

A CROSS-CULTURAL STUDY OF MOOD IN K-POP SONGS

Xiao Hu

University of Hong Kong
xiaoxhu@hku.hk

Jin Ha Lee

University of Washington
jinhalee@uw.edu

Kahyun Choi

University of Illinois
{ckahyu2, jdownie}@illinois.edu

J. Stephen Downie

ABSTRACT

Prior research suggests that music mood is one of the most important criteria when people look for music—but the perception of mood may be subjective and can be influenced by many factors including the listeners' cultural background. In recent years, the number of studies of music mood perceptions by various cultural groups and of automated mood classification of music from different cultures has been increasing. However, there has yet to be a well-established testbed for evaluating cross-cultural tasks in Music Information Retrieval (MIR). Moreover, most existing datasets in MIR consist mainly of Western music and the cultural backgrounds of the annotators were mostly not taken into consideration or were limited to one cultural group. In this study, we built a collection of 1,892 K-pop (Korean Pop) songs with mood annotations collected from both Korean and American listeners, based on three different mood models. We analyze the differences and similarities between the mood judgments of the two listener groups, and propose potential MIR tasks that can be evaluated on this dataset.

1. INTRODUCTION

The mood of music is arguably one of the strongest factors behind people's motivation of listening to music [4]. Recognizing the importance of music mood, an increasing number of studies have been exploring the use of mood data to improve users' access to music. Recent studies in Music Information Retrieval (MIR) have indicated that people from different cultural backgrounds may perceive music mood differently ([2] [9]). In an effort toward establishing a global MIR system that can serve users from different parts of the world, researchers have developed and evaluated algorithms that can work on classifying music from different cultures and/or labeled by listeners from different countries ([17] [14]). Despite the growing interests on cultural influences on MIR ([7] [14]), we still do not have a well-established testbed for cross-cultural MIR tasks where methods proposed by interested researchers can be properly evaluated and compared. Music Information Retrieval Evaluation eXchange (MIREX), which is the primary MIR evaluation venue, has yet to add a cross-cultural evaluation task. This study aims to work toward filling this gap by 1) building a dataset¹ consisting of 1,892 songs from a non-

Western culture (i.e., K-pop or Korean Pop) and labels based on three music mood models annotated by listeners from two distinct cultural groups (i.e., American and Korean); 2) analyzing the differences and similarities between mood labels provided by American and Korean listeners on the same set of 1,892 K-pop songs; and 3) proposing cross-cultural MIR tasks that can be evaluated using this dataset.

2. RELATED WORK

Music is a medium beyond the boundary of languages, countries and cultures. As many MIR systems need to be designed to serve global users, researchers have been paying more attention to cross-cultural issues in MIR. Lee and Hu [9] compared mood labels on a set of 30 Western songs provided by American, Chinese, and Korean listeners and found that cultural background indeed influenced people's perception of music mood. Yang and Hu [17] compared mood labels on U.S. pop songs provided by Western listeners to labels on Chinese pop songs provided by Chinese listeners. The datasets were larger (nearly 500 songs) in their study, although the labels were applied to two separate datasets and thus may not be directly comparable. In this paper, we compare music mood perceptions of the same set of K-pop songs from American and Korean listeners, making the mood annotations directly comparable.

K-pop is increasingly becoming popular with international audiences, as evidenced by the launch of Billboard K-pop Hot 100 chart in 2011², and is actively sought by people from different cultural backgrounds. K-pop has unique characteristics due its history; Korean culture has been heavily influenced by American pop culture since the 1950s, yet is deeply rooted in the long history of East Asia. A recent study by Lee et al. [7] discussed the differences in the perception of K-pop genres by American and Korean listeners based on how they applied genre labels to K-pop music. In this study, we focus on the mood aspect of K-pop, aiming to improve the understanding of how mood can be used as a descriptor for organizing and accessing music by users from different cultures.

Currently there exist several influential datasets in music mood recognition ([3] [5]). However, most of them contain primarily Western music and the cultural background of annotators was either not specified [3] or not controlled [5]. To the best of our knowledge, the dataset built in this study is the first of its kind that is composed



© Xiao Hu, Jin Ha Lee, Kahyun Choi, J. Stephen Downie. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Xiao Hu, Jin Ha Lee, Kahyun Choi, J. Stephen Downie. "A Cross-Cultural Study of Mood in K-Pop Songs", 15th International Society for Music Information Retrieval Conference, 2014.

¹ The tag data will be incorporated into MIREX for use in a variety of MIR evaluations and released incrementally over time when not needed for MIREX.

² <http://www.billboard.com/articles/news/467764/billboard-k-pop-hot-100-launches-sistar-is-no-1-on-new-korea-chart>

of a significant amount of non-Western music, annotated by listeners from two distinct cultures, and labeled based on three music mood models. In MIREX, there have been two mood-related (sub)-tasks: Audio Mood Classification (AMC) starting from 2007 and the mood tag subtask in Audio Tag Classification (ATC) starting from 2009¹. Both tasks consist of Western songs labeled by listeners from unspecified cultural backgrounds [3]. This new dataset will enable evaluation tasks that explore the cross-cultural generalizability of automated music mood recognition systems [17].

3. STUDY DESIGN

3.1 The K-Pop Music Dataset

The dataset consists of 1,892 K-pop songs across seven dominant music genres in K-pop, namely Ballad, Dance/Electronic, Folk, Hip-hop/Rap, Rock, R&B/Soul, and Trot [7]. 30 second music clips were extracted from each song and presented to the listeners for mood annotation. This was to mitigate the cognitive load of annotators and to minimize the effect of possible mood changes during the entire duration of some songs (which can happen for some songs but is beyond the scope of this study).

3.2 Music Mood Models

In representing music mood, there are primarily two kinds of models: categorical and dimensional [5]. In categorical models, music mood is represented as a set of discrete mood categories (e.g., happy, sad, calm, angry, etc.) and each song is assigned to one or more categories. This study adopted two categorical models used in MIREX: 1) the five mood clusters (Table 1) used in the Audio Mood Classification task [3] where each song is labeled with one mood cluster exclusively; and 2) the 18 mood groups (Figure 2) used in the mood tag subtask in Audio Tag Classification where each song is labeled with up to six groups. Besides being used in MIREX, these two models were chosen due to the fact that they were developed from empirical data of user judgments and in a way that is completely independent from any dimensional models, and thus they can provide a contrast to the latter.

Unlike categorical models, dimensional models represent a “mood space” using a number of dimensions with continuous values. The most influential dimensional model in MIR is Russell’s 2-dimensional model [11], where the mood of each song is represented as a pair of numerical values indicating its degree in the *Valence* (i.e., level of pleasure) and *Arousal* (i.e., level of energy) dimensions. Both categorical and dimensional models have their advantages and disadvantages. The former uses natural language terms and thus is considered more intuitive for human users, whereas the latter can represent the degree of mood(s) a song may have (e.g., a little sad). Therefore, we used both kinds of models when annotating the mood of our K-pop song set. In addition to the 5 mood clusters and 18 mood groups, the K-pop songs were also annotated with the Valence-Arousal 2-dimensional model.

Cluster1 (C_1)	passionate, rousing, confident, boisterous, rowdy
Cluster2 (C_2)	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster3 (C_3)	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster4 (C_4)	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster5 (C_5)	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 1. Five mood clusters in the MIREX AMC task.

3.3 Annotation Process

For a cross-cultural comparison, a number of American and Korean listeners were recruited to annotate the mood of the songs. The American listeners were recruited via a well-known crowdsourcing platform, Amazon Mechanical Turk (MTurk), where workers complete tasks requiring human intelligence for a small fee. MTurk has been recognized as a quick and cost-effective way of collecting human opinions and has been used successfully in previous MIR studies (e.g., [6], [8]). In total, 134 listeners who identified themselves as American participated in the annotations based on the three mood models.

For the five-mood cluster model, each “HIT” (Human Intelligence Task, the name for a task in MTurk) contained 22 clips with two duplicates for a consistency check. Answers were only accepted if the annotations on the duplicate clips were the same. Participants were paid \$2.00 for successfully completing each HIT. For the 18-group model, we paid \$1.00 for each HIT, which contained 11 clips with one duplicate song for consistency check. There were fewer clips in each HIT of this model as the cognitive load was heavier: it asked for multiple (up to six) mood labels out of 18. For the Valence-Arousal (V-A) dimensional model we designed an interface with two slide scales in the range of [-10.0, 10.0] (Figure 1). We paid \$1.00 for each HIT, which contained 11 clips with one duplicate song for a consistency check. Consistency was defined such that the difference between the two annotations of the duplicate clips in either dimension should be smaller than 2.0. The threshold was based on the findings in [16] where a number of listeners gave V-A values to the same songs in two different occasions and the differences never exceeded 10% of the entire range. For each of the three mood representation models, three annotations were collected for each music clip. The total cost was approximately \$1800.

As there was no known crowdsourcing platform for Korean people, the nine Korean listeners who participated in the annotation were recruited through professional and personal networks of the authors. The annotation was done with our in-house annotation systems, which are similar to those in MTurk. All instructions and mood labels/dimensions were translated into Korean to minimize possible misunderstanding of the terminology. Similarly, each song received three annotations in each mood model. The payments to annotators were also comparable to those in MTurk. Although the total number of annotators in the two cultural groups differs, each song had exactly

¹ http://www.music-ir.org/mirex/wiki/MIREX_HOME

six independent annotations on which the following analysis and comparisons are based.

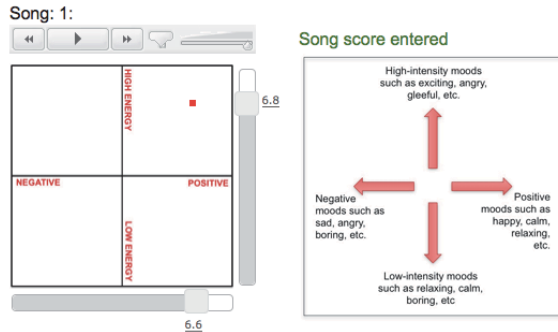


Figure 1. Annotation interface of the VA model (horizontal dimension is Valence, vertical is Arousal).

4. RESULTS

The annotations by American and Korean listeners are compared in terms of judgment distribution, agreement levels, and confusion between the two cultural groups. The Chi-square independence test is applied to estimate whether certain distributions were independent with listeners' cultural background.

4.1 Distribution of Mood Judgment

Table 2 shows the distribution of mood judgment of listeners from both cultural groups across five mood clusters. A Chi-square independence test indicates that the distribution does depend on cultural group ($p < 0.001$, $df = 4$, $\chi^2=396.90$). American listeners chose C_1 (*passionate*) and C_5 (*aggressive*) more often while Korean listeners chose C_2 (*cheerful*), C_3 (*bittersweet*) and C_4 (*silly/quirky*) more often. It is noteworthy that both groups chose C_3 (*bittersweet*) most often among all five clusters. This is different from [9] where both American and Korean listeners chose C_2 (*cheerful*) most often for American Pop songs. This difference may indicate that K-pop songs are generally more likely to express C_3 moods than American Pop songs.

	C_1	C_2	C_3	C_4	C_5	Total
American	1768	897	2225	311	475	5676
Korean	959	1321	2598	453	345	5676

Table 2. Judgment distributions across 5 mood clusters.

With the 18-mood group model, a listener may label a song with up to six mood groups. The American listeners chose 13,521 groups in total, resulting in an average of 2.38 groups per song. The Korean listeners chose 7,465 groups in total, which resulted in 1.32 groups per song. The fact that American listeners assigned almost twice as many groups to each song as Korean listeners did may be related to the individualism/collectivism dichotomy found in psychology and cultural studies [13]; Americans tend to be individualistic and are more flexible in accepting a range of ideas (mood groups in this case) than people from collectivistic cultures (often represented by East Asian cultures). Future studies employing more qualitative approaches are warranted to verify this speculation.

Figure 2 shows the distribution of judgments across the 18 mood groups. A chi-square test verified that the distribution is statistically significantly dependent on cultural backgrounds ($p < 0.001$, $df = 17$, $\chi^2=1664.49$). Americans used “gleeful”, “romantic”, “brooding”, “earnest”, “hopeful”, and “dreamy” more often than Koreans, while Koreans applied “sad” more frequently than Americans. Both groups used “angry” and “anxious” very rarely, probably due to the nature of K-pop songs. Similar observations were made in [17], where mood labels applied to Chinese and Western Pop songs were compared and radical moods such as “aggressive” and “anxious” were applied much more infrequently to Chinese songs than to Western songs. This may indicate a cultural difference in music: Chinese and Korean cultures tend to restrain and/or censor the expression of radical or destructive feelings whereas in Western cultures people are willing and free to expression of all kinds of feelings [10].

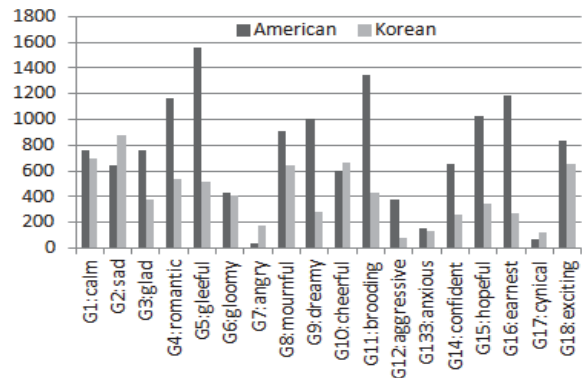


Figure 2. Judgment distributions across 18 mood groups (each group is represented by one representative term).

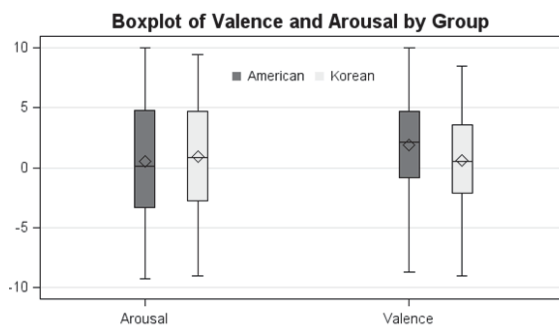


Figure 3. Boxplot of Valence and Arousal values.

Figure 3 shows the boxplot of the annotations based on the VA dimensional space given by the two groups of listeners. The V-A scores given by Americans are more scattered than those by Koreans, suggesting that Americans were more willing to choose extreme values. In addition, the means and medians indicate that Americans rated the songs with lower arousal values but higher valence values than Koreans ($p < 0.001$ in non-paired t -test for both cases). In other words, Americans tended to consider the songs to be less intense and more positive than did Koreans. This may also reflect the cultural difference

that individuals from Western cultures tend to experience and/or express more positive emotions than those from Eastern cultures [12], and Asians present themselves as less aroused compared to Americans and Europeans [1].

4.2 Agreements Within and Across Cultural Groups

In order to find out whether listeners from the same cultural background agree more with each other than with those from another cultural group, we examined the agreement among annotations provided by listeners in each cultural group as well as across cultural groups. The agreement measures used are the Sokal-Michener coefficient and intra-class correlation (ICC). The former is appropriate for categorical data while the latter is used for numerical data in the V-A space.

4.2.1 Sokal-Michener coefficient

The Sokal-Michener (S-M) coefficient is the ratio of the number of pairs with the same values and the total number of variables [2][9], and therefore a higher value indicates a higher agreement. For instance, if two listeners i and j had the same mood judgments on 189 of the 1892 songs, the S-M coefficient between them is approximately 0.1. Table 3 shows the average S-M coefficient aggregated across all pairs of annotators within and across cultural groups on the five-cluster annotations. It is not surprising that Koreans reached a higher agreement than Americans since they are annotating songs originating from their own culture. This is consistent with the findings in [2] and [9], where American listeners reached a higher agreement on the mood of American Pop songs than did Korean and Chinese listeners. The agreement level was the lowest when annotations from American and Korean listeners (cross-cultural) were paired up. The distribution of agreed vs. disagreed judgments is significantly dependent on whether the listeners are from the same cultural group or not, evidenced by the Chi-square test results (Table 3). Listeners from the same cultural group tend to agree more with each other than with those from a different culture.

	American	Korean	χ^2	df	P
American	0.47	0.43	25.35	1	<0.001
Korean	0.43	0.56	249.71	1	<0.001

Table 3. S-M coefficients of the five-cluster annotation within and across cultural groups

The analysis is more complex for the 18 group annotation, as each judgment can associate multiple labels with a song. To measure the agreement, we paired up labels applied to a song by any two annotators, and then calculated the S-M coefficient as the proportion of matched pairs among all pairs. For example, if annotator_1 labelled a song S with g1, g2, g3 and annotator_2 labelled it with g1, g4, then there were six annotation pairs and only one of them matched (i.e., g1 matched g1). The S-M coefficient in this case is $1/6 = 0.17$. Although the denominator increases when more labels are chosen, the chances they get matched also increase. All annotations from all listeners within each cultural group and across cultural groups were paired up in this way, and the result-

ant S-M coefficients are shown in Table 4. Again, the agreement level within Koreans was higher than that within Americans and also across cultural groups. However, the agreement within Americans was at the same level as the cross-cultural agreement, which is further evidenced by the statistically insignificant result of the Chi-square test.

	American	Korean	χ^2	df	p
American	0.11	0.11	3.72	1	0.054
Korean	0.11	0.15	156.88	1	<0.001

Table 4. S-M coefficient of the 18-group annotation within and across cultural groups

4.2.2 Intra-Class Correlation

The intra-class correlation (ICC) is a measure of agreement when ratings are given based on a continuous scale [15]. In the case of V-A annotation in this study, there is a different set of raters (listeners) for each item (song), and thus the one-way random model is used to calculate ICC within each group (3 raters) and across both groups (6 raters), for the valence and arousal dimensions. As shown in Table 5, cross-cultural agreement on valence is lower than within-cultural ones. Unlike five mood cluster annotation, both groups showed similar level of agreement on both dimensions. It is also noteworthy that the agreement on arousal annotation is much higher than valence annotation within- and cross-culturally. This is consistent with earlier MIR literature where valence has been recognized as more subjective than arousal [5].

	American	Korean	Cross-Cultural
Valence	0.27	0.28	0.23
Arousal	0.55	0.54	0.54

Table 5. ICC of Valence Arousal annotations within and across cultural groups

4.3 Confusion Between Cultural Groups

To further our understanding on the difference and similarity of mood perceptions between the two cultural groups, we also examined the disagreement between listeners in the two groups in each of the three types of annotations. For the 5-cluster annotation, Table 6 shows the confusion matrix of the 1,438 songs with agreed labels by at least two listeners in each cultural group. Each cell shows the number of songs labeled as one mood cluster by Koreans (column) and another by Americans (row). The cells on the (highlighted) diagonal are numbers of songs agreed by the two groups, while other cells represent the disagreement between the two groups. The matrix shows that both groups agreed more on C_3 (*bittersweet*) within themselves (661 and 842 songs respectively as shown by the "Total" cells). The bold numbers indicate major disagreements between the two groups. There are 268 songs Korean listeners judged as C_3 (*bittersweet*) that Americans judged as C_1 (*passionate*). The two groups only agreed on C_5 (*aggressive*) on 18 songs, whereas 49 songs judged as C_5 (*aggressive*) by Americans were judged by the Koreans as C_1 (*passionate*).

Table 7 shows the confusion matrix of the seven mood groups (due to space limit) with the most agreed songs by majority vote among the Korean listeners. The biggest confusion/discrepancy is between “exciting” and “gleeful”: 135 songs perceived as “gleeful” by Americans were perceived as “exciting” by Koreans. Other major confusions are between “exciting” and “cheerful”, and “sad” and “mournful.” These moods have similar semantics in terms of valence (both “sad” and “mournful” have low valence values) and arousal (both “exciting” and “gleeful” have high arousal values), which may explain the confusion between these terms. Similarly, there are few songs with disagreement between mood labels with very distinct semantics, such as “exciting” vs. “sad/calm/mournful”; “calm” vs. “cheerful/gleeful”; and “gleeful” vs. “mournful”.

AM \ KO	C_1	C_2	C_3	C_4	C_5	Total
C 1	70	79	268	18	22	457
C 2	41	126	10	11	2	190
C 3	19	53	558	22	9	661
C 4	10	6	5	22	1	44
C 5	49	10	1	8	18	86
Total	189	274	842	81	52	1438

Table 6. Cross-tabulation between 5-cluster annotations across cultural groups

It is interesting to see that a number of songs perceived as “romantic” by Americans were seen as “sad” (31 songs) and “calm” (30 songs) by Koreans. On the other hand, 18 songs perceived as “romantic” by Koreans were viewed as “calm” by Americans. “Romantic” was seldom confused with other high arousal moods such as “exciting” or “cheerful” by either Koreans or Americans, suggesting that both cultures tend to associate “romantic” with low arousal music.

AM \ KO	exciting	sad	cheerful	calm	mournful	gleeful	romantic	Total
exciting	71	2	35	2	2	28	3	143
sad	0	32	0	13	13	0	4	62
cheerful	35	3	32	1	3	7	2	83
calm	0	10	0	25	4	0	18	57
mournful	0	48	0	23	27	0	6	104
gleeful	135	4	98	2	2	55	4	300
romantic	4	31	3	30	18	3	27	116
total	245	130	168	96	69	93	64	865

Table 7. Cross-tabulation between 18-group annotations across cultural groups

For the 2-D annotation, we show the disagreement between the two groups in the four quadrants of the 2-D space (Table 8). Both groups agreed more with listeners from their own cultural group on the first quadrant (+A+V) and the third quadrant (-A-V) (as shown by the “Total” cells). The largest discrepancy was observed between -A+V and -A-V: 116 songs were perceived as

having negative arousal and positive valence (-A+V) by Americans but negative valence (-A-V) by Koreans. Similarly, for the songs perceived as having positive arousal by both groups, 118 of them were again perceived as having positive valence (+A+V) by Americans but negative valence (+A-V) by Koreans. This is consistent with our finding that Korean listeners are more likely to label negative moods than Americans (Section 4.1).

KO \ AM	+A+V	+A-V	-A+V	-A-V	Total
+A+V	495	118	25	34	672
+A-V	8	30	1	17	56
-A+V	51	24	84	116	275
-A-V	10	19	80	178	287
Total	565	191	190	346	1290

Table 8. Cross-tabulation among the four quadrants in 2-D annotations across cultural groups

5. DISCUSSIONS

5.1 Differences and Similarities Between Groups

The results show that mood judgments and the level of agreement are dependent on the cultural background of the listeners. A number of differences were found between the annotations of the two groups. First, Americans assigned a larger number of labels to each song, and applied more extreme valence and arousal values than Koreans (Figure 3). We speculate that perhaps this is related to the fact that the Western culture tends to encourage individualism and divergent thinking more than the Eastern culture [13]. The difference in the number of annotators is another possible explanation. Both of these factors will be further explored in future work. Second, compared to Americans, Koreans were more likely to label songs with negative moods such as “bittersweet”, “sad,” and “mournful” (Table 2, Figure 2), give lower valence values (Figure 3), and agree with each other more often on songs with negative valence (Table 9). These observations were consistent with and supported by findings in previous cultural studies that people from Western cultures tend to experience and/or express more positive emotions than those from Eastern cultures [12]. The fact that Americans in this study could not understand the lyrics of the songs may also have contributed to these results. Sometimes song lyrics and melody may express different moods to invoke complex emotions (e.g., dark humor). In particular, a recent trend among K-pop artists to use faster tempo in Ballad songs may make the melody sound positive or neutral, although the lyrics are sad or melancholy as is the convention for Ballad songs.

It is also found that agreements of within-cultural groups are higher than that of cross-cultural groups based on the comparison of S-M coefficient, and ICC values (on valence only). For within-cultural group agreement, Koreans reached a higher agreement than Americans on 5-cluster annotation, which may be explained by the fact that Koreans were more familiar with the K-pop songs used in this study than Americans. Prior familiarity with

songs was also identified as a factor affecting the agreement level of mood perception in previous studies [2].

Some similarities were also found between the annotations of the two groups: 1) both groups applied and agreed on C_3 (*bittersweet*) more often than other mood clusters (Tables 2 and 8); 2) both groups seldom applied radical mood labels such as “aggressive”, “angry”, “anxious” (Table 2 and Figure 2); and 3) both groups agreed more on songs with +A+V and -A-V values (Table 9). These similarities can potentially be attributed to the nature of the K-pop songs. A previous study comparing mood labels on Western and Chinese Pop songs also found that there were significantly fewer radical mood labels assigned to Chinese Pop songs than to Western songs [17]. This may reflect Eastern Asian preferences for non-aggressive music, perhaps due to their tradition of being more conservative and limiting the expression of feelings [10]. Another likely explanation would be the censorship and regulation¹ that still heavily affects the popular music culture in countries like South Korea and China.

5.2 Proposed MIR Evaluation Tasks

One of the main contributions of this study is to build a large cross-cultural dataset for MIR research. The unique characteristics of the dataset built for this study make it suitable for various evaluation tasks involving cross-cultural components. Specifically, for each of the three annotation sets (i.e., 5-clusters, 18-groups, and 2-dimensions), both within- and cross-cultural evaluations can be performed. For the former, both training and test data can be extracted from the datasets with annotations by listeners from the same cultural group (by cross-validation, for example); for the latter, models can be trained by the dataset annotated by listeners in one culture and applied to the dataset annotated by listeners in another culture. These tasks will be able to evaluate whether mood recognition models often used in Western music can be equally applied to 1) non-Western music, specifically K-Pop songs; 2) K-Pop songs annotated by American and/or Korean listeners; and 3) cross-cultural music mood recognition, for both categorical mood classification [17] and dimensional mood regression [5].

6. CONCLUSIONS AND FUTURE WORK

This study analyzed music mood annotations on a large set of K-Pop songs provided by listeners from two distinct cultural groups, Americans and Koreans, using three mood annotation models. By comparing annotations from the two cultural groups, differences and similarities were identified and discussed. The unique characteristics of the dataset built in this study will allow it to be used in future MIR evaluation tasks with an emphasis on cross-cultural applicability of mood recognition algorithms and systems. Future work will include detailed and qualitative investigation on the reasons behind the differences between mood judgments of these two user groups as well as listeners from other cultural groups.

¹ <http://freemuse.org/archives/7294>

7. ACKNOWLEDGEMENT

Funding support: Korean Ministry of Trade, Industry and Energy (Grant #100372) & the A.W. Mellon Foundation.

8. REFERENCES

- [1] E. Y. Bann and J. J. Bryson: “Measuring cultural relativity of emotional valence and arousal using semantic clustering and twitter,” *Proc. of the Annual Conf. of the Cognitive Sci. Soc.*, pp.1809-1814, 2013.
- [2] X. Hu & J. H. Lee: “A cross-cultural study of music mood perception between American and Chinese listeners,” *ISMIR*, pp.535-540. 2012.
- [3] X. Hu, J. S. Downie, C. Laurier, M., Bay and A. Ehmann: “The 2007 MIREX audio mood classification task: Lessons learned,” *ISMIR*, 2008.
- [4] 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.
- [5] 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,” *ISMIR*, pp. 255–266, 2010.
- [6] J. H. Lee: “Crowdsourcing music similarity judgments using Mechanical Turk,” *ISMIR*, 2010.
- [7] J. H. Lee, K. Choi, X. Hu and J. S. Downie: “K-Pop genres: A cross-cultural exploration,” *ISMIR*, 2013.
- [8] J. H. Lee and X. Hu: “Generating ground truth for mood classification in Music Digital Libraries Using Mechanical Turk,” *IEEE-ACM JCDL*, 2012.
- [9] J. H. Lee and X. Hu: “Cross-cultural similarities and differences in music mood perception,” *Proceedings of the iConference*, 2014.
- [10] 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.
- [11] J. A. Russell: “A circumscript model of affect,” *Journal of Psychology and Soc. Psychology*, vol. 39, no. 6, 1980
- [12] Y. Miyamoto and X. Ma: “Dampening or savoring positive emotions: A dialectical cultural script guides emotion regulation.” *Emotion* 11.6, pp.1346-347, 2011.
- [13] J. J. Schmidt, T. Facca and J. C. Soper: “International variations in divergent creativity and the impact on teaching entrepreneurship,” *Journal of Higher Ed. Theory & Practice*. Vol. 13 Issue 2, pp. 101-109. 2013.
- [14] X. Serra: “A multicultural approach in music information research,” *ISMIR*, 2011.
- [15] P. E. Shrout and J. L. Fleiss: “Intraclass correlations: Uses in assessing rater reliability,” *Psychological Bulletin*, Vol. 86, pp. 420-3428, 1979.
- [16] J. A. Speck, E. M. Schmidt, B. G. Morton and Y. E. Kim: “A comparative study of collaborative vs. traditional annotation methods,” *ISMIR*, 2011.
- [17] Y.-H. Yang and X. Hu: “Cross-cultural music mood classification: A comparison on English and Chinese songs,” *ISMIR*, 2012.