The Mood of Chinese Pop Music: Representation and Recognition

Xiao Hu
Faculty of Education, University of Hong Kong, Runme Shaw Building, Room 209, Pokfulam Road, Hong Kong. E-mail: xiaoxhu@hku.hk

Yi-Hsuan Yang
Research Center for IT Innovation (CITI), Academia Sinica, Taipei, Taiwan. E-mail: yang@citi.sinica.edu.tw

ABSTRACT

Music mood recognition (MMR) has attracted much attention in music information retrieval research, yet there are few MMR studies which focus on non-Western music. In addition, little has been done on connecting the two most adopted music mood representation models: categorical and dimensional. To bridge these gaps, we constructed a new dataset consisting of 818 Chinese Pop (C-Pop) songs, three complete sets of mood annotations in both representations, as well as audio features corresponding to five distinct categories of musical characteristics. The mood space of C-Pop songs was analyzed and compared to that of Western Pop songs. We also explored the relationship between categorical and dimensional annotations and the results revealed that one set of annotations could be reliably predicted by the other. Classification and regression experiments were conducted on the dataset, providing benchmarks for future research on MMR of non-Western music. Based on these analyses, we reflect and discuss the implications of the findings to MMR research.

1. INTRODUCTION

Music mood, as an important metadata type, has not only attracted attention in the research field of Music Information Retrieval (MIR) but has also been utilized in many music websites and services such as AllMusicGuide and Spotify. However, it has been recognized that the field is dominated by studies on Western music. Researchers thus have started investigating cross-cultural issues on music mood such as mood descriptors applied to music in different regions (Lartillot & Toiviainen, 2007), mood perceptions of non-Western listeners (Lee & Hu, 2012), and generalizability of mood recognition models (Yang & Hu, 2012; Hu & Yang, 2016). Meanwhile, the subjective nature of music mood increases the difficulty in building datasets for research tasks related to mood recognition (Trohidis et al., 2008; Hu et al., 2010). In response to the fast growth in research in these areas, more shareable datasets with non-Western music and/or those annotated by non-Western listeners are much needed.

In representing music mood, there are two major types of models: categorical and dimensional ones. The former uses a set of natural language terms to represent different moods, such as “happy” or “angry”; the latter represents music moods with continuous values in a low dimensional space (Kim et al., 2008; Hu et al., 2010). Both models have their own advantages but few studies have explored the relationships between them (Wang et al., 2012), perhaps due to the lack of datasets annotated with both models.

To bridge the gaps, we built a new dataset of Chinese Pop (C-Pop) music, named CH818, for the purpose of investigating the mood of C-Pop. C-Pop broadly refers to popular music made by artists from the Greater China region and/or sung in a Chinese language (Liu & Mason, 2010). C-Pop songs are influenced by both contemporary Western Pop music and Chinese oriental music traditions. They are also found to be the most popular type of music in one of the largest digital music markets in the world (The International Federation of the Phonographic Industry, 2014). The CH818 dataset consists of 818 C-Pop songs, each with three complete sets of annotations from three Chinese music experts in both categorical and dimensional representations, as well as five types of features (loudness, rhythm, pitch, timbre and harmony) extracted from the music audio. To date, CH818 is the largest dataset of C-Pop songs with
comprehensive mood annotations and audio features. All metadata, annotations and extracted audio features will be publicly available for research purposes.

With CH818, we investigated a series of research questions on the representation and automated recognition of C-Pop music:

RQ1: What is the mood space of C-Pop music? How is it compared to the mood space of Western Pop music?

RQ2: What is the relationship between the two main representation models (i.e., categorical and dimensional models) on C-Pop?

RQ3: How well can mood recognition techniques designed for Western music be applied to C-Pop?

The first two questions are on mood representation of C-Pop, whereas the last one is on automated mood recognition. The answers to RQ1 can enhance our understanding on the mood space of C-Pop and how it is similar or different from that of Western Pop music. RQ2 is to bridge the two major kinds of mood representation models (See Related Work) that have been widely used in both MIR and music psychology. Despite the popularity of the two models, the relationship between them has rarely been studied empirically. As the field has been dominated by studies on Western music, little is known about the applicability of mood recognition techniques on non-Western music. RQ3 is to fill the gap. By answering these questions, this study aims to make contributions to MIR research, particularly on mood representation and recognition in the cross cultural context. Besides, this study also constructs a substantial dataset and demonstrates detailed analyses on it. Openly accessible datasets are especially desirable as they could be used for benchmarking across different labs and researchers. However, it is challenging to build shareable datasets, especially in MIR partially due to intellectual property laws (Chen et al., 2014; Hu et al., 2014). With the analysis and experiments, hopefully this study will inspire further research on mood representation and recognition of music in different cultures.

2. RELATED WORK

2.1 Music Mood Representation

Both categorical and dimensional models have been used widely in music mood recognition. A categorical model uses a set of finite discrete terms (usually adjectives) to represent moods in music. Hevner’s “adjective circle” was an early influential categorical model for music mood (Hevner, 1936), where eight groups of adjectives collectively defined eight main categories of moods in (classical) music, such as “vigorous,” “merry,” and “dreamy”. In categorical models, a song is labeled by one or more mood categories, without indications on the extent to which the song is associated with these moods. In MIR research, a widely used categorical model is the five-cluster model used in the audio mood classification (AMC) task in Music Information Retrieval Evaluation eXchange (MIREX), a community-wide campaign for MIR evaluation (Hu et al., 2008). The model contains five mood clusters derived from the editorial mood labels on a widely used online music repository, AllMuscGuide (Hu & Downie, 2007). Reprinted in Table 1, the five clusters consist of 29 mood-related terms, representing five different categories of moods in (mostly Pop) music. This model has been used in various research tasks in MIR, including automated mood classification (Laurier et al., 2008; Hu et al., 2008; Bischoff et al., 2009), user mood perceptions (Hu & Lee, 2012; Lee & Hu, 2014; Hu & Lee, 2016), cross-cultural mood annotations (Singhi & Brown, 2014; Hu et al., 2014), and crowdsourcing for music mood recognition (Lee & Hu, 2012). The categorical model in this current study is developed from this five-cluster model by extending it with new terms suitable for the characteristics of Chinese songs.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mood labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In contrast, dimensional models represent music moods with continuous values in a low dimensional space. The most adopted dimensional model in MIR is Russell's model (Kim et al., 2010; Yang & Chen, 2012; Barthet et al., 2013), which consists of two dimensions, valence (measuring the level of pleasure) and arousal (measuring the level of activity) (Russell, 1980). This study also adopts this model as reprinted in Figure 1.

![Russell's Valence-Arousal Model](image)

**Figure 1.** Russell's Valence-Arousal Model (Russell, 1980, p.1168, annotation added)

Categorical and dimensional models have their complementary advantages. Categorical models are recognized as easy for users to understand since the terms are from natural languages. On the other hand, dimensional models are advantageous for visualizing mood distributions of music and representing various intensity levels of moods. There is a recent trend in studies on music and emotion that strives to integrate both representations (Eeola & Vuoskoski, 2013). Datasets annotated with both models are precious in providing empirical evidence and support for theoretical advance on integrated representation frameworks. In MIR, it is also desirable to combine their advantages to enhance mood-based MIR systems. Multiple studies on music mood attempted to visualize the proximity of various mood categories in 2-dimensional spaces (e.g., Laurier et al., 2009; Yang & Hu, 2012) using dimension reduction techniques. However, the resultant dimensions may have little semantic meanings, marking it difficult for users to understand. It was then proposed to integrate the two representations by visualizing mood categories in a category model in the dimensional space of a dimensional model (Wang et al., 2012). In this study, we also explore the mapping between the two representation models using CH818.

Studies on music mood recognition (MMR) revealed that, for dimensional models, music valence was consistently more difficult to predict than arousal (Yang & Chen 2012; Barthet et al, 2013). For categorical models, it is also found that performances vary across mood categories and the differences might be related to the valence and arousal values corresponding to the categories (Hu et al., 2016). Towards gaining more understanding on automated mood prediction, we conduct experiments with both representations: classification experiments for categorical annotations and regression experiments for dimensional annotations. CH818 also provides a good testbed for further studies on this line of research.

### 2.2 Mood Recognition on Non-Western Music
MMR studies propose algorithms and systems to automatically classify or predict music mood from musical materials, including audio, lyrics, and/or social tags (Kim et al., 2010; Yang & Chen, 2012; Hu, Choi & Downie, 2016). Similar to other tasks in MIR, there is a predominance of Western music and Western listeners in music mood studies. Most experiments were conducted on datasets consisting of Western music and annotated by Western listeners. As music-seeking and consumption have transcended the cultural boundary and become increasingly global, researchers have realized the importance of studying non-Western music and users as well as the possibility of generalizing research findings cross-culturally (Serra, 2011; Hu et al., 2014).

A few recent studies have begun evaluating cross-cultural generalizability of computational models. For instance, Yang and Hu (2012) evaluated to what extent classification models trained with Western songs can be applied to a set of Chinese songs. The last round of MIREX also started a new task on cross-cultural mood classification on a dataset of K-Pop songs (Hu et al., 2014). In the front of mood regression, there are even fewer studies involving non-Western music, partially due to the lack of datasets with annotations in dimensional models. Yang et al. (2008) conducted mood regression experiments on a dataset of 195 Western, Chinese and Japanese songs. Later, they evaluated a new approach on a set of 1,240 C-Pop songs (Yang & Chen, 2011b). To the best of our knowledge, this is the largest non-Western music dataset for MMR but it only contains annotations on valence dimension. A recent study (Hu & Yang, 2016) investigated the cross-cultural and cross-dataset generalizability of mood regression models, using the dimensional annotation of CH818 as one of the datasets. Notwithstanding the contributions of these previous studies, there is a sore lack of MMR studies based on both categorical and dimensional models. The next subsection will summarize existing available datasets for MMR, further illustrating the research gaps.

### 2.3 Existing Datasets in MMR

Through the development of MMR, a few datasets have been made available for the research community. Table 2 summarizes key characteristics of them.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th># of songs</th>
<th>Music</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIREX AMC dataset</td>
<td>(Hu et al., 2008)</td>
<td>600</td>
<td>Western</td>
<td>Categorical: five mood clusters (Table 1)</td>
</tr>
<tr>
<td>CAL500</td>
<td>(Turnbull et al., 2008)</td>
<td>500</td>
<td>Western</td>
<td>Categorical: 18 mood-related social tags</td>
</tr>
<tr>
<td>MIREX ATC dataset</td>
<td>(Hu et al., 2009)</td>
<td>3,469</td>
<td>Western</td>
<td>Categorical: 18 groups of mood-related social tag</td>
</tr>
<tr>
<td>MER60</td>
<td>(Yang &amp; Chen, 2011a)</td>
<td>60</td>
<td>Western</td>
<td>Dimensional: valence, arousal</td>
</tr>
<tr>
<td>MoodSwing</td>
<td>(Speck et al., 2011)</td>
<td>240</td>
<td>Western</td>
<td>Dimensional: valence, arousal</td>
</tr>
<tr>
<td>NTUMIR-1240</td>
<td>(Yang &amp; Chen, 2011b)</td>
<td>1,240</td>
<td>C-Pop</td>
<td>Dimensional: valence</td>
</tr>
<tr>
<td>AMG1608</td>
<td>(Chen et al., 2014)</td>
<td>1,608</td>
<td>Western</td>
<td>Dimensional: valence</td>
</tr>
<tr>
<td>MIREX K-POP dataset</td>
<td>(Hu et al., 2014)</td>
<td>1,892</td>
<td>Korean</td>
<td>18 mood tag groups&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> ATC stands for Audio Tag Classification task in MIREX;<br> <sup>b</sup> the same tag groups as those in the MIREX ATC dataset.

All the above datasets are available to MIR researchers either through direct downloading of audio features extracted from the music files or by participation in MIREX which runs in an “algorithm to data” paradigm (Downie et al., 2014). All but the NTUMIR-1240 and MIREX K-POP datasets consist of Western music. In addition, only the last two datasets contain both categorical and dimensional annotations. However, in both datasets, annotations on different music pieces may be provided by different annotators. This may limit their utility in some research topics such as personalized recommendation or classification. Compared to existing datasets, the CH818 dataset distinguishes itself in the following...
aspects: 1) it consists of a significant amount of C-Pop music; 2) it is annotated with both categorical and dimensional models; 3) it has three sets of annotations given by the same three annotators throughout the dataset; 4) it comes with a comprehensive set of audio features extracting with well-known MIR tools; and 5) it has a maintenance plan in place for long-term access.

3. THE CH818 DATASET

3.1 Data Collection

The CH818 dataset contains 818 C-pop songs released in Taiwan, Hong Kong and Mainland China from 1987 to 2010. Attempting to reflect the current trend of C-Pop, we collected C-Pop albums released in recent years and listed in hit boards (and thus were influential). To diversify songs in the dataset, we randomly selected one song from each of the 818 albums collected. As the mood of a song can change during its course, and the topic of mood variation across time is beyond the scope of this study, a 30-second excerpt from each song was extracted, for the purpose of mitigating the effect of changing moods in a single song (Hu et al., 2014) as well as alleviating the cognitive load of annotators. For the purposes of MMR, the segment with the strongest emotion from each song was automatically detected and extracted (using the sum of square of predicted valence and arousal values). The detection algorithm was based on a regression model built on an external dataset of Western Pop music, as there were no existing dataset of Chinese songs with proper valence and arousal annotations that could be used to train the regression model.

The annotations were collected from three music experts who were Master students or senior undergraduates in a music school at the time of annotation. All were born and raised in Mainland China and thus are regarded as having a Chinese cultural background. After a training session, each of them annotated all 818 songs independently, in both categorical and dimensional representation models.

For categorical annotation, a set of 36 mood labels were used. Out of these, 29 labels were adapted from the five mood clusters in the MIREX AMC dataset (Table 1). Seven additional labels deemed to be representative in C-Pop were also adopted (Yang & Hu, 2012), including "tender", "soothing", "calm/peaceful", "relaxed", "dreamy", "nostalgic" and "encouraging" (Chow & de Kloet, 2010). Each song could be annotated with one or more mood labels, to reflect the realistic situations where each individual song could express multiple moods (Yang & Chen, 2012; Lee & Hu, 2014). All labels were translated into Chinese, with the English originals presented alongside, to avoid possible misunderstanding introduced in translation.

For dimensional annotations, we followed the literature and adopted Russell’s model of valence and arousal dimensions. The values ranged from -10 to 10 and were annotated using two separate slid-bars. The instructions explained what valence and arousal meant and additional notes were presented at both ends and the middle (neutral) point of the slide-bars. Figure 2 illustrates the annotation interface of one clip. On average each annotator spent 25.67 hours and was paid about 15 U.S. dollars per hour.
3.2 Song Characteristics

All but 12 songs in CH818 were released in the first decade of this century, with around 150 songs each year from 2005 to 2009. There are approximately 390 songs with male and female voices respectively, as well as 39 songs with mixed voices (produced by bands or duets). Most of the songs were in Mandarin, while 17 were in Minnanese (Hokkien) and 15 were in Cantonese. It is acknowledged that, despite of our best efforts in collecting and selecting the songs, the dataset may not be comprehensive enough to represent all C-Pop music. Nevertheless, the recency and popularity of songs makes it highly relevant for research and practical purposes.

3.2 Categorical Annotations and Reliability

On average each song received 10.20 labels, with a standard deviation of 2.67. The most popular mood labels applied were "tender" (1,027 times), "dreamy" (658) and "relaxed" (647), followed by "rousing" (645) and "passionate" (629). The least applied labels include "volatile" (18), "autumnal" (59), "nostalgic" (60), "fiery" (67) and "aggressive" (76).

To quantify the inter-rater reliability across annotators, we calculated the agreement ratio between each pair of annotators. In CH818, each song could be annotated with multiple mood labels by each annotator. Agreement ratio between two sets of annotations on a song is calculated as the number of identically tagged labels divided by the number of all labels applied to the song (Nowak & Rüger, 2010; Hu & Lee, 2012). The agreement ratios of all songs are then averaged to show an aggregated value. The averaged agreement ratio in CH818 is 0.37 (Table 3, first row). It is comparable to the agreement ratio of 0.35 between Chinese listeners on Western songs in (Lee & Hu, 2014) and higher than the ratio of 0.15 between Korean listeners on K-Pop songs (Hu et al., 2014).

To improve the reliability of the annotations, we calculated the majority voted annotations on each song. That is, only mood labels applied two or three times to a song are counted. In this way, the average number of labels for each song is 4.03, with a standard deviation of 1.96. The number of songs with each mood label is presented in Table 5, and the agreement ratios are shown in Table 3 (second row) which are much higher than those between the annotators (first row). In the subsequent analysis, we will use the majority voted categorical annotations unless otherwise specified.

Table 3: Agreement ratio between each pair of annotations.

(A1, A2, A3 represents the three annotators)
3.3 Dimensional Annotations and Reliability

Figure 3 illustrates the scatter plot of the annotations in the valence-arousal space. There are more values in the first quadrant than others, indicating songs with positive valence and positive arousal annotations are the most popular in the dataset. There are no songs in the bottom right corner of the space (near [10, -10]), indicating high valence and low arousal values. These observations are consistent with the MER60 (Yang & Chen, 2011a) and AMG1608 (Chen et al., 2015) datasets, which both consist of Western Pop songs.

Following existing research (Yang & Chen, 2011a; Hu et al., 2014), we measured agreement across annotators on dimensional annotations using Krippendorff’s alpha (Gwet, 2010). The results were 0.49 and 0.62 for valence (V) and arousal (A) respectively, which fall in “fair” and “moderate” agreement (Gwet, 2010). Similar to the findings in previous studies (Kim et al., 2010; Yang & Chen, 2012; Hu et al., 2014), valence annotations are harder to reach agreement than arousal. The alpha values of CH818 are higher than those of AMG1608 (V: 0.31, A: 0.46) (Chen et al., 2014) and the MIREX K-POP mood dataset (V: 0.28, A: 0.54) (Hu et al., 2014)³, while comparable to those of MER60 (V: 0.39, A: 0.70).

It is noteworthy that due to the subjectivity of music mood, inter-annotator agreement for music mood is usually moderate. We have thus split CH818 into three subsets with controlled annotation reliability levels. Initial experiment has shown better prediction performances on subsets with higher reliability levels which is in accordance with the results in (Hu & Yang, 2016). The subsets are also released with CH818 to facilitate further studies.

3.4 Audio Features

To capture a range of musical characteristics that are recognized to be related to music mood (Juslin, 2000), five categories of acoustic features were extracted from the songs in CH818 using tools that are well recognized in the MIR field. These features can be broadly categorized into loudness, pitch, rhythm, timbre and harmony. Table 4 lists the features, their dimensionality and the music audio processing tools that are used to extract them. For specific descriptions of the features, please refer to (Hu & Yang, 2016).

Table 4. Extracted audio features (Hu & Yang, 2016).
(M: MIR toolbox (Lartillot & Toiviainen, 2007), P: PsySound (Cabrera, 1999), T: Tempogram toolbox (Grosche & Müller, 2011) and C: Chroma toolbox (Müller & Ewert, 2011))
3.5 Maintenance of the Dataset

The metadata, three sets of annotations in both categorical and dimensional models, and audio features presented will be publicly available for research purposes. Due to copyright restrictions, the audio clips cannot be shared. To compensate for this disadvantage, we plan to continue extracting new audio features, and add them to the dataset. For this purpose, we welcome MIR researchers to share their feature extraction programs with us so that we can run the programs against the audio files of CH818 and share the newly extracted features by adding them to the webpage of the dataset.

It is our plan to maintain the datasets for long-term availability and values (Donnelly, 2014). While preserving research datasets in MIR is beyond the scope of this paper, we are seeking opportunities to deposit stable versions of the CH818 dataset to a data or institutional repository for the sake of long-term sustainability.

4. RQ1: MOOD SPACES OF CHINESE AND WESTERN POP MUSIC

The mood space of C-Pop is constructed using the categorical annotations on CH818. It is then compared to that of Western Pop music (as represented in (Yang & Hu, 2012)) in three aspects: 1) distributions of mood labels; 2) relative distance among mood labels; and 3) mood clusters.

4.1 Mood Space of CH818

To visualize the mood space of CH818, we projected the mood labels based on their distances to a 2-D space using multidimensional scaling (MDS), the same method used in (Yang & Hu, 2012). The distance between a pair of mood categories was calculated based on the common songs shared by them in the majority voted categorical annotations. As shown in Figure 4, the proximity of labels reflects their semantic closeness in most cases (e.g., “aggressive” is close to “volatile” and away from “dreamy”).
Inspired by the MIREX five-cluster mood model (Table 1) where individual labels were merged into clusters, we also constructed a set of mood clusters for CH818, using the same agglomerative clustering technique used on the MIREX AMC dataset (Hu et al., 2008). The resultant dendrogram showed six distinct clusters, but two labels, “autumnal” and “nostalgic” were only merged in very late iterations (i.e., they are not very close to any clusters). Therefore, these two labels were excluded from the resultant clusters. This exclusion did not result in any songs being excluded because only few songs were annotated with them and all those songs also had other labels. As shown in Table 5, overall labels in each of the clusters are semantically consistent. Among the clusters, C_6 contains five of the seven labels added in this study. The numbers of songs in the clusters are quite balanced except for C_5 and C_6, which may reflect that there are more calm and tender C-Pop songs than aggressive ones (Hu, 2014; Hu & Lee, 2016).

Table 5. Clustering results of mood labels in CH818.

<table>
<thead>
<tr>
<th>Cluster (#. of songs)</th>
<th>Mood labels (#. of songs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH818_C_1 (293)</td>
<td>Passionate (189), rousing (212), confident (181), boisterous (86), encouraging (86), rowdy (41)</td>
</tr>
<tr>
<td>CH818_C_2 (219)</td>
<td>Rollicking (51), cheerful (72), fun (95), sweet (51), amiable/good natured (63), humorous (32), witty (41)</td>
</tr>
<tr>
<td>CH818_C_3 (206)</td>
<td>Pignant (50); brooding (166); bittersweet (81)</td>
</tr>
<tr>
<td>CH818_C_4 (194)</td>
<td>Silly (66), campy (52), quirky (68), whimsical (142), wry (13)</td>
</tr>
<tr>
<td>CH818_C_5 (75)</td>
<td>Aggressive (8), fiery (13), volatile (2), tense/anxious (24), intense (58), visceral (61)</td>
</tr>
<tr>
<td>CH818_C_6 (506)</td>
<td>Calm/peaceful (143), tender (361), relaxed (194), dreamy (181), soothing (130), literate (101), wistful (173)</td>
</tr>
</tbody>
</table>

4.2 Cross-cultural Comparison

With a substantial number of C-Pop songs and thorough mood annotations, CH818 provides a representative case for exploring the mood space of C-Pop. The space can then be compared to that of music from other cultures, providing systematic insights on how music moods are similar or different across cultures. As the mood labels in this study are originally from AllMusicGuide, the Western song dataset in (Yang & Hu, 2012) is used for comparison, which consists of 1,520 “Top Songs” associated with mood labels on AllMusicGuide. Figure 5 shows the number of times each mood label is applied to both da-
It should be noted that in AllMusicGuide, each mood label can at most have 100 top songs. Nonetheless, some patterns of differences are clear to see: “dreamy,” “relaxed” and “calm” are popular in CH818 but not in the Western dataset whereas “wry,” “aggressive” and “fiery” are the opposite. This observation may be attributed to the songs. Influenced by the Chinese culture which values modesty and introversion, Chinese music may favor music elements that sound mellow, mellifluous and light-spirited (Liu & Mason, 2010; Hu & Lee, 2016).

The mood space of CH818 shown in Figure 4 can be compared to that of the Western songs in (Yang & Hu, 2012, reprinted in Figure 6). In both spaces, labels sharing similar semantics are gathered together, such as the group of “aggressive,” “fiery,” and “volatile” and that of “confidence,” “passionate” and “rousing”. The two spaces also differ in several ways. The group of “aggressive” and that of “confident” are close in the Western space but far apart in the CH818 space. This is possibly related to Chinese people’s tendency of disliking radical moods (e.g., “aggressive”). In addition, the middle parts of both spaces are crowded by many labels, but the group of “relaxed”, “dreamy”, “calm” and that of “humorous”, “silly”, “quirky” are separated from the middle in the two spaces respectively, indicating they represent quite distinct moods in each culture, but not the other.

We also compare the mood clusters of CH818 (Table 5) to that of the MIREX AMC five-cluster representation model (Table 1). The two sets of clusters are comparable as both were derived from expert annotations using similar methods. The Comparison discloses some commonalities: 1) the fifth clusters (C_5) in both sets completely match; 2) all labels in the first clusters (C_1) match except for “encouraging” which is new in CH818 and shares similar semantics with other labels in C_1; 3) C_2 and C_4 in both spaces are very similar, except two labels, “humorous” and “witty” are moved from C_4 of the MIREX model to C_2 of the CH818 model. In fact, a previous study (Lee & Hu, 2012) has found that MIREX_C_2 and MIREX_C_4 were indeed the most confusing pairs for MIREX evaluators, which can at least patri-
This is a pre-print version of the following article: Hu, X. & Yang, Y.-H. (In press). The mood of Chinese Pop music: Representation and recognition. *Journal of the Association for Information Science and Technology*.

...cianly explain the change of cluster membership of the two labels; 4) All three labels in CH818_C_3 are in MIREX_C_3, while two labels in the latter, "literate" and "wistful" move to CH818_C_6, which is unique to CH818. The biggest difference between the two sets of clusters is probably the new emergence of CH818_C_6. This suggests that calm and relaxed moods are more prominent in C-Pop than Western Pop songs.

5. RQ2: RELATIONSHIP BETWEEN CATEGORICAL AND DIMENSIONAL MOOD MODELS

5.1 Mapping between the models

It is well recognized that both categorical and dimensional models of music mood have their own advantages and disadvantages, and thus it is desirable to have both annotations. Alternatively, as human annotations are expensive and time-consuming to obtain, in a more realistic sense, it would be helpful if one set of annotations could be inferred from the other. Given that both kinds of models are widely used in the field, MIR researchers have been interested in the relationships between them (Wang et al., 2012). As the first C-Pop dataset with annotations in both categorical and dimensional models, CH818 provides an excellent opportunity to investigate the mappings between them.

Figure 7 illustrates the 36 categorical labels in the valence-arousal (VA) 2-D space where the position of each label is decided by the mean valence and arousal values of all songs annotated with that label. The distribution of labels in Figure 7 is in accordance with song distribution plotted in Figure 3, in that most songs and labels fall in the first and third quadrants (+V+A and −V−A). In addition, the relative positions of the mood labels in the VA space generally match the cluster compositions shown in Table 5. These observations support that the two sets of annotations in this dataset are well connected. For instance, "passionate," "confident" and "rousing" are placed together (C_1 in Table 5) in the first quadrant while "wistful," "soothing" and "literate" (C_6 in Table 5) cluster near the zero point (with slightly negative valence and arousal). It is noteworthy that the arousal axis in Figure 7 is slightly shifted to the left compared to Figure 1 which is based on Western music. Some mood categories such as "aggressive" are of negative valence in Russell’s model (Figure 1) but carry positive values here in Figure 7. This difference may be attributed to the sparseness of aggressive C-Pop songs and/or Chinese listeners’ low threshold for perceiving a piece as “aggressive”. Future studies are warrant to further verify the reasons.

![Figure 7](image-url)

**Figure 7.** CH818 mood labels plotted in valence-arousal space.
5.2 Prediction between models

We conducted experiments to test the extent to which one set of annotations on CH818 could be predicted by the other. A linear regression model was constructed to predict valence or arousal annotations of the songs based on the binary variables of the categorical labels. The $R^2$ of the two models are 0.53 for valence and 0.70 for arousal which are comparable or even higher than regression performances based on audio features (Kim et al., 2010; Yang and Chen, 2011). A closer examination of the significant predictor variables (i.e., the categorical labels) in both models reveals that those unique to each model do bear semantics indicating valence or arousal respectively. For example, "volatile," "cheerful" and "encouraging", which clearly indicate negative or positive sentiment, are unique to the valence prediction model. In contrast, labels with energy implications such as "fiery," "boisterous," "calm/peaceful," and "dreamy" are unique to the arousal model.

For classification of mood labels, we combined the labels into clusters shown in Table 5, to reduce the number of models and to avoid problems caused by data sparseness. Six logistic regression models were built based on the valence and arousal values to classify whether or not a song belongs to the clusters. Compared to a baseline of random chance (accuracy = 50%), the results show that CH818_C_1, C_3, C_5 and C_6 are highly predictable (accuracy = 89.73%, 88.14%, 93.64% and 87.90%, respectively). C_2 and C_4 performed less ideal (accuracy = 74.82% and 75.31% respectively), which again is probably related to the fact that the terms in these two clusters were somewhat confusing to human annotators (Lee & Hu, 2012). The results are comparable to audio-based classifiers (See next section) and those reported in the literature (Kim et al., 2010; Yang and Chen, 2011). It is particularly encouraging given the fact that there are only two predictor variables (i.e., valence and arousal).

The experiments show that the relationship between the two sets of annotations in CH818 is reliable. Annotations of one model could predict those of the other. This makes it one step further towards the goal of unified user interfaces that can benefit from both mood representation models. To this end, CH818 itself could serve as a ground truth on which various techniques can be evaluated. Furthermore, this could be in conjunction with cross-cultural and cross-dataset experiments, to find out whether and to what extent the predictability between annotation models can transcend across such boundaries as culture, genre, etc.

6. RQ3: MUSIC MOOD RECOGNITION (MMR) FOR C-POP

With both categorical and dimensional annotations, as well as rich audio features extracted (Table 4), CH818 can serve as a benchmarking dataset for evaluating the tasks of mood classification (predicting mood labels) and mood regression (predicting valence and arousal values) for C-Pop songs. To illustrate the utility of CH818 in these tasks and benchmark the performances, we conducted two sets of classification and regression experiments on this dataset. The first was to evaluate the classification and regression performances on the combination of all extracted audio features. The second was to compare the performances of the five sets of features summarized in Table 4. Performances are compared to the state-of-the-art of MMR on Western music.

6.1 MMR on All Extracted Features

As Support Vector Machines (SVM) have been widely adopted in MMR studies and shown superior performances, we used SVM classification and regression (SVR) models. For classification, we used the clusters in Table 5 as the prediction target. As each song may be in multiple clusters, we constructed the problem into a binary one. That is, one classifier was built for each cluster, with positive examples being songs labelled with any terms contained in this cluster. The equal number of negative examples was then randomly selected from the remaining songs 4. The sampling process was repeated 20 times and the averaged results were reported. The measures of accuracy and Cohen’s Kappa are used to gauge the performances. For binary classification with balanced data, a trivial prediction by chance would result
in an accuracy of 50% and a Kappa of 0. For regression, the valence and arousal values averaged across annotators were used as the ground truth. The measures of $R^2$ and root mean squared error (RMSE) are calculated. Both the classification and regression experiments used the RBF kernel with default parameter settings, and were conducted with 10-fold cross validation. The results are shown in Table 6.

Table 6. Classification and regression experiment results.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Accuracy</th>
<th>Kappa (κ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH818_C_1</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td>CH818_C_2</td>
<td>0.74</td>
<td>0.50</td>
</tr>
<tr>
<td>CH818_C_3</td>
<td>0.82</td>
<td>0.64</td>
</tr>
<tr>
<td>CH818_C_4</td>
<td>0.81</td>
<td>0.63</td>
</tr>
<tr>
<td>CH818_C_5</td>
<td>0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>CH818_C_6</td>
<td>0.92</td>
<td>0.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension</th>
<th>$R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>0.25</td>
<td>3.54</td>
</tr>
<tr>
<td>Arousal</td>
<td>0.79</td>
<td>2.31</td>
</tr>
</tbody>
</table>

The classification performances on C_1, C_5, and C_6 are fairly good, not only with high accuracies of over 90%, but also with kappa values in the “very good agreement” level (κ = 0.80 – 1.00) with the ground truth (Gwet, 2010). The accuracy levels of these clusters are also close to the latest result of the MIREX Audio Tag Classification (ATC) task (90%) where the task was also binary classification. The performances on C_3 and C_4 are in “good agreement” with the ground truth (κ = 0.60 – 0.79), while C_2 is only in the “moderate agreement” (κ = 0.40 – 0.59). The less satisfactory performance on C_2 may be due to the inherent difficulty of prediction on some moods from audio features (e.g., the “happy” group in the ATC task). The performances of regression are comparable to the literature where $R^2$ is usually from 0.17 to 0.30 and 0.58 to 0.80 for valence and arousal respectively (Yang & Chen, 2012; Guan, Chen, & Yang, 2012; Kim et al., 2010). These results verified the feasibility of evaluating music mood recognition with CH818 as well as the applicability of the audio features proposed from Western music on C-Pop songs.

6.2 Comparison of Feature Sets

To further investigate the relative advantages of the different feature sets on C-Pop songs, we conducted similar classification and regression experiments on each of the five feature sets: loudness, pitch, rhythm, timbre and harmony. Figures 8 and 9 show the results. For classification, timber and pitch features performed better than other features, and such advantages are more obvious on the worse performing clusters, C_2 and C_3. Timber features have been shown effective in MMR of Western music (Barthet et al., 2013; Yang & Chen, 2012). The fact that pitch features (designed from Western music) worked well also evidences that C-Pop is influenced by the 12 pitch class structure in Western music. For regression, timber, pitch and loudness features were good for arousal predictions, whereas rhythm and timber features were good for valence prediction. These results are in accordance to those in (Hu & Yang, 2016) where similar features were also evaluated on Western music. The applicability of state-of-the-art audio features on C-Pop not only supports the evolutionary relationship between C-Pop and Western Pop (Liu & Manson, 2010), but also inspires further research on comparative studies of different feature sets on music of different cultural origins.
7. CONCLUSIONS AND FUTURE WORK

This study investigated a few long-standing questions on music mood recognition, using the CH818 dataset of C-Pop music. The comparison of mood spaces between Chinese and Western Pop music demonstrated the high consistency between the two, as well as the distinct and salient cluster of "calm", "soothing", "relaxed" mood in C-Pop music. Such a comparison can also be made with music from other cultural origins. Collectively, the results will show a more complete picture of the mood spaces of music in the world and contribute towards the goal of global access to music. The mapping between categori-
cal and dimensional models verifies the internal consistency of the two kinds of models which provides a theoretical base of computational work on automated prediction of one representation from the other.

The applicability of MMR techniques for both classification and regression on CH818 is verified by a set of experiments. The comparison of performances across different audio feature sets helps answer the question of which features are good for the prediction of different mood categories or dimensions for C-Pop.

It is noteworthy that the songs in CH818 span mainly across one decade and thus may not be sufficiently comprehensive to represent the entire landscape of C-Pop. Nonetheless, by exploring the raised research questions using CH818, we hope this study can stimulate further explorations and innovations related to MMR. We will further develop the CH818 dataset by adding more dimensions of information into it, such as lyrics, metadata and/or editorial/social tags, so that it can be used for multi-modal MMR (Hu et al., 2016) and to help explore the question of the effects of lyrics and melodies on music mood (Ali & Peynircioğlu, 2006). As datasets are foundations of development of a field, it is also an intriguing and important direction to explore how an MIR data repository can be built up. This will be beneficial to the entire MIR community.

ACKNOWLEDGEMENT
This study is partially 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.

FOOT NOTE
1http://www.allmusic.com/
2The dataset is accessible at http://ccmir.cite.hku.hk/data/CH818/
3The MIREX K-Pop mood dataset reported intra-class correlation (ICC) with the one-way random model which is equivalent to Krippendorff's alpha (Gwet, 2010).
4The random sampling process is reversed for CH818_C_6, as there are more positive examples than negative ones.
5http://nema.lis.illinois.edu/nema_out/mirex2014/results/atg/subtask2_report/bin/summary.html

REFERENCES


