Cross-Dataset and Cross-Cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs

Xiao Hu and Yi-Hsuan Yang, *Member, IEEE*

Abstract—In music mood prediction, regression models are built to predict values on several mood-representing dimensions such as valence (level of pleasure) and arousal (level of energy). Many studies have shown that music mood is generally predictable based on music acoustic features, but these experiments were mostly conducted on datasets with homogeneous music. Little research has been done to explore the generalizability of mood regression models cross datasets, especially those with music in different cultures. In the increasingly global market of music listening, generalizable models are highly desirable for automated processing, searching and managing music collections with heterogeneous characteristics. In this study, we evaluated mood regression models built on fifteen acoustic features in five mood-related musical aspects, with a focus on cross-dataset generalizability. Specifically, three distinct datasets were involved in a series of five experiments to examine the effects of dataset size, reliability of annotations and cultural backgrounds of music and annotators on mood regression performances and model generalizability. The results reveal that the size of the training dataset and the annotation reliability of the testing dataset affect mood regression performances. When both factors are controlled, regression models are generalizable between datasets sharing a common cultural background of music or annotators.

Index Terms—Cross-dataset, cross-cultural, evaluation, generalizability, music mood prediction, regression model

1 Introduction

The availability of large amounts of digital music calls for automated means to facilitate music information organization and access. The affective aspect of music, often referred to as music mood or emotion[1], has been recognized as an important criterion when people manage or seek for music [1], [2]. Consequently, techniques for automated music mood classification and regression have been developed and achieved substantial results [3], [4]. Major evaluation events in music information retrieval (MIR) also include mood-related tasks [5], [6].

With music increasingly consumed by a global audience, cultural factors have attracted much attention in MIR [7], [8], [9], [10]. In particular, it is found that cultural context, in addition to musical features, affects how people feel about music mood [2]. Studies have shown that people from different cultural backgrounds may perceive the mood of the same music in significantly different ways [8], [11], [12], [13], [14].

In experimenting with cross-cultural generalizability of mood prediction techniques, datasets of different cultural backgrounds are involved. However, the datasets often differ in other factors in addition to cultural backgrounds, such as size and annotation reliability level. Therefore, a thorough discussion of this topic needs to include both *cross-dataset* and *cross-cultural* aspects [15], [16]. This study significantly extends previous studies by systematically evaluating 15 acoustic features in five musical aspects, experimenting with three music datasets with unique cultural backgrounds, as well as conducting a series of experiments with varied configurations on data characteristics. The goal of this study is to further our understanding on cross-dataset and cross-cultural generalizability of music mood regression models.

As will be explained in detail in the next section, music mood can be represented by a set of categories or by numerical values in low-dimensional mood spaces. For the former, the automated technique used to assign music pieces to mood categories is *classification*. For the latter, the technique is *regression* that automatically predicts values in the mood dimensions for music pieces. Both classification and regression are prediction problems in machine learning. In mood regression, the regression models are usually built on acoustic features extracted from the audio files of the music pieces. Existing studies have attempted to evaluate the effectiveness of different features in mood regression [3], [4], but there have been few studies on cross-dataset and cross-cultural generalizability of mood regression models built on various feature sets. To fill the gaps in previous research, this study aims to investigate the following two research questions:

RQ1: *Which music acoustic features are most effective for music mood regression, within- and cross-datasets?*

1. Following the convention in Music Information Retrieval, we use the terms *mood* and *emotion* interchangeably in this study, although they bear different meanings and implications in psychology.

- *X. Hu is with the University of Hong Kong, Hong Kong. E-mail: xiaoxhu@hku.hk.*
- *Y.-H. Yang is with the Research Center for Information Technology and Innovation, Academia Sinica, Taipei, Taiwan. E-mail: yang@citi.sinica.edu.tw.*

Manuscript received 3 June 2015; revised 14 Dec. 2015; accepted 23 Jan. 2016. Date of publication 28 Jan. 2016; date of current version 6 June 2017. Recommended for acceptance by A. Potamianos. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2016.2523503

An extensive comparison is conducted on the most commonly used feature sets as well as those recently proposed. Some of them (e.g., chroma and tempogram features) have not been systematically evaluated in mood regression tasks. Even though some feature sets in the loudness, timbre and harmony categories have been widely used in automated MIR tasks, different studies extracted the feature sets using different toolboxes and the feature sets were rarely compared side by side or in a cross-dataset scenario. Answers to this research question will be instructive for MIR systems intended to work with different kinds of music and serve listeners in various situations and backgrounds.

RQ2: Can mood regression models trained with music in one dataset be applied to music in another? How do factors of dataset size, annotation reliability level and cultural background affect the cross-dataset applicability of the models?

With three distinct yet comparable datasets, this study examines the effects of size, annotation reliability level and cultural backgrounds (of both music and annotators) of the datasets by constructing sub-datasets and conducting a series of regression experiments in various controlled conditions. Answers to this question will contribute to cross-dataset and cross-cultural MIR, in providing empirical evidence on the extent to which music mood regression models could transcend various boundaries between different collections of music.

The rest of the paper is organized as follows. After discussing related work, we will describe the three datasets with distinct characteristics and the five types of acoustic features evaluated in this study. Experiments designed to answer the two research questions will then be presented, followed by detailed discussions on the results. We then draw conclusions by summarizing findings of this study and envisioning future research.

2 Related Work

2.1 Music Mood Representation and Prediction

Music mood is often represented in categorical and dimensional models in music psychology and MIR. Categorical models use a finite set of discrete labels (e.g., sad, angry) to categorize music mood (e.g., [17]), whereas the dimensional models use continuous values in a low-dimensional space to capture the mood of music (e.g., [18]). In a dimensional model, each dimension is a psychological factor of mood. The most widely used dimensions in music mood analysis are valence (i.e., level of pleasure) and arousal (i.e., level of energy) [1], [3], [4], [9], as presented in the well-known Russel's model [18]. When music mood is represented in categorical models, the technique used to predict a mood label is classification, which has been explored in many studies [3], [4], [5]. The Music Information Retrieval Evaluation eXchange (MIREX) has set up an "Audio Mood Classification" task [5] and an "Audio Tag Classification Task – Mood subtask" since 2007 and 2009, respectively [19]. We refer to the prediction of mood labels as *music mood classification* in this paper.

There are also many studies using dimensional representation of music mood, particularly using the valence and arousal dimensions [3], [20], [21], [22], [23]. In these studies, regression models are built to predict numerical values in the valence and arousal dimensions for each music piece. Starting from 2013, the MediaEval Benchmarking Initiative for Multimedia Evaluation has included music mood regression tasks [6]. We refer to the prediction of numerical mood values as *music mood regression* in this paper. Moreover, we use *music mood prediction* as a more general term that includes both the classification and regression subtasks.

2.2 Cross-Cultural and Cross-Dataset Music Mood Prediction

As most existing studies on computational modeling of music mood have focused on Western music [1], [3], [4], [6], [19], researchers now are interested in finding out whether and to what extent techniques designed for Western music can be applied to non-Western music [7], [9], [15], [16]. In music mood classification, the study by [16] compared mood categories and mood classification models for English and Chinese Pop songs. Classification models were trained with songs in one dataset and tested with those in the other. The result showed that the classification models were generally applicable cross-dataset even though within-dataset classification still performed better [16]. In the most recent round of MIREX, a cross-cultural mood classification task was initiated with a Korean Pop (K-POP) music dataset annotated by both American listeners and Korean listeners [9]. It aims to investigate whether mood classification models developed on Western music can be applied to K-POP music and whether classification models can be effective on classifying both mood labels given by American annotators and those given by Korean annotators. The task attracted 10 participating systems and the results are currently under analysis. One main difference between this current study and [9], [16] lies in that of mood regression versus mood classification.

There have been few studies on cross-dataset music mood regression. Eerola [24] explored the generalizability of mood regression models across nine datasets in different genres including classical, film, pop and mixed genres. It was found that arousal was moderately generalizable across genres but valence was not. Although Eerola exhaustively evaluated nine datasets, all of them were composed of Western music. In addition, the different sizes and annotation reliability levels of the datasets were not considered as factors that might have affected model generalizability. A recent study by [15] explored the generalizability of mood regression models in a cross-cultural and cross-dataset setting, and found that prediction on the arousal dimension was generally feasible across datasets, whereas cross-dataset prediction on the valence dimension only worked when the music or annotators of the datasets shared the same cultural background. Notwithstanding its important and encouraging findings as a first study in cross-cultural music mood regression, [15] only considered one set of acoustic features in each of the mood-related musical aspects (e.g., loudness, rhythm, harmony). Given the fact that each aspect of music characteristics can be more or less captured by different sets of acoustic features, this current study compares multiple acoustic feature sets within each of the musical aspects related to music mood. Furthermore, an automated feature selection mechanism is applied to search for the most effective combination of multiple feature sets. In addition, due to the scarcity of comparable datasets

with mood annotations, one of the datasets in [15] consisted of the audio part of music video clips for which the mood annotations were based on both audio and moving image components. To eliminate possible effects introduced by non-music stimuli, this study only uses datasets of music audio, with comparable instructions of annotations. Two of the datasets are newly built, of significant sizes, and have not been used in cross-dataset and cross-cultural music mood regression before. Last but not least, to investigate possible factors that might affect regression model generalizability, a series of experiments are carried out in this study with varied data sizes and annotation reliability levels.

In summary, this current study fills the gaps in existing literature by 1) evaluating cross-dataset and cross cultural generalizability of music mood regression models; 2) comparing recently proposed acoustic features; 3) using new and comparable datasets suitable for the task; and 4) exploring the effects of data size and annotation reliability level on model generalizability.

3 The Datasets

3.1 MER60

Developed by Yang and Chen [25] for mood regression tasks, this dataset consists of 60 English Pop songs. A 30-second clip was manually selected from the chorus parts of each song, and was annotated by 40 university students (in non-music major). The annotators were born and raised in Taiwan and thus had a Chinese cultural background. The annotators were instructed (in Chinese) to give real values ranging between $[-5, 5]$ to the valence and arousal dimensions by clicking on a two-dimensional mood space displayed on a computer screen. The groundtruth values were the average across all annotators after outliers were removed. The inter-annotator reliability of the annotations was measured by Krippendorff's α [26], and the results were 0.387 and 0.704 for valence and arousal respectively, indicating "fair" and "moderate" agreement. It is well acknowledged that valence annotations are harder to reach agreement than arousal annotations [3], [21].

3.2 CH818

Newly built for this study, this dataset contains 818 clips of Chinese Pop songs released in Taiwan, Hong Kong and Mainland China. Each of the clips is 30-second long and was algorithmically extracted from the segment with the strongest affective content [16]. Specifically, the algorithm applied 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 ($|\text{valence}|^2 + |\text{arousal}|^2$) value was chosen to represent each song. Each clip was annotated by three music experts who were born and raised in Mainland China and thus were with a Chinese cultural background. The annotation was done with an interface consisting of two sliding bars of continuous real values between $[-10, 10]$ for valence and arousal respectively. The instructions were written in Chinese and a training session was conducted to ensure a precise understanding of the annotation task. On average each annotator spent 25.67 hours in annotating these clips. Krippendorff's α of valence and arousal annotations were 0.491 ("fair") and 0.617 ("moderate"). It is acknowledged that the number of annotators was smaller than the other two datasets, but the annotations were considerably consistent. The Pearson's correlation coefficient between each annotator's annotations and the averaged annotations is quite close among the three annotators ($r = 0.842 \sim 0.907$ for arousal and $0.794 \sim 0.833$ for valence) [27]. Therefore, similar to MER60, the average values across the three annotations were used as the groundtruth. The annotations and extracted audio features of the CH818 dataset will be publicly available for research purposes.

TABLE 1
Characteristics of the Three Dataset

		MER60	CH818	AMG1608
Music	Format	Audio (mp3)	Audio (mp3)	Audio (mp3)
	Size	60	818	1608
	Culture	Western	Chinese	Western
	Length	30 seconds	30 seconds	1 minute
	Segment selection	Chorus part; manual selected	Segment with the strongest emotion	Audio previews from 7 digital
Annotators	Type	Volunteers	Experts	MTurk workers
	Culture	Chinese	Chinese	Western
	Number	40 per clip	3 per clip	15–32 per clip
Annotations	Scale	Continuous $[-5, 5]$	Continuous $[-10, 10]$	Continuous $[-1, 1]$
	Dimensions	V. A.	V. A.	V. A.
	Interface	2-D interactive interface	two separate sliding bars	2-D interactive interface
	Emotion	Intended	Intended	Intended
	α	V: 0.387; A: 0.704	V: 0.491; A: 0.617	V: 0.306; A: 0.458

Acronyms: V: valence; A: arousal.

3.3 AMG1608

This dataset was built by [28], consisting of 1,608 preview clips (each 30-second long) of Western songs available on 7digital, a popular music stream service. The valence and arousal emotion annotations were collected using Amazon Mechanic Turk (MTurk), a crowdsourcing platform. Workers in MTurk received instructions similar to those in the other two datasets. As only workers who were American people and lived in the U.S. were invited, the annotators of this dataset had a Western cultural background. Similar to MER60, a two-dimensional space of continuous real values between [–1, 1] was provided for annotators to indicate the valence and arousal values at the same time. A validation question was carefully designed to ensure the quality of the annotations. Specifically, one in every 10 clips was duplicated and a worker's annotations were only accepted if the annotations of the duplicated clips were within 0.2 (10 percent of the given value range). Each clip was annotated by 15-32 annotators and their averages were used as the groundtruth. The Krippendorff's α was 0.306 for valence and 0.458 for arousal, both in the range of "fair" agreement.

The three datasets are suitable for this study because they were annotated with the same scheme (i.e., continuous values in valence and arousal dimensions) and each of them has a homogeneous cultural background in terms of either music or annotators. Table 1 summarizes the characteristics

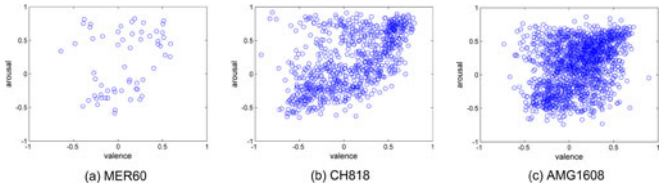

Fig. 1. Distribution of valence and arousal values of the three datasets. (All values were normalized to [−1, 1].)

of the three datasets, in terms of music, annotators and annotations. There are commonalities in the distributions of valence and arousal annotations in the three datasets (Fig. 1): more songs are in the first quadrant and few songs are in the right bottom corner in the valence arousal (VA) space (near (1, –1)). In addition, the left bottom corner of the space (near (−1, −1)) is rather unoccupied in MER60 and AMG1608, which reflects the finding in existing studies that it is unlikely for someone to feel unpleasant and tranquil at the same time by an acoustic stimulus [29], [30]. There are mainly three reasons to use datasets of English and Chinese Pop songs: first, there are large audiences for English and Chinese Pop songs; second, they may represent differences between Western and Eastern cultures; third, datasets of music in these two cultures are more available to us than those in other cultures.

4 Music Acoustic Features

Studies in music psychology have discovered that music mood is associated with multiple musical traits [31]. Therefore, we extract a variety of acoustic features based on which mood regression models are built. The features cover five aspects of musical characteristics including loudness, pitch, rhythm, timbre and harmony which are recognized as having consistent relationships with perceived mood [32], [33]. These features and the tools used to extract them from music audio signals are described in this section and summarized in Table 2. We used the default parameter settings of the tools in feature extraction, as they have been shown effective in many music signal processing tasks [34], [35], [36], [37]. Before feature extraction, an audio signal was re-sampled at 22 kHz. Different tools used short time windows of different lengths and frame rates in computing the short-time Fourier transform. Unless otherwise specified, all the features are computed frame-by-frame and then aggregated across time into clip-level features by taking the mean and standard deviation values.

4.1 Loudness

Loudness is the perceptual intensity of sound. It depends primarily on the physical intensity of sound, but is also related to other physical properties of sound, such as frequency and duration. In this study, we adopt two sets of loudness features.

Loudness_mirtb (l_m) contains energy features extracted by MIR Toolbox, a widely used tool in MIR [34]. This feature set contains the mean and standard deviation of the frame-level root mean square values of the audio signal voltage and waveforms across all the frames of a music piece, with 50ms frame size and 50 percent overlap between successive frames.

TABLE 2
Summary of Acoustic Features Used in This Study

Category	Code	Feature	Dim.	Toolbox
Loud-ness	l_m	RMS Energy	2	M
	l_p	Loudness	18	P
Pitch	p	Pitches	88	C
	pCP	Chroma-pitch	24	C
	pCLP	Chroma-log-pitch	24	C
	pCE	Chroma energy normalized statistics	24	C
	pCR	ChromaDCT-reduced log pitch	24	C
Rhythm	rCTD	Fourier-based cyclic tempogram	80	T
	rCTA	ACF-based cyclic tempogram	80	T
	r_m	rhythm strength, regularity, clarity, avg. onset frequency, avg. tempo	5	M
Timbre	t_mf	Mel-frequency cepstral coefficients	120	M
	t_mi	Spectrum characteristics	28	M
	t_p	Dissonance	8	P
Harmony	h_m	Key clarity, mode, and HCDF	6	M
	h_p	Tonalness, multiplicity, and chord change likelihood	8	P

Acronyms: "Dim" stands for Dimensionality, C = Chroma Toolbox M = MIR Toolbox, P = PsySound, T = Tempogram Toolbox, l_m = loudness_mirtb, l_p = loudness_psysound, p = pitch, pCP = pitchCP, pCLP = pitchCLP, pCE = pitchCENS, pCR = pitchCRP, rCTD = rhythmCTDFT, rCTA = rhythmCTACF, r_m = rhythm_mirtb, t_mf = timbre _mfcc, t_mi = timbre_mirtb, t_p = timbre_psysound, h_m = harmony_mirtb, h_p = harmony_psysound, ACF = autocorrelation function.

Loudness_psysound (l_p) contains features extracted by the PsySound toolbox, a unique tool extracting not only the physical measurement of music audio signals (e.g., voltage and amplitude) but also human perceptions of sound [35]. This loudness feature set captures the sensation of loudness in the human auditory system by modeling the outer and middle ear transfer function and the response of the basilar membrane within the cochlea of the inner ear. Specifically, five types of loudness features are computed: the total loudness, the sharpness of sound based on two psychoacoustic models, the timbral width, and the spaciousness (volume) of sound based on another psychoacoustic model. Also included are sound pressure levels with and without weighting, as well as those with 'fast' and 'slow' integration times using exponential temporal integration. The PsySound features are computed for every 4,096 samples, without overlaps.

4.2 Pitch

Pitch is the auditory attribute of sound that can be ordered on a scale from low to high. Western music has 88 different pitches, usually notated as A0 to C8. The relationship between two or more simultaneous pitches constitutes harmony of music, whereas music melody is related to the temporal succession of pitches [36]. In this study, we evaluate a range of pitch feature sets extracted by Chroma toolbox, a recently developed tool focusing on pitch-based audio features [36].

Pitch (p). This set of features represent the 88 absolute pitches using the energy level of frequency bands corresponding to the pitches within the audio signal. Specifically, the short-time mean-square power of each pitch subband is calculated and summed over time, leading to an 88-bin histogram corresponding to the 88 notes.

Pitch_CP (pCP). In Western music, pitches are usually represented by twelve pitch classes (C, C#, . . . B). Each pitch class encompasses all pitches that are octaves apart (e.g., C4 and C5) and is referred to as a chroma. This Chroma-Pitch

(CP) feature set aggregates the short-time energy of the pitches in each chroma and represents each chroma by the mean and standard deviation of its energy levels over time.

Pitch_CLP (pCLP). Chroma-Log-Pitch feature set is similar to the last one except that it applies a logarithmic transformation to the pitch features, so as to take into account the logarithmic sound sensation detected by the human auditory system.

Pitch_CENS (pCE). Chroma energy normalized statistics features are short-time statistics of energy distributions within each chroma. This set of features is strongly correlated to the short-time harmonic content of music and is claimed to be robust to variations of other musical properties such as dynamics, timbre and articulation.

Pitch_CPR (pCR). Chroma DCT-Reduced log Pitch features further improve the timbre invariance of the chroma features by removing part of the cepstral coefficients closely related to the timbral aspect of music.

4.3 Rhythm

Rhythm is an important musical trait that depicts tempo or pulse of a music piece. The recently developed Tempogram Toolbox [37] is adopted in this study for extracting rhythmic features, because it implements the novel concept of cyclic tempograms that is inspired by the concept of chroma features. This approach forms tempo equivalence classes by grouping tempi that differ by a power of two. The resultant cyclic tempo features are robust to weak note onset and changing tempo [38]. In this study, we consider the following two sets of cyclic tempogram features.

Rhythm_CTDFT (rCTD). Cyclic tempograms based on a short-time discrete Fourier transformation. According to [38], this feature set responds to harmonies (multiples) of the beats-per-minutes (BPM) measures while suppressing subharmonies (fractions) of them. We divide the tempo range from 30 to 480 BPM to 40 bins and take the mean and standard deviation values across time.

Rhythm_CTACF (rCTA). Cyclic tempograms based on the autocorrelation function. Unlike Rhythm_CTDFT, this feature set indicates subharmonics while suppressing harmonics [38].

Rhythm_mirtb (r_b). For comparison purposes, we also include the conventional rhythm features extracted from the MIR Toolbox which have been shown to be relevant to both valence and arousal perception [39]. It consists of five features: rhythm strength (the average onset strength of the onset detection curve), rhythm regularity and clarity (calculated based on the autocorrelation on the onset detection curve), average onset frequency (the number of note onsets per second), and the estimated average tempo.

4.4 Timbre

Timbre often refers to the quality or texture of sound that distinguishes a musical sound from another, even when the sounds share the same pitch and loudness. Timbre features have been widely used in music genre and mood classifications [3], [4], [40]. Certain properties of the music audio signals can be used to model the timbre aspect of music such as spectrum, intensity and roughness. In this study, we use both MIR Toolbox and PsySound to extract different sets of timbre features.

Timbre_mfcc (t_mf). The Mel-frequency cepstral coefficients (MFCC) have been widely used and highly effective in many MIR tasks, due to its use of the nonlinear Mel-frequency scale that approximates the human auditory system's responses [41]. We compute the first 20 MFCCs, excluding the DC term, along with its first-order and second-order instantaneous derivatives (i.e., the Δ and $\Delta\Delta$ of the features) for each 50-millisecond frame.

Timbre_mirtb (t_mi). Timbre features extracted from the MIR Toolbox capture characteristics of the spectrum transformed from the audio signals of a music piece, including statistic measures of the spectrum: centroid, skewness, kurtosis; distribution of sound energy: brightness, flatness, roll-off; and changes of the spectrum: irregularity, zero crossing rate, and spectral flux.

Timbre_psysound (t_p). The dissonance features extracted from PsySound measuring the "harshness" or "roughness" of the sound. In the context of tonal music, a chord or a note is said to be consonant when it implies stability, and dissonant when it implies instability. Dissonance is often used in combination with consonance to increase the expressiveness of music. This feature set includes spectral and dissonance features generated by four psychoacoustic models.

4.5 Harmony

Harmony refers to the relationship between two or more simultaneous pitches in music. In this study, we use both MIR toolbox and PsySound to extract different sets of harmony features.

Harmony_tonal (h_m). Extracted from the 12-bin chroma vectors by the MIR Toolbox, this feature set contains key clarity, musical mode, and harmonic change detection function (HCDF). Key clarity indicates how clearly the set of pitches in a frame is organized into a harmonic structure. It is calculated by comparing the chroma vector of a short-time frame to the pre-computed profiles of the 24 major and minor keys. The similarity (or key strength) between the chroma vector and the best matched key profile is returned as key clarity. Musical mode is the key strength difference between the best major and minor keys. The major-minor tonality is the predominant musical system in Western classical music as well as most modern popular music. Finally, the HCDF indicates the magnitude of difference in harmonic content between successive frames.

Harmony_psysound (h_p). From the chroma vectors, PsySound calculates the tonalness (the extent to which the sound follows pitch progressions in tonal music) and multiplicity (the number of pitches heard) of the music signals [35]. Moreover, PsySound uses a correlation based algorithm to perform chord recognition and calculates the likelihood of chord changes, which can be considered as a HCDF-like measure of harmonic variation.

5 Experiment Design

To answer the first research question, we conducted regression experiments with each of the aforementioned feature sets on the three given datasets, in within- and cross-dataset scenarios. Existing studies suggest that combinations of multiple feature sets may improve regression performances as different feature sets may capture different aspects of information and thus could possibly compensate for one another [5], [42].

TABLE 3
Krippendorf's α of the Datasets Before and After Selection in Experiments 3 and 4

	MER60	CH818	AMG1608
Complete set	V: 0.387; A: 0.704	V: 0.491; A: 0.617	V: 0.306; A: 0.458
Selected subset	V: 0.387 A: 0.704	V: 0.458 A: 0.629	V: 0.403 A: 0.627

Therefore, we conducted systematic feature selection using the step-wise forward feature selection algorithm [43] to find out a best performing combination of multiple feature sets. Specifically, the algorithm started from the best-performing single feature set (e.g., loudness_psysound). Then for each round, the algorithm selected another feature set which achieved best performance when combined with those already selected. The algorithm stopped when there was no performance improvement by adding new feature set. As the goal of feature selection is to identify one feature combination that is good for both within and cross dataset regression, the performance measure used here is the average performance across all nine dataset combinations (i.e., the last row in Tables 5 and 7, indicated with "ALL"). It is a greedy approach that finds a local maximum approximating a global optimal solution in a reasonable time. Cross-validation was used for each performance evaluation in the process.

To answer the second research question, as the datasets vary in different aspects, we conducted a series of experiments with controlled conditions and compared the performances on within- and cross-dataset predictions (Experiments 1 to 5). Experiment 1 used the original datasets, but with the selected combined feature set. To investigate the effect of dataset sizes, in Experiment 2 we built regression models with the same size of training datasets by randomly selecting 60 examples from the AMG1608 and CH818 datasets (as the MER60 dataset has 60 samples). The random selections were repeated 20 times and the averaged regression performances are presented.

Another factor that might affect regression performances is the annotation reliability level of the datasets. Presumably clips with higher agreement among human annotators would likely be clearer in mood expression, and therefore could be easier for automated regression. In Experiments 3 and 4, we constructed subsets of 60 clips from the CH818 and AMG1608 datasets with controlled reliability levels that approximate those of the MER60 dataset (see Section 3). For AMG1608, as it has lower reliabilities on both valence and arousal compared to MER60, we sampled 60 clips that had higher agreement levels and were equally distributed in the four quadrants of the VA space. Specifically, we calculated the sum of variances in valence and arousal annotations for each of the clips, and obtained 100 samples with the smallest sum of variances in each quadrant, followed by a random sampling of 15 clips from each quadrant. We repeated the sampling 10 times and chose the one with the reliability values closest to those of MER60. For CH818, as its valence reliability is higher than that of MER whereas arousal reliability is lower, we ranked the clips by the difference between average distance of arousal annotations and that of valence annotations, so as to select clips with smaller differences on arousal annotations and larger differences on valence annotations. As this dataset has fewer samples in the fourth and second quadrants, we obtained 150 samples from the first and third quadrants, 125 from the second and 90 from the fourth before randomly sampling 15 clips from each of the four quadrants. We also repeated 10 times of random sampling, and selected the one with reliability values closest to those of MER60.

TABLE 4
Summary of Experiment Design

	Training data	Testing data	Factor examined
Exp. 1	Complete set	Complete set	Combined features
Exp. 2	Selected set (random)	Complete set	Size of training dataset
Exp. 3	Selected set (controlled)	Complete set	Reliability level (close to that of MER 60) of training dataset
Exp. 4	Selected set (controlled)	Selected set (controlled)	Reliability level (close to that of MER 60) of both training and testing dataset
Exp. 5	Selected sets (varied)	Selected sets (varied)	Reliability level (varied) of both training and testing dataset

Table 3 lists the reliability values of the original and selected datasets in Experiments 3 and 4. In Experiment 3, the selected subsets were used as training data only. The unselected samples were used as testing data for within-dataset predictions, and all samples in the other datasets for cross-dataset predictions. In contrast, Experiment 4 used the selected subsets for both training and testing. That is, a cross validation on the selected subset was conducted for within-dataset prediction, whereas in cross-dataset prediction a model was trained on the selected subset of one dataset and tested on that of another.

Finally, Experiment 5 was for quantifying the effect of annotation reliability levels on regression performances. Unlike the case in Experiments 3 and 4, this time both CH818 and AMG1608 datasets were sampled multiple times to construct subsets with varying reliability levels (see more in Section 6.2). Table 4 summarizes the experiment designs and the factor each experiment is designed to examine.

As the Support Vector Regression model (SVR) with the radial basis function (RBF) kernel often outperformed other regression models in music mood prediction [3], [22], [44], we adopt this regression model throughout experiments in this study. The experiments were conducted with the scikit-learn package [45]. The parameters (*C* and *gamma*) were optimized by grid searches using the training data in each experiment except for the step-wise forward feature selection. Due to the high computational complexity involved in the feature selection process, the parameter values used were those tuned in the model of the best-performing single feature set (which was selected in the first round) for each corresponding dataset combination. When grid search was used in other parts of the experiments, the process was only conducted within training examples (with an additional layer of cross validation), and thus the regression models constructed were valid in qualifying the generalizability of the models.

Performance is measured by R^2, the square of correlations between predicted values and groundtruth values.

TABLE 5
Regression Performance (in R^2) of Individual Feature Sets on Valence Dimension

Training	Testing	l_m	l_p	p	pCP	pCLP	pCE	pCR	rCTD	rCTA	r_m	t_mf	t_mi	t_p	h_m	h_p
MER	MER	0.14	0.10	0.06	0.27	0.15	0.19	0.15	0.21	0.24	0.08	**0.28**	0.08	0.18	0.25	0.30
CH	MER	0.00	**0.22**	0.00	0.01	0.01	0.04	0.00	0.02	0.10	0.03	0.02	0.03	0.19	0.13	0.07
AMG	MER	0.03	**0.29**	0.09	0.04	0.02	0.05	0.01	0.03	0.09	0.04	0.19	0.16	0.28	0.11	0.21
MER	CH	0.00	0.06	0.03	0.01	0.03	0.01	0.00	0.04	0.07	0.03	0.02	0.06	0.08	**0.09**	0.03
CH	CH	0.11	**0.25**	0.20	0.15	0.22	0.14	0.07	0.23	0.24	0.12	0.22	0.23	0.19	0.16	0.18
AMG	CH	0.00	**0.21**	0.02	0.05	0.07	0.04	0.01	0.12	0.18	0.07	0.1	0.01	0.11	0.04	0.06
MER	AMG	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	**0.04**	0.03	**0.04**	0.01	0.00
CH	AMG	0.00	0.04	0.00	0.02	0.02	0.03	0.00	0.02	**0.06**	0.04	0.01	0.01	0.03	0.04	0.02
AMG	AMG	0.01	0.09	0.04	0.05	0.07	0.06	0.02	0.02	0.07	0.08	**0.12**	0.06	0.10	0.08	0.06
WITHIN		0.09	0.15	0.10	0.16	0.15	0.13	0.08	0.15	0.18	0.09	**0.21**	0.12	0.16	0.16	0.18
CROSS		0.01	**0.14**	0.03	0.02	0.03	0.03	0.00	0.04	0.09	0.04	0.06	0.05	0.12	0.07	0.07
ALL		0.03	**0.14**	0.05	0.07	0.07	0.06	0.03	0.08	0.12	0.06	0.11	0.07	0.13	0.10	0.10

The best performance in each row is highlighted with bold font. l_m = loudness_mirtb, l_p = loudness_psysound, p = pitch, pCP = pitchCP, pCLP = pitchCLP, pCE = pitchCENS, pCR = pitchCRP, rCTD = rhythmCTDFT, rCTA = rhythmCTACF, r_m = rhythm_mirtb, t_mf = timbre _mfcc, t_mi = timbre_mirtb, t_p = timbre_psysound, h_m = harmony_mirtb, h_p = harmony_psysound.

Within its range of [0, 1], the higher R^2 is, the better the performance is. For *within-dataset experiment* (i.e., models trained and tested in one dataset), performances were averaged across 10-fold cross validation. For *cross-dataset experiment* (i.e., models trained in one dataset and tested in another), performances were averaged across 10 segmentations of the testing dataset. The two tailed *t*-test with heterogeneous variations was applied to compare the performances. In determining whether a model is generalizable within- and cross-datasets, we compare its performance to the performance levels reported in recent literature which could be seen as reflecting the state-of-the-art in music mood regression. Admittedly, sometimes it could be somewhat arbitrary to set a lower bound for "acceptable" performances, and thus for cases in cross-dataset predictions, we also compare the performances to those of within-dataset predictions on the same testing datasets.

6 Results and Discussions

6.1 RQ1: Audio Features for Mood Regression

6.1.1 Valence

Table 5 reports the performances of each feature set on the valence dimension, with different combinations of training and testing datasets. It also includes the averaged performances of three within-dataset regressions, six cross-dataset regressions, and all nine regressions. The following observations can be made.

1. The two loudness feature sets showed large difference in performance (0.03 versus 0.14 on overall performance). Among all the fifteen feature sets, loudness_psysound was the best performing feature set for cross-dataset and overall regressions whereas loudness_mirtb performed nearly the worst for both within- and cross-dataset cases. In comparing these two loudness feature sets (cf. Section 4.1), we can see the benefit of modeling the human auditory system as done by PsySound in valence prediction.
2. Pitch features did not perform well for cross-dataset prediction, although some of them (pitchCP and pitchCLP) were quite good for within-dataset cases. Harmony features demonstrated similar patterns: they also performed well for within-dataset predictions, but were mediocre for cross-dataset situations. These results may suggest that the relationships between pitch and valence as well as those between harmony and valence might be cultural or dataset specific. Models worked well for one dataset may not be generalized to another.
3. The fact that the rhythmCTACF feature set worked well for both within- and cross-dataset situations seems in accordance with findings in music psychology which associate flowing/fluent rhythm with positive valence, and firm rhythm with negative valence [46]. The result is also benefited from the adoption of the new tempogram features produced by the Tempogram toolbox, especially when considering the mediocre performances of the conventional rhythm features (rhythm_mirtb). In addition, the CTACF (autocorrelation-based cyclic tempogram) features performed better than the CTDFT (Fourier-based cyclic tempogram) features consistently across the dataset combinations, suggesting sub-harmonies of the beats (i.e., fractions of whole beats) might be more salient than harmonies of them (i.e., multiples of whole beats) in valence prediction.
4. Timbre features worked well for both within- and cross-dataset predictions: the MFCC features generated by MIR toolbox performed the best for within-dataset predictions, whereas the dissonance features generated by PsySound performed very well for cross-dataset predictions ($R^2 = 0.12$). The latter may be related to the fact that the sensation of stability and dissonance is found related to not only music content itself but also social and cultural factors [47]. The traditional spectrum features extracted by the MIR Toolbox (timbre_mtb) did not perform well, which is consistent with findings of previous studies [3] that spectrum features *alone* were not very helpful for valence prediction.

The step-wise forward feature selection procedure chose the following three feature sets for combination: loudness_psysound, harmony_psysound, and timbre_psysound. The

TABLE 6
Regression Performances (in R^2) of Experiments 1-4 on Valence Dimension Using the Combined Feature Set

Training	Testing	Exp. 1	Exp. 2	Exp .3	Exp. 4
MER	MER	**0.18**	**0.18**	**0.18**	**0.18**
CH	MER	**0.27**	0.19	0.25	0.25
AMG	MER	**0.40**	0.22	0.20	0.20
MER	CH	0.13	0.13	0.13	**0.53**
CH	CH	0.25	0.15	0.10	**0.28**
AMG	CH	**0.21**	0.06	0.04	0.20
MER	AMG	0.08	0.08	0.08	**0.33**
CH	AMG	0.07	0.02	0.02	**0.13**
AMG	AMG	0.14	0.03	0.06	**0.19**
WITHIN		0.19	0.12	0.11	**0.22**
CROSS		0.19	0.12	0.12	**0.28**
ALL		0.19	0.12	0.12	**0.26**

Models were built using the compound features comprising of loudness_psysound, timbre_psysound, and harmony_psysound. The best performance in each row is highlighted with bold font.

performances of the combined feature set are shown in Table 6 (the column of "Exp. 1"). With the combined feature set, cross-dataset performance was considerably improved from those with single feature sets (from 0.14 in Table 5 to 0.19). In particular, models trained on AMG1608 outperformed those trained with individual feature sets for large margins. In particular, the prediction on MER60 using the model trained with AMG1608 achieved a very good performance score ($R^2 = 0.40$) which is seldom seen in studies on music valence prediction [3], [14], [21]. The fact that the combined feature set helped on valence prediction seems to suggest that different feature sets could compensate for one another.

6.1.2 Arousal

Table 7 reports the regression performances of each feature set on the arousal dimension. The following observations are made.

1. Similar to valence prediction, the loudness features extracted by PsySound outperformed those by MIR Toolbox and achieved nearly the highest performance across all within- and cross-dataset predictions. In contrast, traditional loudness features (i.e., loudness_mirtb) did not perform well, which is in accordance to previous studies in mood regression [15] and classification [16]. Therefore, modeling the human auditory system as done by PsySound is helpful for arousal prediction as well.
2. Similar to valence prediction, pitchCLP feature set performed well on within-dataset arousal prediction. However, unlike in valence prediction, pitchCLP performed well on cross-dataset arousal prediction as well. This result suggests pitchCLP features captured some commonality related to the arousal aspect across these datasets. The fact that the performances of pitch feature set (i.e., 88 absolute pitches) did not perform well suggests that chroma features (i.e., the 12 classes of combined pitches) are better than features of absolute pitches for arousal prediction.
3. Unlike in valence prediction, none of the individual rhythm or harmony feature sets performed well, either in within- or cross-dataset predictions. Although studies in music psychology indicate rhythm is related to the arousal dimension (e.g., faster songs with high arousal values and slower ones with low arousal values) [48], studies in MIR [14], [25], [49], [50] have found similar results that acoustic rhythm features *alone* rarely outperformed timbre or pitch features in arousal prediction.
4. Timbre features performed generally well, but MFCC, which performed the best in within-dataset valence prediction, was not as good as the other two timbre feature sets here. Both timbre features extracted by MIR toolbox and those extracted by PsySound performed very well, with the latter being the best among all individual feature sets.

Together with the results of valence prediction, we can see the relative strengths of *individual* feature sets in mood regression. Loudness (extracted by PsySound) and timbre features (extracted by MIR Toolbox and PsySound) worked well by themselves for both valence and arousal. However, rhythm features (extracted by Tempogram toolbox) alone worked well only for valence prediction. Pitch features (chroma-abased features extracted by Chroma toolbox)

TABLE 7
Regression Performance (in R^2) of Individual Feature Sets on Arousal Dimension

Training	Testing	l_m	l_p	p	pCP	pCLP	pCE	pCR	rCTD	rCTA	r_m	t_mf	t_mi	t_p	h_m	h_p
MER	MER	0.30	0.78	0.48	0.70	0.73	0.69	0.34	0.59	0.29	0.46	0.68	**0.81**	0.73	0.44	0.36
CH	MER	0.11	0.81	0.39	0.71	0.75	0.67	0.19	0.51	0.42	0.50	0.67	0.73	**0.83**	0.44	0.20
AMG	MER	0.22	0.80	0.61	0.71	0.71	0.63	0.25	0.61	0.39	0.59	**0.84**	0.83	0.73	0.52	0.40
MER	CH	0.03	0.70	0.52	0.49	0.70	0.43	0.08	0.26	0.30	0.33	0.59	0.69	**0.71**	0.34	0.24
CH	CH	0.30	**0.77**	0.68	0.62	0.76	0.54	0.38	0.58	0.62	0.47	0.71	0.76	0.75	0.57	0.68
AMG	CH	0.09	0.70	0.45	0.61	**0.71**	0.54	0.20	0.46	0.55	0.39	0.57	0.67	0.66	0.49	0.46
MER	AMG	0.00	0.50	0.27	0.35	0.51	0.34	0.04	0.06	0.07	0.35	0.51	0.54	**0.59**	0.36	0.27
CH	AMG	0.02	0.42	0.12	0.48	0.55	0.44	0.08	0.14	0.27	0.4	0.4	0.53	**0.58**	0.39	0.18
AMG	AMG	0.04	0.62	0.46	0.51	0.57	0.47	0.26	0.22	0.31	0.51	0.66	0.63	**0.68**	0.46	0.39
WITHIN		0.21	0.72	0.54	0.61	0.69	0.57	0.33	0.46	0.41	0.48	0.68	0.73	0.72	0.49	0.48
CROSS		0.08	0.66	0.39	0.56	0.66	0.51	0.14	0.34	0.33	0.43	0.60	0.67	**0.68**	0.42	0.29
ALL		0.12	0.68	0.44	0.58	0.67	0.53	0.20	0.38	0.36	0.44	0.63	0.69	**0.70**	0.45	0.35

The best performance in each row is highlighted with bold font. l_m = loudness_mirtb, l_p = loudness_psysound, p = pitch, pCP = pitchCP, pCLP = pitchCLP, pCE = pitchCENS, pCR = pitchCRP, rCTD = rhythmCTDFT, rCTA = rhythmCTACF, r_m = rhythm_mirtb, t_mf = timbre _mfcc, t_mi = timbre_mirtb, t_p = timbre_psysound, h_m = harmony_mirtb, h_p = harmony_psysound.

TABLE 8
Regression Performances (in R^2) of Experiments 1-4 on Arousal Dimension Using the Combined Feature Set

Training	Testing	Exp. 1	Exp. 2	Exp .3	Exp. 4
MER	MER	**0.68**	**0.68**	**0.68**	**0.68**
CH	MER	**0.80**	0.74	0.75	0.75
AMG	MER	**0.75**	0.67	0.65	0.65
MER	CH	**0.70**	**0.70**	**0.70**	0.59
CH	CH	0.77	0.72	0.74	**0.81**
AMG	CH	**0.68**	0.67	0.63	0.66
MER	AMG	0.61	0.61	0.61	**0.73**
CH	AMG	0.66	0.57	0.59	**0.71**
AMG	AMG	**0.73**	0.63	0.63	0.72
WITHIN		0.73	0.68	0.68	**0.74**
CROSS		**0.70**	0.66	0.66	0.68
ALL		**0.71**	0.67	0.67	0.70

Models were built using the compound features comprising of loudness_psysound, pitchCLP and timbre_mirtb. The best performance in each row is highlighted with bold font.

alone worked only for arousal prediction. Harmony features (extracted by either MIR Toolbox or PsySound) worked well only for within-dataset valence predictions. These findings are largely in agreement with user studies in MIR. For example, Kaminskyj and Uitdenbogerd [51] found that listeners' music mood perception was affected by music features of intensity/energy (i.e., loudness), tempo and beat strength (i.e., rhythm) but not so much by pitch and tonality (i.e., harmony). Lee et al. [2] also found tempo (i.e., rhythm), instrumentation (often captured by timbre features) affected listeners' mood perception.

Similar to valence prediction, we applied stepwise forward feature selection to identify a combined feature set for arousal prediction. Two individual feature sets were chosen: timbre_psysound and rhythm_mirtb. Although individual rhythm feature sets did not perform well by themselves, it seems they could contribute to arousal prediction and be complementary to timbre features. This result complies with findings in music psychology that rhythm is related to arousal [48]. The performances of the selected combined feature set are presented in Table 8 (the column of "Experiment 1"). In general, the combined feature set achieved slightly better performances than the best-performing single feature set for most dataset combinations. The fact that the combined feature set did not significantly outperform the best single feature set might suggest that performances at this level may have reached the optima of arousal regression across the given datasets.

All the following experiments were conducted with the combined feature sets for valence and arousal respectively.

6.2 RQ2: Cross-Dataset Generalizability

6.2.1 Valence

Table 6 presents the performances of Experiments 1 to 4 on valence dimension. Results on the complete datasets (Experiment 1) show that the models trained on AMG1608 worked much better on MER60 ($R^2 = 0.40$) than on CH818 ($R^2 = 0.21, p < 0.001$). Similarly, the models trained on CH818 worked better on MER60 ($R^2 = 0.27$) than on AMG1608 ($R^2 = 0.07, p < 0.001$). These observations seem to suggest that MER60 lies in the "middle ground" between the CH818 and AMG1608 datasets, in that models built either with CH818 or AMG1608 could be applied to MER60. It is also noteworthy that AMG1608 seems a good training dataset, as the trained model can be applied to all three datasets with reasonably good performances ($R^2 = 0.40, 0.21$ and 0.14).

Experiment 2 limited the size of training datasets to 60 randomly selected samples. The results show that the performance levels generally reduced from those of the original datasets (Experiment 1, $p = 0.01$). This indicates that the size of the training datasets indeed mattered, and more training data was beneficial to both within- and cross-dataset prediction of music valence. In particular, the cross-dataset generalizability from AMG1608 to CH818 did not hold any more, with a performance ($R^2 = 0.06$) significantly worse than that of within-dataset prediction on CH818 ($R^2 = 0.21$, $p < 0.001$). It is also noteworthy that the performance of training on AMG1608 and testing on MER60 is reduced from 0.40 in Experiment 1 (with 1,608 training examples) to 0.22 in Experiment 2 (with 60 training examples). Therefore, the generalizability of valence regression models trained on the AMG1608 dataset did benefit from the considerably large size of this dataset. This finding is in accordance with those in previous studies that a higher number of samples helps classification or regression models to better deal with the noise underlying the data [52].

Experiment 3 controlled the annotation *reliability level* of the *training* datasets in addition to limiting their size to 60 samples. The testing datasets were still all the samples (excluding the training samples in within-dataset cases). The results did not improve in comparison with those of Experiment 2. Although the reliability of the AMG1608 subset ($\alpha = 0.403$) increased from that of the complete dataset ($\alpha = 0.306$) to a remarkable extent, the performances of models trained on the AMG1608 subset did not improve. It seems *improving the reliability level of training dataset did not help as much as having more training examples for valence prediction*.

Experiment 4 controlled the reliability levels of both training and testing datasets such that they were the same (for within-dataset predictions) or close to each other (for cross-dataset predictions). The results show increased performances across most dataset combinations. In particular, predictions on CH818 and AMG1608 increased remarkably from previous experiments. Both cross-dataset and within-dataset predictions achieved the best average performances so far. It is noteworthy that the reliability of the CH818 subset ($\alpha = 0.458$) is actually reduced from that of the complete CH818 dataset ($\alpha = 0.491$, Experiment 1), whereas the performances on CH818 did not reduce, but increased. Therefore, these results indicate that *equalizing annotation reliability of training and testing datasets help improve performances for valence prediction*.

As the effects of annotation reliability levels on music mood regression performances have rarely been studied before, in Experiment 5, we aimed at quantifying the extent to which annotation reliability contributed to regression performances. Specifically, we sampled CH818 and AMG1608 five times respectively, with controlled valence annotation reliability levels ranked from high to low (Table 9). For CH818, songs were ranked by average differences between valence annotations given by different annotators; whereas

TABLE 9
Regression Performances (in R^2) of Five Subsets of CH818 and AMG1604 on Valence Dimension (Experiment 5)

		Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
α	CH	0.964	0.896	0.771	0.232	−0.223
α	AMG	0.636	0.447	0.282	0.171	0.120
Training	Testing					
MER	MER	**0.18**	**0.18**	**0.18**	**0.18**	**0.18**
CH	MER	0.17	**0.23**	0.16	0.16	0.15
AMG	MER	0.15	**0.31**	0.27	0.17	0.20
MER	CH	0.14	0.20	**0.32**	0.24	0.10
CH	CH	0.15	**0.19**	0.18	0.11	0.13
AMG	CH	**0.33**	0.17	0.17	0.17	0.20
MER	AMG	**0.31**	0.27	0.27	0.14	0.27
CH	AMG	0.31	**0.34**	0.13	0.16	0.28
AMG	AMG	0.17	**0.18**	0.10	**0.18**	0.10
WITHIN		0.17	**0.18**	0.15	0.16	0.14
CROSS		0.23	**0.25**	0.22	0.17	0.20
ALL		0.21	**0.23**	0.20	0.17	0.18

The best performance in each row is highlighted with bold font.

for AMG1608, songs were ranked by variance among valence annotations. Following the sampling method used in Experiments 3 and 4, each of the resultant subsets consisted of 60 samples, with equal number of them with positive or negative valence values. The reliability values and regression performances of each subset are presented in Table 9.

Statistical tests show that subset 1 outperformed subset 5 ($p = 0.04$) and subset 2 outperformed subsets 4 and 5 ($p = 0.02$ and 0.01). As Experiment 4 was conducted with the same settings as Experiment 5, only with different subsets, the results of Experiment 4 are also compared and shown to be significantly better that those of subsets 3 and 4 ($p = 0.03$ and 0.02). The general trend is that regression performances were better on datasets with higher reliability levels. It is noteworthy that the best performances for cross-dataset predictions between CH818 and MER60 were achieved in Experiment 4 where the reliability levels of the two subsets were the closest among all subsets of CH818 ($\alpha = 0.458$ for CH and 0.387 for MER60). Similar observations are found for cross-dataset predictions between AMG1608 and MER60: the best performances occurred when the reliability levels of the two subsets were fairly close (subset 2 and Experiment 4). However, cross-dataset performances between CH818 and AMG1608 in Experiment 4 where the reliability levels were the closest were much worse than those of subset 1 where the reliability levels of both subsets were the highest.

As the size and annotation reliability level of the datasets were controlled in Experiments 4 and 5, we may consider the performance differences from a cross-cultural perspective. Cross-dataset generalizability held between MER60 and AMG1608 ($R^2 = 0.24$ 0.31 on most of the subsets) as well as between MER60 and CH818 ($R^2 = 0.25$ and 0.53 in Experiment 4). This may be due to the fact that MER60 shares the same music cultural background with AMG1608 and the same annotator cultural background with CH818. The results that cross-dataset predictions between CH818 and AMG1608 could only work well when both subsets had high reliability (subset 1, $R^2 = 0.31$ and 0.33), coupled with the fact that the CH818 and AMG1608 datasets have neither music or annotators' cultural background in common, seem to suggest that *cross-cultural generalizability for valence prediction cannot be warranted when the datasets differ in culture background of both music materials and annotators, unless the annotation reliability of both datasets is high.*

6.2.2 Arousal

The results of Experiments 1 to 4 on the arousal dimension are shown in Table 8. The original datasets (Experiment 1) performed generally well for both within- and cross-dataset predictions, with performances ($R^2 = 0.73$ and 0.70) comparable to those in the literature (e.g., [53] reported 0.71). For MER60, the models trained with the CH818 and AMG1608 datasets even outperformed the within-dataset prediction.

Similar to the results in valence prediction, Experiment 2 shows that the size of the training datasets mattered. The overall performances were significantly worse compared to those in Experiment 1 ($p = 0.01$) The performances of models trained with the CH818 and AMG1608 subsets degraded significantly after limiting the size of training datasets to 60. Controlling the reliability level of training subsets (Experiment 3) did not help the performance either. When both training and testing datasets were controlled (Experiment 4), the averaged within- and cross-dataset performances improved over Experiments 2 and 3. However, different from valence prediction, the overall performance of Experiment 4 ($R^2 = 0.70$) is not significantly different from that of the original dataset (Experiment 1, $R^2 = 0.71, p = 0.20$), which again supports that *datasets of larger sizes are desirable for music arousal prediction*. On the other hand, the performances of cross-dataset predictions on the AMG1608 subset increased remarkably over those in Experiment 1 (R^2 increased from 0.61 to 0.73 for training on MER60 and testing on AMG1608; R^2 increased from 0.66 to 0.71 for training on CH818 and testing on AMG1608). This is probably due to that fact that the reliability level of the testing data increased remarkably compared to that of the original dataset (α from 0.458 to 0.627).

Similar to valence prediction, a series of subsets of CH818 and AMG1608 were sampled with varying reliability levels in Experiment 5, with similar sampling method to that used in the valence counterpart. The results are shown in Table 10. Statistical tests indicate that subset 1 significantly outperformed all other subsets including those in Experiment 4 ($p = 0.00 \sim 0.04$). In addition, subset 5 was significantly worse than all other subsets including those in Experiment 4 ($p < 0.001$). These indicate a general trend that regression performance on arousal would benefit from higher reliability levels of the datasets. Upon looking closer to cross-dataset predictions, we can see that the predictions between MER60 and CH818 performed the best on subsets 1 and 2. Those between MER60 and AMG1608 had the best performance on subset 1. These best-performing subsets were not those with closest reliability values to that of MER60 ($\alpha = 0.704$). The best performances between CH818 and AMG1604 also occurred on subset 1, not when their reliability levels were the closest (in Experiment 4). Therefore, these results suggest a different conclusion from that in valence prediction: *for arousal*

TABLE 10
Regression Performances (in R^2) of Five Subsets of CH818 and AMG1604 on Arousal Dimension (Experiment 5)

		Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
α	CH	0.987	0.933	0.802	0.321	−0.255
α	AMG	0.781	0.646	0.494	0.365	0.201
Training	Testing					
MER	MER	**0.68**	**0.68**	**0.68**	**0.68**	**0.68**
CH	MER	**0.72**	**0.72**	0.69	**0.72**	0.36
AMG	MER	0.71	0.66	0.73	**0.74**	**0.74**
MER	CH	**0.82**	0.79	0.75	0.44	0.27
CH	CH	**0.85**	0.82	0.75	0.48	0.25
AMG	CH	0.80	**0.83**	0.69	0.45	0.15
MER	AMG	**0.80**	0.61	0.70	0.75	0.46
CH	AMG	**0.75**	0.69	0.70	0.71	0.20
AMG	AMG	**0.79**	0.75	**0.79**	0.73	0.51
WITHIN		**0.78**	0.75	0.74	0.63	0.48
CROSS		**0.77**	0.72	0.71	0.64	0.36
ALL		**0.77**	0.73	0.72	0.63	0.40

The best performance in each row is highlighted with bold font.

prediction, higher reliability levels are preferred rather than those equalized between training and testing datasets.

As Experiments 4 and 5 had the dataset size and reliability level controlled, we may examine the cross-dataset performances from a cross-cultural perspective. Different from valence prediction, in arousal prediction, when the reliability levels of the datasets are sufficiently high (e.g., subsets 1-3 in Experiment 5 and that in Experiment 4), all cross-dataset predictions worked well. Therefore, *annotation reliability level seems the most important factor for cross-dataset generalizability of models on arousal prediction*. For subset 4, performances with MER60 and AMG1608 subsets as testing data were generally good, whereas performances with CH818 subset as testing data were remarkably worse, regardless of the training dataset. This reduction of performances on CH818 subset seems to be in accordance with the reduction of reliability level on this dataset (α from 0.801 in subset 3 to 0.321 in subset 4). One may also observe that the reliability level of the AMG1608 subset was similar ($\alpha = 0.365$), but performances on the AMG1608 subset were much better ($R^2 = 0.71$ to 0.75). We conjecture this may possibly be related to the small number of annotators with highly similar background involved in CH818 which may have boosted the values of reliability measure while the data instances being annotated were still quite heterogeneous. In subset 5, the performances are generally low for both within- and cross-dataset predictions, which corresponds to the low reliability levels of the subsets. It is noteworthy that, except for the cases involving the CH818 subset (whose reliability level, $\alpha = -0.255$, is exceptionally low), cross-dataset performances were comparable to within-dataset performances.

In summary, the results of arousal prediction seem to support: 1) reliability level of dataset is very important for both within- and cross dataset regression on arousal; and 2) cross-cultural generalizability of arousal regression models is largely supported as long as the reliability levels of dataset are not too low. In other words, music arousal is generally predictable from acoustic features across datasets with music in different cultures or annotated by listeners with different cultural backgrounds.

7 Conclusion and Future Work

This study explored the cross-dataset generalizability of music mood regression models in the valence and arousal dimensions. Fifteen audio feature sets in five musical aspects were extracted from and evaluated for three unique datasets. The results revealed that loudness and timbre features worked well by themselves for both valence and arousal prediction. Features of music rhythm alone were effective for valence prediction. Chroma features alone were helpful for arousal prediction and within-dataset prediction for valence. Harmony features by themselves were good for within-dataset valence prediction but did not help for arousal prediction. In addition, the study also found that the combination of multiple feature sets further improved regression performances, especially on the valence dimension. The effectiveness of feature sets evidenced in this study is instructive for future research on the computational modeling of music mood.

A series of experiments were conducted to examine different factors of the datasets and their effects on music mood regression. In general, within-dataset predictions outperformed cross-dataset predictions except for Experiment 4 on valence predictions in which both training and testing datasets had controlled size and annotation reliability levels. The comparison of experimental results show that a larger size of training datasets helped regression performances on both valence and arousal dimensions. In addition, balanced reliability levels of training and testing datasets seemed to help cross-dataset valence prediction, whereas higher reliability level in both training and testing datasets helped on arousal prediction.

When both the size and reliability factors were controlled, cultural factors of the datasets were considered. The results provide evidence that a common cultural background in the datasets is important for cross-dataset predictions in *valence* dimension. When the two datasets contain music in the same cultural background (i.e., stimuli cultural similarity) or mood annotations made by listeners in the same cultural background (i.e., subject cultural similarity), the regression models trained on one dataset could be applied to the other with acceptable performances or performances comparable to the within-dataset prediction. However, when the two datasets differed in both music and annotator's cultural background, the generalizability of trained models did not hold, unless both datasets had high annotation reliability.

For *arousal* dimension, there was no clear observation on the effect of cultural background of dataset on regression performances. When the reliability levels of involved datasets were not too low, cross-dataset performances on arousal prediction were similar to within-dataset performance. The results seem to indicate that the arousal aspect of music mood can be consistent across Western and Chinese Pop songs, and annotators with different cultural backgrounds may perceive music arousal in similar ways.

The findings of this study also raise further questions for future study. For example, although it is rarely seen in MIR to include multiple independent and substantial datasets, it

is our hope to investigate cross-dataset and cross-cultural MIR with more and diverse datasets. All three datasets used in this study essentially consist of Pop music, which might have contributed to the generalizability reflected by the results. It remains a question whether and to what extent the conclusions would hold for datasets with specific cultural traditions such as Indian classical music, Beijing Opera or Flamenco. The fact that the model trained on Chinese songs (the CH818 dataset) can be applied to English songs (the MER60 dataset) may not necessarily suggest that the model could be applied to a set of Indian songs, even if they are also annotated by Chinese listeners. Therefore, further extensions of this study could include evaluation of more datasets, preferably those annotated by listeners unfamiliar to the music. Such evaluations will provide further empirical evidence and directions toward cross-dataset and cross-cultural music information access and retrieval.

Acknowledgments

The authors thank the anonymous reviewers for their helpful and detailed suggestions on improving the manuscript. This study is supported in part by a Seed Fund for Basic Research from the University of Hong Kong and Grant NSC 102-2221-E-001-004-MY3 from the Ministry of Science and Technology of Taiwan.

References

[1] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception: An Interdisciplinary J.*, vol. 30, no. 3, pp. 307–340, 2013.

[2] J. H. Lee, T. Hill, and L. Work, "What does music mood mean for real users?" in *Proc. iConf.*, Feb. 2012, pp. 112–119.

[3] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, p. 40, May 2012.

[4] M. Barthet, G. Fazekas, and M. Sandler, "Music emotion recognition: From content to context-based models," in *Sounds to Music and Emotions*. Berlin, Germany: Springer-Verlag, 2013, pp. 228–252.

[5] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. Ehmann, "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. Int. Conf. Music Inform. Retrieval*, 2008, pp.462–467.

[6] A. Aljanaki, Y. H. Yang, and M. Soleymani, "Emotion in music task at mediaeval 2014," in *Proc. MediaEval Workshop*, Barcelona, Spain, vol. 1263, 2014.

[7] X. Serra, "A multicultural approach in music information research," in *Proc. Int. Conf. Music Inform. Retrieval*, 2011, pp. 151–156.

[8] J. H. Lee and X. Hu, "cross-cultural similarities and differences in music mood perception," in *Proc. iConf.*, 2014, pp. 259–269.

[9] X. Hu, J. H. Lee, K. Choi, and J. S. Downie, "A cross-cultural study on the mood of K-POP songs," in *Proc. Int. Conf. Music Inform. Retrieval*, 2014, p. 385–390.

[10] A. Singhi and D. G. Brown, "On cultural, textual and experiential aspects of music mood," in *Proc. Int. Conf. Music Inform. Retrieval*, 2014, pp. 1–6.

[11] A. H. Gregory and N. Varney, "Cross-cultural comparisons in the affective response to music," *Psychol. Music*, vol. 24, no. 1, pp. 47–52, 1996.

[12] P. C. Wong, A. K. Roy, and E. H. Margulis, "Bimusicalism: The implicit dual enculturation of cognitive and affective systems," *Music Perception*, vol. 27, no. 2, p. 81, 2009.

[13] X. Hu and J. H. Lee, "A cross-cultural study of music mood perception between American and Chinese listeners," in *Proc. Int. Soc. Music Retrieval*, 2012, pp. 535–540.

[14] K. Kosta, Y. Song, G. Fazekas, and M. B. Sandler, "A study of cultural dependence of perceived mood in Greek music," in *Proc. Int. Soc. Music Retrieval*, 2013, pp. 1–6.

[15] X. Hu and Y. H. Yang, "A study on cross-cultural and cross-dataset generalizability of music mood regression models," in *Proc. Sound Music Comput. Conf.*, 2014, pp. 1149–1155.

[16] Y. H. Yang and X. Hu, "Cross-Cultural music mood classification: A comparison on English and Chinese songs," in *Proc. Int. Conf. Music Inform. Retrieval*, Oct. 2012, pp. 19–24.

[17] K. Hevner, "Experimental studies of the elements of expression in music," *Amer. J. Psychol.*, vol. 48, no. 2, pp. 246–268, 1936.

[18] J. A. Russell, "A circumplex model of affect," *J. Psychol. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[19] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, "The music information retrieval evaluation exchange: Some observations and insights," in *Advances in Music Information Retrieval*. Berlin, Heidelberg: Springer, 2010, pp. 93–115.

[20] J. Madsen, J. B. Nielsen, B. S. Jensen, and J. Larsen, "Modeling expressed emotions in music using pairwise comparisons," in *Proc. Int. Symp. Comput. Music Model. Retrieval*, 2012, pp. 526–533.

[21] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. Int. Soc. Music Inf. Retrieval*, 2010, pp. 255–266.

[22] A. Huq, J. P. Bello, and R. Rowe, "Automated music emotion recognition: A systematic evaluation," *J. New Music Res.*, vol. 39, no. 3, pp. 227–244, 2010.

[23] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a Gaussian mixture model," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 56–68, Jan.–Mar. 2015.

[24] T. Eerola, "Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *J. New Music Res.*, vol. 40, no. 4, pp. 349–366, 2011.

[25] Y. H. Yang and H. H. Chen, "Predicting the distribution of perceived emotions of a music signal for content retrieval," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2184–2196, Sep. 2011.

[26] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA, USA: Sage, 2012.

[27] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.* , Mar. 2012, pp. 341–344.

[28] Y. A. Chen, Y. H. Yang, J. C. Wang, and H. H. Chen, "The AMG1608 dataset for music emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2015, pp. 693–697.

[29] K. Drossos, A. Floros, A. Giannakoulopoulos, and N. Kanellopoulos, "Investigating the impact of sound angular position on the listener affective state," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 27–42, Jan. 2015.

[30] M. M. Bradley and P. J. Lang, "International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings," The Center for Research in Psychophysiology, Univ. Florida, Gainesville, FL, USA, Tech. Rep. B-2, 1999.

[31] P. N. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception," *J. Exp. Psychology: Human Perception Perform.*, vol. 26, no. 6, p. 1797, 2000.

[32] A. Gabrielsson and E. Lindström, "The role of structure in the musical expression of emotions," in *Handbook of Music and Emotion: Theory, Research, Applications*, London, U.K.: Oxford Univ. Press, 2010, pp. 367–400.

[33] E. Coutinho and A. Cangelosi, "Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements," *Emotion*, vol. 11, no. 4, p. 921, 2011.

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

[35] D. Cabrera, "Psysound: A computer program for psychoacoustical analysis," in *Proc. Australian Acoustic Soc. Conf.*, 1999, pp. 47–54.

[36] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proc. 12th Int. Conf. Music Inform. Retrieval*, 2011, pp. 215–220.

[37] P. Grosche and M. Muller, "Extracting predominant local pulse information from music recordings," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1688–1701, Aug. 2011.

[38] P. Grosche, M. Muller, and F. Kurth, "Cyclic tempogram—a mid-level tempo representation for music signals," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2010, pp. 5522–5525.

[39] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 5–18, Dec. 2006.
[40] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
[41] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE J. Select. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, Oct. 2011.
[42] X. Hu, K. Choi, and J. S. Downie, "A framework for evaluating multimodal music mood classification," *J. Association Inf. Sci. Technol.*, 2016.
[43] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. New York, NY, USA: Springer- Verlag, 2014.
[44] M. Caetano and F. Wiering, "The role of time in music emotion recognition," in *Proc. Int. Symp. Comput. Music Modeling Retrieval*, 2012, pp. 287–294.
[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[46] A. Gabrielsson and E. Lindström, *The Influence of Musical Structure on Emotional Expression*. New York, NY, USA: Oxford Univ. Press, 2001.
[47] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. New York, NY, USA: Springer-Verlag, 2002.
[48] A. Gabrielsson and E. Lindström, *The Influence of Musical Structure on Emotional Expression*. New York, NY, USA: Oxford Univ. Press, 2001.
[49] B. Schuller, C. Hage, D. Schuller, and G. Rigoll, "'Mister D.J., cheer me up!' Musical and textual features for automatic mood classification," *J. New Music Res.*, vol. 39, no. 1, pp. 13–34, 2010.
[50] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," in *Proc. Int. Conf. Music Inform. Retrieval*, 2012, pp. 523–528.
[51] I. Kaminskyj and A. L. Uitdenbogerd, "A study of human mood tagging of musical pieces," in *Proc. Int. Conf. Music Commun. Sci.*, 2007, p. 68.
[52] Y.-H. Yang and H.-H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech Language Process.*, vol. 19, no. 4, pp. 762–774, May 2011.
[53] D. Guan, X. Chen, and D. Yang, "Music emotion regression based on multi-modal features," in *Proc. Int. Symp. Comput. Music Model. Recog.*, 2012, pp. 70–77.

Xiao Hu received the PhD degree in library and information science from the University of Illinois in 2010. She is an assistant professor in the Division of Information and Technology Studies, Faculty of Education, University of Hong Kong. Her research interests include music mood recognition, information retrieval, and affective computing. She has received the Best Student Paper award in the ACM Joint Conference on Digital Libraries (2010) and Best Student Paper award in the iConference (2010). She was a tutorial speaker on music affect recognition (2012) and a Conference Co-chair (2014) in the International Society for Music Information Retrieval Conference.

Yi-Hsuan Yang (M'11) received the PhD degree in communication engineering from the National Taiwan University in 2010, and joined Academia Sinica as an assistant research fellow in 2011. He is an associate research fellow with Academia Sinica. He is also an adjunct associate professor with the National Tsing Hua University, Taiwan. His research interests include music information retrieval, machine learning, and affective computing. He received the 2011 IEEE Signal Processing Society (SPS) Young Author Best Paper Award, the 2012 ACM Multimedia Grand Challenge First Prize, and the 2014 Ta-You Wu Memorial Research Award of the Ministry of Science and Technology, Taiwan. He is an author of the book *Music Emotion Recognition* (CRC Press 2011) and a tutorial speaker on music affect recognition in the International Society for Music Information Retrieval Conference (ISMIR 2012). In 2014, he served as a Technical Program Co-chair of ISMIR, and a guest editor of the *IEEE Transactions on Affective Computing* and the *ACM Transactions on Intelligent Systems and Technology*. He is a member of the IEEE.