and implications in psychology.

based on the categorical model where one or more mood class labels are assigned to each music piece [1, 6, 7].

The second is dimensional models where mood is represented as continuous values in a low-dimensional space. Each dimension is a psychological factor of moods. Models may vary in the dimensions considered but most of them include dimensions of *arousal* (i.e., level of energy), *valence* (i.e., level of pleasure) [8], and sometimes *dominance* (i.e., level of control). Dimensional models are also very popular in MIR where regression models are built to predict numerical values in the dimensions for each music piece [4, 7, 9-13].

Both categorical and dimensional models have their own advantages and disadvantages. The semantics of mood class labels in categorical models is the most natural for human users while dimensional models can represent the degree of mood (e.g., a little vs. very much pleased), for example. Therefore, to obtain a more complete picture of music mood, it is better to consider both types of representations [7].

### 2.2 Cross-cultural Music Mood Classification

In recent years cross-cultural issues have garnered much attention in the music computing research community (e.g., [1, 3]). In particular, as most existing research has been focused on Western music, researchers are interested in finding out whether and to what extent conclusions drawn on Western music can be applied to non-Western music. In music mood classification, a recent study [1] compared mood categories and mood classification models on English Pop songs and Chinese Pop songs. Classification models were trained with songs in one culture and tested with those in the other culture. The result showed that although within-cultural (and thus within-dataset) classification outperformed cross-cultural (and thus cross-dataset) classification, the accuracy levels of cross-cultural classification were still acceptable.

Motivated by [1], this study is to investigate whether cross-cultural generalizability holds when music mood is represented in a dimensional space. Moreover, the present study goes even one step further to examine cross-dataset applicability which is more general and covers more factors in addition to cultural background.

### 2.3 Cross-genre Mood Regression in Western Music

When music mood is represented in dimensional spaces, the technique used to predict a numerical value in each dimension is regression [7]. To our best knowledge, there have been very few studies on cross-cultural or cross-dataset music mood regression, and most of them have been on Western music. In [14], Eerola explored cross-genre generalizability of mood regression models and concluded that arousal was moderately generalizable across genres but valence was not. Although Eerola exhaustively evaluated nine datasets of music in different genres, all the datasets were composed of Western music [14]. In contrast, our study focuses on the generalizability across different cultures with culture being defined with regard to music (stimuli) and annotators (subjects), and across datasets with different characteristics.

## 3. THE DATASETS

Three datasets are adopted in this study. All of them were annotated in the valence and arousal dimensions. Each song clip in these datasets was associated with a pair of valence and arousal values that represent the overall emotional expression of the clip, rather than a time-series trajectory that depicts mood variation as time unfolds [6, 7]. In other words, the mood of a clip is assumed to be not time-varying, and the investigation of time-varying moods is left as a future work. In what follows, we provide detailed descriptions of the datasets and compare them from several factors that may affect model generalizability.

### 3.1 The CH496 Dataset

This Chinese music dataset contains 496 Pop song clips sampled from albums released in Taiwan, Hong Kong and Mainland China. Each of the clips was 30-second long and was algorithmically extracted such that the chosen segment was of the strongest emotion as recognized by the algorithm [1]. The clips were then annotated by three experts who were postgraduate students in Music major and were born and raised up in Mainland China. The annotation interface contained two separate questions on valence and arousal and was written in Chinese to minimize possible language barriers in terminology and instructions. For each clip, the experts were asked to give two real values between [−10, 10] for valence and arousal. To ensure reliability across annotators, the three experts had a joint training session with an author of the paper where example songs with contrasting valence and arousal values were played and discussed till a level of consensus was reached.

Pearson's correlation, a standard measure of inter-rater reliability for numerical ratings [15], was calculated between each pair of annotators. The average Pearson's correlation across all pairs of annotators was 0.71 for arousal and 0.50 for valence. The former is generally acceptable and regarded as high agreement level [15]. While the agreement level on valence can only be regarded as moderate at best, it is comparable to other studies in the literature where the subjectivity of music valence has been well acknowledged (e.g., [7, 11-13]). Therefore, the average values across the three annotators were used as the groundtruth. As the annotators were experts who have been trained for the task and come from the same cultural background, this dataset is deemed as highly suitable for the task in question.

### 3.2 The MER60 Dataset

This English music dataset was developed by Yang and Chen [13]. It consists of 60 pieces of 30-second clips manually selected from the chorus parts of English Pop songs. Each clip was annotated by 40 non-experts recruited from university students who were born and raised up in Taiwan and thus had a Chinese cultural background. The subjects were asked to give real values ranging between [–5, 5] to the valence and arousal dimensions at the same time. The values were entered by clicking on an emotion space displayed on a computer screen. With this interactive interface, a subject was able to compare the annotations of different clips she or he just listened to and possibly refined the annotations. The groundtruth values were the average across all subjects after outliers were removed. With an advanced annotation interface and a large number of subjects from the same cultural background, this dataset is deemed as of high fitness to the task as well.

### 3.3 The DEAP120 Dataset

The DEAP dataset [16] contains 120 pieces of one-minute music video clips collected from YouTube (http://www.youtube.com). The music video featured songs of European and North American artists and thus was of Western cultural background. Each clip was annotated by 14–16 European student volunteers whose cultural background could be identified as Western. The subjects were asked to annotate valence, activation (equivalent to arousal), and dominance separately on a discrete 9-point scale for each video clip using an online self-assessment tool. The annotated values on each clip were then aggregated and normalized using z-score ($\mu/\sigma$).

It is noteworthy that the original stimuli of this dataset were music video and thus the annotations were applied to both the audio and the moving image components. To be able to perform cross-dataset evaluation in this study, we only extracted features from the audio component. Therefore, some important cues might be lost. In addition, the discrete annotation values may not be as accurate as real values in the other two datasets, and thus this dataset is regarded as medium level suitability to the task of this study.

We also note that the emotional expression of music can be further divided into emotions that are considered being expressed in the music piece (i.e. *intended* emotion) or emotions that are felt in response to the music piece (i.e. *felt* emotion). The first two datasets considered in this study were labeled with intended emotion [1, 13], whereas the last one was labeled with felt emotion [16]. Therefore, this is another important difference among the three datasets.

### 3.4 Qualitative Comparison of the Three Datasets

Table 1 summarizes the characteristics of the three datasets from the perspectives of stimuli, subjects, and annotation methods. Any pair of the datasets is cross-cultural in terms of stimuli, subjects, or both. Some combinations of the datasets are also cross stimulus type and annotation methods. Therefore, experiments on these datasets would shed light on the effect of these different factors on the generalizability of mood regression models.

| | | **CH496** [1] | **MER60** [13] | **DEAP120** [16] |
|---|---|---|---|---|
| **Stimuli** | Type | Music | Music | Music video |
| | Size | 496 | 60 | 120 |
| | Culture | Chinese | Western | Western |
| | Length | 30 seconds | 30 seconds | 1 minute |
| | Segment selection | With strongest emotion; automatic | Chorus; manual selection | With strongest emotion; automatic |
| **Subjects** | Type | Experts | Volunteers | Volunteers |
| | Culture | Chinese | Chinese | Western |
| | Number | 3 per clip | 40 per clip | 14–16 per clip |
| **Annotation** | Scale | Continuous | Continuous | Discrete |
| | Dimensions | V. A. | V. A. | V. A. D. |
| | Interface | Annotate dimensions separately | 2-D interactive interface | Annotate dimensions separately |
| | Emotion | Intended | Intended | Felt |
| | Fitness to the task | High | High | Medium |

**Table 1**. Characteristics of the three datasets. Acronyms: V.: valence, A.: arousal, D.: dominance.

Table 2 presents the numbers of music clips in each quadrant of the 2-dimensional space across datasets. A chi-square independence test [17] on the three distributions indicates the distribution is dataset-dependent ($\chi^2$ = 30.70, d.f. = 6, $p$-value < 0.001). In other words, the distributions of music clips in the four quadrants of the valence-arousal space are significantly different across the datasets. Pair-wised chi-square independence tests show that the distributions of CH496 and MER60 are not significantly different ($\chi^2$ = 2.10, d.f. = 3, $p$-value = 0.55), neither are MER60 and DEAP120 ($\chi^2$ = 4.37, d.f. = 3, $p$-value = 0.22). However, DEAP120 is significantly different from CH496 ($\chi^2$ = 30.43, d.f. = 3, $p$-value < 0.001). The test results are very interesting in that the MER60 dataset seems to be in between of the other two datasets whose sample distributions are very different from each other. When looking at the dataset characteristics (Table 1), MER60 indeed situates in the middle: it shares the same music cultural background with DEAP120 and the same annotator cultural background with CH496.

| | V+A+ | V-A+ | V-A- | V+A- | Total |
|---|---|---|---|---|---|
| CH496 | 228 | 82 | 130 | 56 | 496 |
| MER60 | 23 | 11 | 16 | 10 | 60 |
| DEAP120 | 33 | 20 | 31 | 36 | 120 |

**Table 2**: Distributions of audio clips in the 2-d valence (V) arousal (A) space. V+A+ stands for the first quadrant of the space, V−A+ stands for the second quadrant, etc.

Figure 1 is the scatter plots of the three datasets in the valence−arousal space (normalized to the scale of [−1, 1]). Each point represents the average valence and arousal ratings for a music piece across the annotators. There are certain patterns in common across the plots: for example, no samples in the bottom right corner (very low arousal and very positive valence). However, CH496 is relatively more skewed toward the first quadrant, suggesting that there is possibly a bias toward happy and upbeat songs in the Chinese dataset. By comparing MER60 and DEAP120, we see that the samples of the former dataset are farther away from the origin of the space, showing that either the stimuli in MER60 have stronger emotion, the subjects regarded songs in MER60 had stronger emotion, or the subjects had higher degree of consensus on the mood of music in MER60 (so the annotated values did not cancel out in the aggregation process of taking the average of the subjects' ratings).

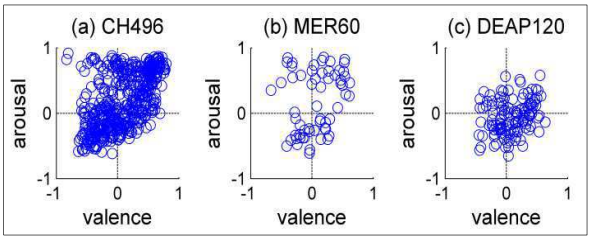


**Figure 1**. Scatter plots of the distribution of valence and arousal values in the three datasets.

## 4. REGRESSION EXPERIMENTS AND RESULTS

As in previous studies on music mood regression, separate regression models were built for valence and arousal. All nine combinations of the three datasets were evaluated in this study, with one dataset for training and the other for testing. When the same dataset was used as training and test data (*within-dataset* regression), 10 fold cross validation was applied. In contrast, when different datasets were used (*cross-dataset* regression), the data sizes were balanced by random sampling from the larger dataset. In both cases, the regression experiment was repeated 20 times for a stable, average performance. The regression model used in this study was Support Vector Regression (SVR) with the Radial Basis Function (RBF) kernel, which has been shown as highly effective and robust in previous research on music mood regression [7]. The parameters of SVR were determined by grid searches on the training data. The performance measure used in this paper is squared correlation coefficient ($R^2$). Moreover, the pair-wise student *t*-test is used in comparing the differences of performances.

### 4.1 Audio Features

In music mood classification and regression, it is still an open question which audio features are most effective. In order to see the effectiveness and generalizability of different acoustic cues, we followed [1] and compared six widely used audio feature sets which are reprinted in Table 3, along with abbreviations. Although employing features from the lyrics of songs might lead to a better accuracy (especially for the valence dimension [11]), we did not explore this option in this study due to the difference in the languages of the stimuli.

| Feature | Type | Dim | Description |
|---|---|---|---|
| RMS | Energy | 2 | The mean and standard deviation of root mean square energy |
| PHY | Rhythm | 5 | Fluctuation pattern and tempo |
| TON | Tonal | 6 | Key clarity, musical mode (major/minor), and harmonic change (e.g., chord change) |
| PCP | Pitch | 12 | Pitch class profile: the intensity of 12 semitones of the musical octave in Western twelve-tone scale |
| MFCC | Timbre | 78 | The mean and standard deviation of the first 13 MFCCs, delta MFCCs, and delta delta MFCCs |
| PSY | Timbre | 36 | Psychoacoustic features including the perceptual loudness, volume, sharpness (dull/sharp), timbre width (flat/rough), spectral and tonal dissonance (dissonant/consonant) of music |

**Table 3.** Acoustic feature sets used in this study ("Dim" stands for number of dimensions of a feature sets).

Table 4 shows within- and cross-dataset performances across all feature sets, averaged across various dataset combinations. It can be seen that the psychoacoustic features (PSY) outperformed other feature sets on predicting both arousal and valence values. This is the same as in [1] where PSY was the best performing feature sets for both within- and cross- cultural mood classification.

Across all feature sets, within-dataset performances were consistently higher than cross-dataset ones. PSY and MFCC feature sets are more generalizable across datasets in that the reductions from within- to cross-dataset performances on these feature sets were smaller than those of other feature sets. This might due to the nature of the feature sets, or because of the fact that PSY and MFCC are of higher dimensions among the considered feature sets. In contrast, TON feature set seems less generalizable across datasets, as evidenced by the large differences between within- and cross-dataset performances.

For arousal prediction, the performance differences between PSY features and other feature sets were all significant ($p$-value < 0.005). However, it is noteworthy that the PCP features, with only 12 dimensions, performed as well as the famous MFCC features for arousal. This might be due to the fact that the 12 chroma intensity features captured the pitch level and contour of music pieces that are recognized as related to arousal [5].

For valence prediction, it is not surprising that the performances were much inferior to those of arousal. All previous research has found that valence values are much harder to predict than arousal values [11, 12, 14], partially because the subjectivity in annotating valence values. Among all the six feature sets, the differences between PSY, MFCC and TON on valence prediction were not significant at $p$-value = 0.05 level. It is also noteworthy that the TON features, with only 6 dimensions, achieved the same level of performances for valence prediction as MFCC and PSY features. This perhaps can be explained by findings in music psychology that connect the mode (i.e., major vs. minor) and harmony (consonant vs. dissonant) factors to valence [5].

| | | RMS | RHY | TON | PCP | MFCC | PSY |
|---|---|---|---|---|---|---|---|
| **Arousal** | Within- | 0.17 | 0.50 | 0.49 | 0.61 | 0.60 | 0.67 |
| | Cross- | 0.16 | 0.41 | 0.16 | 0.57 | 0.57 | 0.63 |
| | Avg. | 0.17 | 0.44 | 0.27 | 0.58 | 0.58 | 0.64 |
| **Valence** | Within- | 0.08 | 0.14 | 0.26 | 0.19 | 0.17 | 0.19 |
| | Cross- | 0.12 | 0.09 | 0.11 | 0.10 | 0.15 | 0.18 |
| | Avg. | 0.11 | 0.10 | 0.16 | 0.13 | 0.16 | 0.18 |

**Table 4.** Performances (in $R^2$) of different feature sets. Acronyms: "within-" and "cross-" stand for within- and cross-dataset performances, "Avg." stands for average performances across all the nine dataset combinations.

Notwithstanding that one might be able to obtain better performance on these three datasets through feature engineering and model optimization, we opt for using simple features and simple machine learning models and focusing on the general trends. The following analysis on arousal prediction will be based on the performances obtained on the PSY feature set, while the analysis on valence prediction will be based on the performances obtained on a combined feature set of top performing features: PSY, MFCC and TON.

### 4.2 Cross-dataset Performances on Arousal

Table 5 summarizes the regression performances on different combinations of the datasets. The columns list the test dataset and the rows list the training dataset.

The first two columns show the results when CH498 and MER60 were used for testing. Not surprisingly, the best performance on each of the two datasets was achieved when the models were trained on the dataset itself (i.e. within-dataset). When using the other dataset as training data, the performances decreased but not at a significant level ($p$-value = 0.103 for CH496; $p$-value = 0.052 for MER60). Also, the reduced performances are still comparable or even better than other studies on predicting arousal values for music (e.g., Guan *et al.* [11] reported 0.71). Therefore, cross-dataset prediction between CH496 and MER60 can be considered feasible. The fact that the two datasets contain music from different cultures indicates regression models on arousal can be generalized cross the cultural boundary given both datasets are annotated by listeners from the same cultural background.

| **Arousal [PSY]** | **CH496 [test]** | **MER60 [test]** | **DEAP120 [test]** | **Avg.** |
|---|---|---|---|---|
| **CH496 [train]** | 0.80 | 0.73 | 0.42 | 0.65 |
| **MER60 [train]** | 0.77 | 0.77 | 0.47 | 0.67 |
| **DEAP120 [train]** | 0.67 | 0.70 | 0.44 | 0.60 |

**Table 5**. Regression performances (in $R^2$) on arousal.

When using DEAP120 as training data (i.e. the third row), performances on CH496 and MER60 further reduced to 0.67 and 0.70, respectively. Although the performances are significantly different from within-dataset performances ($p$-value < 0.001 for CH496; $p$-value = 0.003 for MER60), the performance values are still acceptable. However, when using DEAP120 as test data (i.e. the third column), the performances were not good regardless of 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, or because DEAP120 is concerned with felt emotion rather than intended emotion. While validation of such conjectures is beyond the scope of this study, it is safe to say stimulus type or suitability of the annotation to the task does play a role in arousal prediction.

So far, we have looked at the absolute performance values with regard to whether they are acceptable empirically. For the generally unacceptable performances on DEAP120 (i.e. the third column in Table 5), it is worthwhile to examine the relative performances using different training datasets. The model trained on MER60 ($R^2$ = 0.47) even outperformed the within-dataset prediction on DEAP120 ($R^2$ = 0.44), while the model trained on CH496 ($R^2$ = 0.42) performed significantly worse than within-dataset prediction ($p$-value = 0.04). The difference between MER60 and CH496 lies in cultural background of stimuli (Chinese songs in CH496 vs. Western songs in MER60). Therefore, when the test data are of a different stimulus type or the annotations are not highly suitable to the task, the model trained on music from the same cultural background has better generalizability than that trained on music from a different culture.

In summary, although cross-dataset performances are generally lower than within-dataset prediction, cross dataset prediction of arousal seems generally feasible, espe-

cially when the training and testing datasets are annotated by subjects from the same cultural background. When the test dataset is of a different stimulus type (e.g., music versus music video), only models trained with music of the same cultural background can be applied without significant performance degradation.

### 4.3 Cross- dataset Performance on Valence

Table 6 presents the $R^2$ performances on various combinations of the datasets. Similar to arousal prediction, cross-dataset predictions between CH496 and MER60 seem feasible as the performances were comparable to those of within-dataset predictions and to other related studies [7]. The music stimuli in these two datasets were from different cultures but the difference might have been compensated by the shared cultural background of the annotators.

The cross-dataset predictions between MER60 and DEAP120 even outperformed within-dataset predictions of both datasets. The model trained on DEAP120 and tested on MER60 achieved significantly higher performance ($R^2 = 0.23$) than within-dataset performance ($R^2 = 0.15$, p-value < 0.001). In addition, the model trained on MER60 can be applied to DEAP120 with a relatively high performance ($R^2 = 0.31$). Therefore, unlike in arousal prediction, stimulus type does not seem to be a barrier for cross-dataset valence prediction. In fact, also unlike the results in arousal prediction, the within-dataset prediction on DEAP120 achieved fairly good performance ($R^2 = 0.22$) compared to the literature [7]. 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.

| Valence [PSY+MFCC+TON] | CH496 [test] | MER60 [test] | DEAP120 [test] | Avg. |
|---|---|---|---|---|
| **CH496 [train]** | 0.26 | 0.16 | 0.08 | 0.17 |
| **MER60 [train]** | 0.24 | 0.15 | 0.31 | 0.23 |
| **DEAP120 [train]** | 0.12 | 0.23 | 0.22 | 0.19 |

**Table 6**. Regression performances (in $R^2$) on valence.

The worst cross-dataset performances occurred between CH496 and DEAP120. Either training/testing combination resulted in significantly lower $R^2$ values ($R^2 = 0.12$ and $R^2 = 0.08$) compared to within-data predictions ($R^2 = 0.26$, $R^2 = 0.22$, $p$-value < 0.001). If not considering stimulus type which has been regarded as not a barrier for cross-dataset valence prediction, these two datasets differ in the cultural backgrounds of both music (stimuli) and annotators (subjects). Based on these observations, we may conclude that cross-dataset regression on valence is feasible when the datasets consist of music in different cultures (CH496 and MER60) or when the datasets are annotated by listeners in different cultural groups (MER60 and DEAP120), but not both (CH496 and DEAP120).

In summary, valence prediction is generally much more challenging than arousal prediction. The factors of cultural background of music (stimuli) and annotators (subjects) are more important for cross-dataset generalizability on valence prediction than stimuli type and annotation method.

## 5. CONCLUSIONS AND FUTURE WORK

In this study, we have investigated cross-cultural and cross-dataset generalizability of regression models in predicting valence and arousal values of music pieces. Three distinct datasets were evaluated and compared to disclose the effects of different factors. The distributions of valence and arousal values of the three datasets on the 2-dimensional mood space shared common patterns, suggesting that the 2-dimensional representation of music mood can be applicable to both Western and Chinese Pop music.

Six different acoustic features were evaluated and the psychoacoustic features outperformed other features in both arousal and valence predictions, while MFCC and tonal features also performed well in valence prediction.

Cross-cultural and cross-dataset generalizability is well supported for arousal prediction especially when the training and test datasets are annotated by annotators from the same cultural background. When the test dataset is of a different stimulus type, only models trained with music in the same culture can be applied.

Cultural backgrounds of music stimuli and annotators are important for cross-dataset prediction on valence. In other words, in order to generalize valence prediction models between datasets, the two datasets should consist of music in the same culture or should be annotated by annotators with the same cultural background.

These findings provide empirical evidences and insights for building cross-cultural and cross-dataset music mood recognition systems. For future work, it would be interesting to investigate the generalizability of regression models in predicting time-series trajectory of music mood [18]. In addition, findings of the study can be further verified and enriched by considering music from other cultures.

## 6. AKNOWLEDGMENT

This study is supported in part by a Seed Fund for Basic Research from the University of Hong Kong and Grant NSC 102-2221-E-001-004-MY3 from the Ministry of Science and Technology of Taiwan.

## 7. REFERENCES

[1] Y.-H. Yang and X. Hu: “Cross-cultural music mood classification: A comparison on English and Chinese

songs," in *Proc. International Conference on Music Information Retrieval*, pp. 19–24, 2012.

[2] X. Hu and J.-H. Lee: "A cross-cultural study of music mood perception between American and Chinese listeners," in *Proc. International Conference on Music Information Retrieval*, pp. 535–540, 2012.

[3] X. Serra: "A multicultural approach in music information research," in *Proc. International Conference on Music Information Retrieval*, pp. 151–156, 2011.

[4] J. Madsen, J. B. Nielsen, B. S. Jensen, and J. Larsen: "Modeling expressed emotions in music using pairwise comparisons," in *Proc. International Symposium on Computer Music Modeling and Retrieval*, pp. 526–533, 2012.

[5] A. Garbrielesson and E. Lindstrom: "The role of structure in the musical expression of emotions," in *Handbook of Music and Emotion*, ed. P. N. Juslin and J. A. Sloboda, Oxford 2010

[6] M. Barthet, G. Fazekas, and M. Sandler: "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models," in *Proc. International Symposium on Computer Music Modeling and Retrieval*, pp. 492–507, 2012.

[7] Y. E. Kim, E. M. Schmidt, R. Migenco, B. G. Morton, P. Richardson, J. J. Scott, J. A. Speck, and D. Turnbull: "Music emotion recognition: A state of the art review," in *Proc. International Conference on Music Information Retrieval*, pp. 255–266, 2010.

[8] J. A. Russell: "A circumspect model of affect," *Journal of Psychology and Social Psychology*, vol. 39, no. 6, 1980.

[9] S. Beveridge and D. Knox: "A feature survey for emotion classification of western popular music," in *Proc. International Symposium on Computer Music Modeling and Retrieval*, pp. 508–517, 2012.

[10] M. Caetano and F. Wiering: "The role of time in music emotion recognition," in *Proc. International Symposium on Computer Music Modeling and Retrieval*, pp. 287–294, 2012.

[11] D. Guan, X. Chen and D. Yang: "Music emotion regression based on multi-modal features," in *Proc. International Symposium on Computer Music Modeling and Retrieval*, pp. 70–77, 2012.

[12] A. Huq, J. P. Bello, and R. Rowe: "Automated music emotion recognition: A systematic evaluation," *Journal of New Music Research*, Vol. 39, No. 3, pp. 227–244, 2012.

[13] Y.-H. Yang and H. H. Chen: "Predicting the distribution of perceived emotions of a music signal for content retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2184–2196, 2011.

[14] T. Eerola: "Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journal of New Music Research*, Vol. 40, No. 4, pp. 349–366, 2011.

[15] K. L. Gwet: *Handbook of Inter-Rater Reliability* (2nd Edition), Gaithersburg : Advanced Analytics, LLC, 2010.

[16] S. Koelstra, C. Muhl, M. Soleymani, J.-S., Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras: "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, 3, 1, 18–31, 2012.

[17] R. R. Sokal and C. D. Michener: "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin,* Vol. 38, pp. 1409–1438, 1958.

[18] F. Weninger, F. Eyben1, and B. Schuller: "On-line continuous-time music mood regression with deep recurrent neural networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2014.