CROSS-CULTURAL MUSIC MOOD CLASSIFICATION: A COMPARISON ON ENGLISH AND CHINESE SONGS

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

Xiao Hu
University of Denver
xiao.hu@du.edu

ABSTRACT

Most existing studies on music mood classification have been focusing on Western music while little research has investigated whether mood categories, audio features, and classification models developed from Western music are applicable to non-Western music. This paper attempts to answer this question through a comparative study on English and Chinese songs. Specifically, a set of Chinese pop songs were annotated using an existing mood taxonomy developed for English songs. Six sets of audio features commonly used on Western music (e.g., timbre, rhythm) were extracted from both Chinese and English songs, and mood classification performances based on these feature sets were compared. In addition, experiments were conducted to test the generalizability of classification models across English and Chinese songs. Results of this study shed light on cross-cultural applicability of research results on music mood classification.

1. INTRODUCTION

There have been a number of studies on music mood classification in the Music Information Retrieval (MIR) community in recent years [7][17]. However, most of existing studies have focused on Western music, in particular English songs. The two mood-related tasks in the Music Information Retrieval Evaluation eXchange (MIREX): Audio Mood Classification (AMC) and Audio Tag Classification (mood tag set) have been using datasets consisting of Western music [3]. Although such research activities have shown promising performances on classifying Western music by mood, there is little research on whether and how the mood categories and techniques applied to Western music can be equally well applied to non-Western music. This study aims to bridge the gap using Chinese contemporary pop songs as a case of non-Western music. In particular, three research questions are answered in this study:

1) How well mood categories developed from English songs can be applied to Chinese songs and what are the differences of mood distributions among Chinese and English songs?
2) Are audio features commonly used in mood classification of Western songs applicable to Chinese songs?
3) Can prediction models built using English songs be reliably applied to Chinese songs?

Answers to these questions will further our understanding on cross-cultural generalizability of music mood categories, audio features and classification models.

2. RELATED WORK

2.1 Mood Categories in Western and Chinese Music

In building datasets for evaluating mood classification algorithms, MIR researchers have used a variety of mood categories. Quite a number of studies have used four categories derived from the four quadrants of Russell's *valence-arousal* dimensional model [12]: *contentment, depression, exuberance, and anxious/frantic* (or similar ones). It is noteworthy that these four categories have been used for both Western music [10][14] and non-Western (e.g., Chinese) music [4][14].

In the industry, online music repositories may organize music by mood. For example, allmusic.com, a large and influential online repository for Western popular music, provides 182 mood labels that are applied to songs and albums by professional music editors. In fact, the mood categories used in the AMC task in MIREX (cf. Table 1) were based on the most popular mood labels on allmusic.com [3]. However, no research has been done to investigate whether these mood labels are suitable for non-Western music. In this study, we take on this challenge using Chinese pop music as a case (see Section 3).

2.2 Mood Classification of Western Music

Quite a number of studies have been conducted on automated mood classification on Western Music, as reviewed in [7][17]. Most of these studies extracted acoustic features from music audio files. Some studies combined acoustic features with features extracted from other information sources such as lyrics and social tags [2][9]. As one of the first studies comparing mood classification techniques on Western and non-Western music, this paper focuses on acoustic features and leave it to future work to compare approaches using combined information modals.

The classification models often used include neural network, k-nearest neighbor (k-NN), maximum likelihood, decision tree, and support vector machines (SVM). Of these, SVM seems the most popular due to its reliable classification performance. In this study, we use SVM as

the classification model for all the experiments as our focus is on the generalizability of acoustic features.

2.3 Mood Classification of Chinese Music

There have been few studies on mood classification of non-Western music and they are predominately on Chinese music. Hu *et al*. [4] and Xia *et al.* [15] classified Chinese pop songs using lyric features. Yang *et al*. combined lyrics and audio features to classify Chinese pop songs [17]. It is noteworthy that all these previous studies used four or fewer mood categories (e.g., two categories were used in [15], "lighthearted" and "heavy-hearted."). Moreover, none of these studies compared Chinese music to Western music on either mood categories or mood classification techniques and performance.

2.4 Comparison between Mood Classification on Western and Chinese Music

To the best of our knowledge, [14] is the only existing study that evaluated the same mood classification techniques on both Western and Chinese music. Wu *et al.* [14] extracted three acoustic feature sets of pitch, rhythm and timbre from 20 pieces of Chinese traditional instrumental music and 20 pieces of Western classical music. They used a Bayesian probabilistic structure to classify the moods in these pieces into four mood categories based on the Russell model and the results indicated better classification performance on Western pieces. At the same time, the authors found "different identification about the moods between the Oriental and Western culture" (p.152).

Our study differs from this prior work in the following aspects. First, the music we focus on is popular music from Western and Chinese cultures. Second, while [14] used four mood categories, we compare more than 30 mood labels in terms of their distributions in Chinese and English songs. Third, the dataset in [14] was evidently too small (20 pieces in each culture) to draw reliably conclusions, whereas our study uses a dataset of 500 Chinese pieces and even more English pieces. Fourth, in addition to the rhythm and timbre features examined in [14], our study also examines dynamics, MFCCs, psychoacoustic and tonal features. Fifth, besides comparing classification performances on music in each culture, our study also investigates the generalizability of classification models built with music in one culture being applied to music in another culture (cf. Section 4.2).

3. MOOD CATEGORIES IN CHINESE SONGS

This section describes how we determine the mood categories and obtain the mood annotations of Chinese songs from the experts we recruited and those of English songs from allmusic.com.

3.1 Allmusic.com Mood Labels and Translation

To answer our first research question, we chose mood labels contained in the five mood clusters used in the AMC task in MIREX [3]. This is due to the following reasons. First, the AMC mood clusters were derived from the most popular mood labels on allmusic.com, one of the most popular sites for Western music, and thus they are representative to the mood of Western songs. Second, allmusic.com provides "top songs" for each of its mood labels, which empowers us to obtain English songs with expert-annotated mood labels. Third, the 29 mood labels in AMC mood clusters (see Table 1) are of finer granularity than the commonly used four categories based on Russel's model [12]. Previous studies based on the four categories were not able to discern the difference of mood distributions among Western and Chinese songs [14]. Fourth, as pointed out by music psychology research [6], classical models like Russell's may not reflect the social context of everyday music listening. Since allmusic.com serves a large quantity of listeners in real life, its mood labels are closer to the reality of music listening at present time.

C1	Rowdy, rousing, confident, boisterous, passionate
C2	Amiable/good natured, sweet, fun, rollicking, cheerful
C3	Literate, wistful, bittersweet, autumnal, brooding, poignant
C4	Witty, humorous, whimsical, wry, campy, quirky, silly
C5	Volatile, fiery, visceral, aggressive, tense/anxious, intense

Table 1. Mood categories used in MIREX AMC task [3].

It is noteworthy that, in discussions with music experts (see below), seven additional labels were added in because the experts believed they might be important to represent moods in Chinese pop songs: "calm/peaceful," "dreamy", "encouraging," "nostalgic," "relaxed," "soothing," and "tender," making 36 mood categories in total. Except for "encouraging" and "tender", the other five labels had appeared on allmusic.com at various time points (as allmuic.com changes its mood labels from time to time).

The mood categories were translated into Chinese for consistent understanding among the annotators who are native Chinese speakers. The translation was first done by one of the authors who is a native Chinese speaker and fluent in English. The translation and the original English terms were then examined by three expert annotators. The four of them discussed several difficult cases (e.g., "autumnal," "visceral") before reaching agreements on the translations.

3.2 Selection of Chinese Songs

The Chinese pop songs were collected from an in-house collection which contains albums released in Taiwan, Hong Kong and Mainland China during the years from 1987 to 2010. To maximize the diversity of songs, we selected one song from each album to be included in this study. Songs with non-Chinese (mostly English, some in Japanese) titles were eliminated because they were not in Chinese despite being sung by Chinese artists. This process resulted in 500 Chinese songs.

An excerpt of 30 seconds was extracted from each song, for the purposes of limiting the burden of human annotators and mitigating the cases where mood changes during the entire course of a song [7]. Using 30 second excerpts rather than entire songs is a common practice in MIR, but it remains in debate which 30 second segment should be chosen [17]. For the purpose of mood classification, we decided to choose the segment with strongest emotion from each song. Specifically, we used a sliding window of 30 seconds to exhaustively extract all 30 second segments from each song, and then used a regression model to predict the *valence* and *arousal* values of each segment. The segment with highest ($|valence|^2$ + $|arousal|^2$) value was chosen to represent each song. Please note that the regression model was built on an external dataset of Western pop music [16], as there were no Chinese songs with proper *valence* and *arousal* annotations that could be used to train the regression model.

3.3 Annotation of Chinese Songs

To ensure the quality of annotation and to reduce the variance of annotations across annotators, this study adopts the approach of expert annotation.

Three experts were recruited from a university in the United States. All of them were female, Chinese (Mandarin) native speakers, raised in Mainland China, majored in music, and fans of Chinese pop music. Table 2 shows their demographic and background information.

ID	Age	Specialty	Year in college	Years in the US	Freq. of listening to Chinese pop music
1	23	Theory	Senior	4	Several times a week
2	25	Violin	Graduate	0.5	Daily
3	25	Vocal	Graduate	3	Several times a week

Table 2. Information about the experts.

The annotation was conducted through a web-based survey system. Figure 1 shows the interface of annotating a piece. One or more mood labels could be applied to each music piece. This is more in accordance to the reality where a song can express multiple moods [2].

Figure 1. Screenshot of the annotation interface.

Before the annotation started, all three experts met together with one of the authors for a training session. Annotation requirements were made clear during the session, including "focusing on the mood expressed by the pieces instead of mood induced in yourselves;" "making use of the lyrics but without looking them up anywhere;" "ignoring the order by which the mood labels are arranged." Also in the training session, the experts listened to eight 30 seconds long pieces and discussed which moods each piece expressed. Half of the eight pieces were English songs with very different moods selected from a previous study [16] and the other half were randomly selected from the Chinese pieces to be annotated. Through the discussion, the experts reached better agreement on musical meanings of the mood labels.

Each piece was assigned to one expert and the three experts annotated 150, 200 and 150 pieces respectively. Using one human judge's opinion as gold standard for evaluation has been verified to be effective in the domain of text information retrieval [13]. Admittedly, this has yet to be verified in MIR, which will be our future work. The experts reported that each song took each of them about 1 minute to annotate and were paid for their work on an hourly basis.

3.4 Mood Categories in Chinese and English songs

A total of 2,453 mood labels were applied to the 500 pieces, with one piece getting 1 to 13 labels. On average, each piece had 5 labels (standard derivation: 2.08). Each of the 36 mood labels were applied to 6 to 202 pieces, with an average of 68.14 pieces each label (standard derivation: 40.85). The most popular mood labels among the 500 Chinese songs were "tender" (202), "wistful" (164), and "passionate" (128). The least popular mood labels were "volatile" (6), "wry" (6), and "nostalgic" (8). The distribution of Chinese songs across mood categories is shown in the left panel of Figure 2.

As comparison, we counted the number of "Top Songs" provided by allmusic.com for each of the mood labels, as shown in the right panel of Figure 2. Note that the numbers of Chinese songs in Figure 2 are limited to the songs available to us. While allmusic.com has a much larger song pool, they only provided up to 100 top songs for each mood label. Despite this, it is still clear from Figure 2 that there are more Chinese songs than English songs labeled with "relaxed," "wistful," "passionate," "brooding," or "rousing." On the other hand, there are much fewer Chinese songs labeled with "wry," "volatile," "humorous," "aggressive," and "fiery." Such difference might reflect the differences between the Chinese and Western cultures: the Chinese culture tends to restrain the expression of feelings and Chinese people are more introverted compared to Western people [11]. Another possible reason for the absent of radical moods (e.g., "aggressive," "fiery") in Chinese songs might be that Chinese popular music has a shorter history comparing to Western one and the whole mood spectrum is yet to be developed.

To further illustrate the applicability of the mood categories to Chinese songs, we also examined and compared the relative distance among mood categories based on the two datasets. The distance between a pair of mood categories was calculated based on the common songs shared

by them. We then projected the mood labels based on their distances to a 2-D space using multidimensional scaling (MDS) for each of the two datasets, as shown in Figure 3. It can be found that the relative positions of mood labels for both sets share some similarity: "aggressive," "fiery," "intense" and "volatile" are clustered together and away from low arousal moods such as "amiable" or "wistful." Another two clusters shared by both plots are also in line with common sense: "rollicking," "cheerful," "passionate" and "rousing"; "literate," soothing" and "bittersweet".

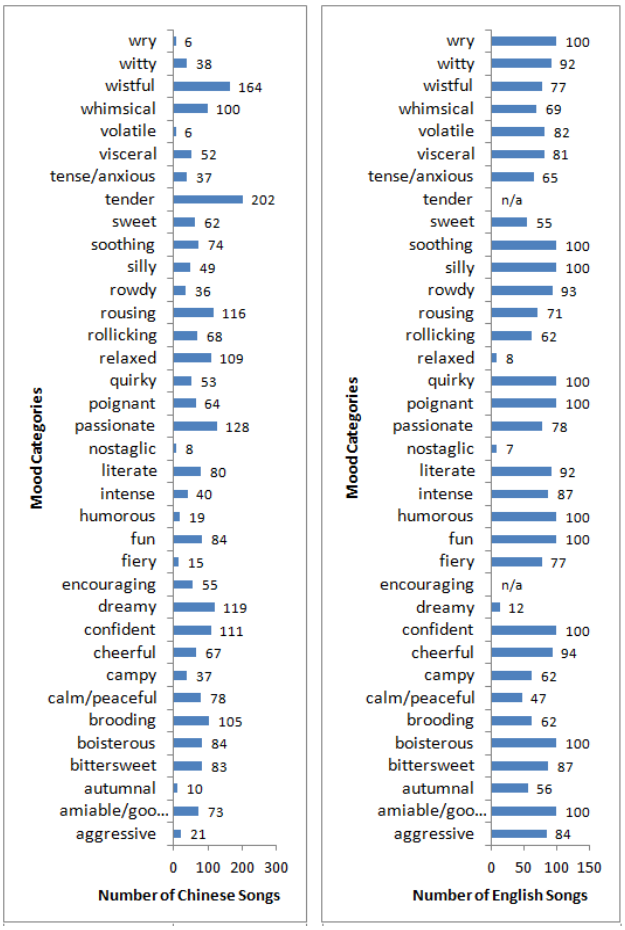
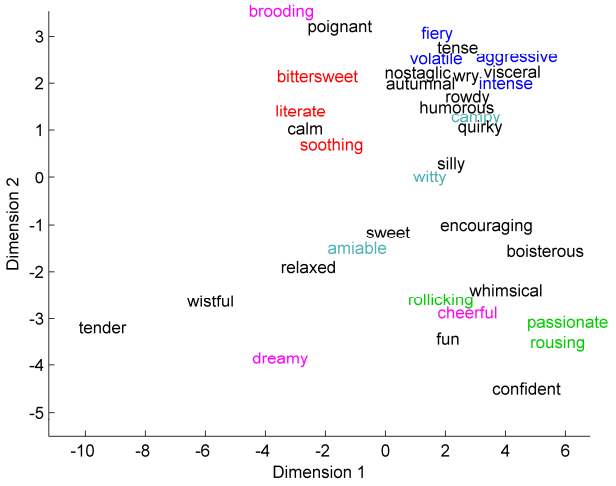
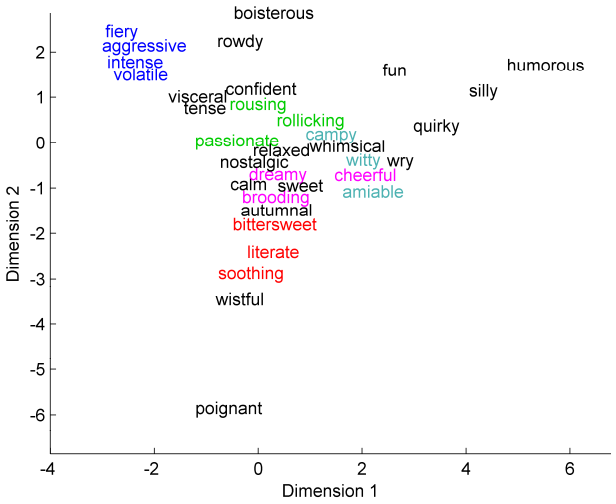


Figure 2. Song distribution across categories.

The two plots in Figure 3 also differ in several ways. "Dreamy," "brooding" and "cheerful" are close to one another in the English dataset but are separated out in the Chinese dataset which seems more intuitive. As another example, the English set puts "campy" and "witty" together with "cheerful" and "amiable" while the Chinese set separates them apart. This difference might be related to the cultural contexts. The Chinese experts interpreted the first two words as neutral and the last two as positive. However, an English native speaker said he would associate the four terms to "fun" although the first two terms were more of a kind of "dark" fun. In sum, the relative distance among mood labels in the Chinese set, although somewhat different from that in the English set, seems agreeable with the semantic meanings of the terms.



(a) Mood distance in the Chinese dataset



(b) Mood distribution for the English dataset

Figure 3. Projection of mood categories to a 2-D space.

From the comparisons of song distributions across mood categories as well as the relative distance between them, we can see that the mood categories used to describe English songs are generally applicable to Chinese songs with exceptions of radical moods such as "fiery" and "volatile."

4. CLASSIFICATION EXPERIMENTS

To answer research questions 2 and 3, we conducted mood classification experiments on Chinese and English songs. To build a dataset of English songs, we collected the "Top song" lists from allmusic.com for mood labels used in this study and then obtained 30-second audio previews from 7digital.com which boasts itself as "a leading digital media delivery company." A total of 1,520 English song clips were collected.

All experiments were set up as binary classifications for each mood category, without considering possible correlations between categories. Positive examples of a mood category are songs labelled with that category while negative examples are randomly selected from songs labelled with other categories. Positive and nega-

tive examples are balanced both in training and test sets. The classification is carried out by SVM, with RBF kernel and the two parameters C and γ tuned. Each experiment was repeated 20 times and average accuracy values are reported. In our evaluation, we only used the mood classes with more than 20 positive examples in both Chinese and English datasets. Throughout the experiments, the datasets were split into 50% training and 50% test.

4.1 Effectiveness of Acoustic Features

In view of the complexity of mood perception, it is difficult to find a universal feature representation that well characterizes every mood. In addition, the perceptions of different moods in music are usually associated with different patterns of acoustic cues [5]. We therefore extracted acoustic features that represent various perceptual dimensions of music listening and trained classifiers using features of each perceptual dimension separately.

A total of six feature sets were examined in this study, as summarized in Table 3. They were extracted using the MIR toolbox [8] and the PsySound toolbox [1]. These features have been used extensively in previous work on mood classification [7][17].

Feature	Type	Dim	Description
RMS	Energy	2	The mean and standard deviation of root mean square energy
PHY	Rhythm	5	Fluctuation pattern and tempo [8]
PCP	Pitch	12	Pitch class profile, the intensity of 12 semitones of the musical octave in Western twelve-tone scale [8]
TON	Tonal	6	Key clarity, musical mode (major/minor), and harmonic change (e.g., chord change) [8]
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 [1]

Table 3. Feature representations adopted in our study.

Figure 4(a) shows the average classification accuracy of the binary classification tasks on the two datasets. The six audio descriptors performed well in both sets and even better for the Chinese songs. The relative performances among the feature sets are similar in both datasets: timbre descriptors performed better than energy or rhythm related descriptors. The fact that PCP and TON features worked well for Chinese songs reflects that the composition of contemporary Chinese pop songs is influenced by the Western twelve-tone scale. In addition, we find that PSY performed the best for both datasets with an average accuracy of 74.7% and 65.8%, respectively. The performance differences between PSY and the first four features sets are significant (pair-wise t-test, $p < 0.001$). This shows that the psychoacoustic features seem to be generally applicable to Chinese songs [1].

It is interesting that significantly better performance (pair-wise t-test, $p < 0.001$) is obtained for Chinese songs than for English songs. Although both datasets were annotated by experts (recruited by allmusic.com and by us, respectively), the allmusic.com song lists contain "tier1" (more representative) and "tier2" (less representative) songs. Our inclusion of tier2 songs may have introduced noise to the English dataset. In contrast, all the Chinese songs were annotated with the same criteria. The performance difference may also result from the fact that the excerpts of the Chinese songs were segments with strong emotion, whereas the English excerpts were provided by 7digital which may not represent the full songs annotated by allmusic.com experts.

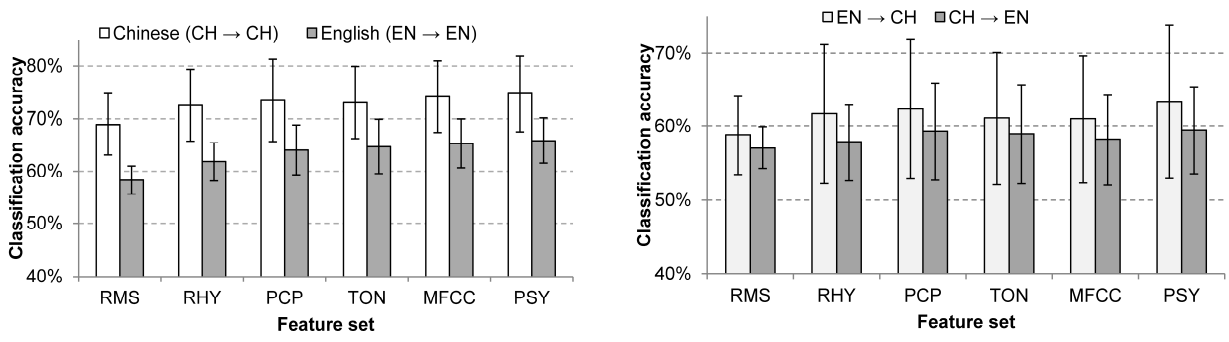
4.2 Cross-cultural Applicability of Classifiers

In this subsection, we report the result of using classifiers trained from English songs to classify Chinese songs, and vice versa. This set of experiments is designed to study the cross-cultural applicability of classification models which has rarely been addressed in the literature. Figure 4(b) shows the performances across the six feature sets.

PSY was again the best performing feature set, with average accuracies of 63.3% and 59.4% for Chinese and English test data, respectively. The performance differences between PSY and other feature sets are significant when Chinese songs are used as the test set (pair-wise t-test, $p < 0.001$). While the performances are comparable to those in the literature [3][17], they are significantly worse than those of last experiments where training and testing data were drawn from the same dataset (pair-wise t-test, $t = 5.92$ for Chinese songs; $t = 5.09$ for English songs, $df = 25$, $p < 0.001$). This means the datasets in the two cultures have significant difference. Whenever possible, it is better to use songs in the same culture to train classification model. However, in cases when training data from the same culture as test data are not available, it is still an acceptable alternative to use classification models built with data in the other culture.

Using English songs as training data to classify Chinese songs performed significantly better than the other way around (pair-wise t-test, $t = 3.70$, $df = 25$, $p = 0.001$). This may be because there were more English songs making larger training sets and/or mood classification of Chinese songs seemed to be easier (c.f. Section 4.1).

To further examine possible differences in acoustic characteristics between English and Chinese songs, we applied MDS to project the PSY features of the songs to a 2-D space (Figure 5). The fact that the two datasets overlap to a large extent indicates that Chinese songs and English songs in our datasets have similar perceptual timbre quality (as depicted by the PSY features [1]). This may partially explain the fact that PSY features performed well in both English and Chinese songs as well as the cross-cultural experiments.



(a) (b)

Figure 4. Average accuracy of different feature sets for mood classification of (a) intra-cultural classification using Chinese and English datasets and (b) inter-cultural classification using the other set as training data.

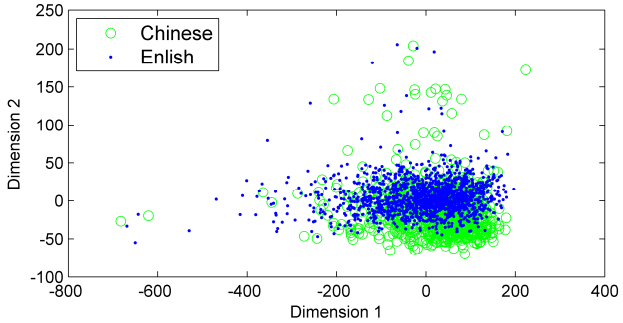


Figure 5. Visualizing PSY features of the two datasets.

5. CONCLUSIONS AND FUTURE WORK

This study investigates the cross-cultural applicability of mood categories, acoustic features and classification models in the case of English and Chinese songs. Results show that mood categories found in English songs are generally well applicable to Chinese songs except for several categories representing radical moods. It also seems feasible to apply feature descriptors developed for English songs to represent the audio content of Chinese songs, possibly due to the overlap of Psychoacoustic timbre features in both datasets. Our cross-cultural evaluation showed significant degradation of classification performance compared to the result of within-culture evaluation, although the absolute accuracy values are still comparable to the state-of-the-art in the literature.

In future work, we will examine the cross-cultural applicability of audio features and classification models on individual mood categories. We also plan to explore the problem of predicting the valence and arousal values of Chinese songs and investigate whether techniques that worked for Western music will work for Chinese music.

6. REFERENCES

[1] D. Cabrera: "Psysound: A computer program for psychoacoustical analysis," in *Proc. Australia Acoustic Society Conf.*, 1999.

[2] X. Hu and J. S. Dowine: "Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio," in *Proc. ACM/IEEE Joint Conf. Digital Libraries* (*JCDL*), pp. 159–168, 2010.

[3] X. Hu, J. S. Dowine, C. Laurier, M. Bay, and A. F. Ehmann: "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. ISMIR*, 462–467, 2008.

[4] Y. Hu, X. Chen, and D. Yang, "Lyric-based song emotion detection with affective lexicon and fuzzy clustering method," in *Proc. ISMIR*, 2009.

[5] P. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception," *J. Experimental Psychology*, vol. 16, no. 6, 2000.

[6] P. N. Juslin and P. Laukka, P.: "Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening," *J. New Music Research*, vol. 33, no. 3, pp. 217–238, 2004.

[7] Y. E. Kim *et al*: "Music emotion recognition: A state of the art review," in *Proc. ISMIR*, 2010.

[8] O. Lartillot and P. Toiviainen: "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *Proc. ISMIR*, pp. 127–130, 2007.

[9] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Proc. Int. Conf. Machine Learning and Applications*, pp. 1–6, 2008.

[10] L. Lu, D. Liu, and H. J. Zhang: "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006

[11] R. R. McCrae, P. T. Costa, and M. Yik: "Universal aspects of Chinese personality structure," in M. H. Bond (Ed.) *The Handbook of Chinese Psychology*, pp. 189–207. Hong Kong: Oxford University Press, 1996.

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

[13] E. M. Voorhees and D. K. Harman: *TREC: Experiments in Information Retrieval Evaluation*, MIT Press, 2005.

[14] W. Wu and L. Xie: "Discriminating mood taxonomy of Chinese traditional music and Western classical music with content feature sets," in *Proc. IEEE Congress on Image and Signal Processing*, pp. 148–152, 2008.

[15] Y. Xia, L. Wang, K. Wong, and M. Xu: "Sentiment vector space model for lyric-based song sentiment classification," in *Proc. ACL*, pp. 133–136, 2008.

[16] Y.-H. Yang and H.-H. Chen: "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2184–2196, 2011.

[17] Y.-H. Yang and H.-H. Chen: "Machine recognition of music emotion: A review," *ACM Trans. Intelligent Systems and Technology*, vol. 3, no. 3, 2012.