Music Emotion Recognition: The Role of Individuality

Yi-Hsuan Yang, Ya-Fan Su, Yu-Ching Lin, and Homer H. Chen
Graduate Institute of Communication Engineering, National Taiwan University
1 Roosevelt Rd. Sec.4, Taipei, 10617, Taiwan R.O.C.
886-2-3366-3549
affige@gmail.com, b92901017@ntu.edu.tw, b92901058@ntu.edu.tw, homer@cc.ee.ntu.edu.tw

ABSTRACT
It has been realized in the music emotion recognition (MER) community that personal difference, or individuality, has significant impact on the success of an MER system in practice. However, no previous work has explicitly taken individuality into consideration in an MER system. In this paper, the group-wise MER approach (GWMER) and personalized MER approach (PMER) are proposed to study the role of individuality. GWMER evaluates the importance of each individual factor such as sex, personality, and music experience, whereas PMER evaluates whether the prediction accuracy for a user is significantly improved if the MER system is personalized for the user. Experimental results demonstrate the effect of personalization and suggest the need for a better representation of individuality and for better prediction accuracy.

Categories and Subject Descriptors
H.5.5 [**Sound and Music Computing**]: *systems*

General Terms
Algorithms, Performance, Experimentation, Human Factors.

Keywords
Music emotion recognition, individuality, personalization.

1. INTRODUCTION

As the multimedia content explosion continues, it becomes important to develop computers that can appraise the emotion of multimedia content and provide easy and effective information access [1]. Among the various multimedia types, the detection of emotion in music is of particular interest since music is increasingly pervasive and it is the finest language of emotion [2]. Music can bring us tears, console us when we are grief, or drive us to love and hate. Therefore, music classification and retrieval by perceived emotion is natural and functionally powerful. Making computers capable of recognizing the emotion of music also enhances the way human and computer interacts. For example, a wearable device such as MP3 player or cellular phone equipped with MER function can play a song best suited to the emotion of the user [3], [4]; a smart space (e.g. restaurant, conference room, residence) can play background music best suited the people inside it. Automatic prediction of emotion in music is referred to as *music emotion recognition* (MER) throughout the paper.

Though the relationship between music and perceived emotion has been studied by psychologists for decades, the boom of MER can be dated back to within the last ten years [5]–[12]. Being at its preliminary stage, many critical issues of MER are yet to be solved. As [9]–[12] have pointed out, one of the critical issues of MER is the *subjectivity problem*, which stems from the fact that music perception is intrinsically subjective and is under the influence of many individual factors such as personality, sex, cultural background etc. Failure to cope with the subjectivity nature of emotion perception seriously affects the performance of an MER system when a common agreement of the recognition result can not be reached for music selection. In other words, the recognition result may not hold true when the perceived emotions of a song differ a lot for different individuals.

Despite that the subjectivity nature of MER is well recognized, few systems can deal with the subjectivity problem. In particular, to our best knowledge, none of the previous works has explicitly taken the individuality (or individual factors) into consideration in the development of an MER system. Since the subjectivity problem has a great impact on the performance of an MER system, the primary purpose of this paper is to study whether the recognition accuracy can be significantly improved if individuality is considered and whether individuality indeed plays such an important role in MER.

We formulate MER as a regression problem [12] and adopt the support vector regression (SVR) model [13] to train regressors. A baseline regressor is trained based upon the average opinions from the subjective test (also refer to as the *general regressor*). Then we evaluate the prediction accuracy of the following two approaches:

HCM'07, September 28, 2007, Augsburg, Bavaria, Germany.

1) Group-wise MER (GWMER): We group users according to individual factors including sex, academic background, music experience and personality, and train *group-wise regressors* for each user group. The prediction accuracy of each group-wise regressor is then interpreted as the importance of the associated individual factor. For example, if the prediction accuracy is significantly improved when different regressors are used for male and female, it implies human perception of music has distinguishable difference between the two sexes.

2) Personalized MER (PMER): Another approach to resolve the subjectivity problem is to personalize the MER system. Typically the personalization is based on listening history [14] or user feedback mechanisms such as relevance feedback [15], but these methods cannot be applied to MER directly. PMER personalizes the MER system by asking a user to explicit annotate his/her perceived emotions of a number of music selections, and using these annotations as ground truth to train a *personalized regressor*. The performance of the personalized regressor should be better than the general regressor for the particular user if individuality indeed plays such an important role.

The paper is organized as follows. Section 2 reviews previous MER works. Section 3 presents the GWMER and PMER approaches in detail. Section 4 describes the MER system. Experimental results of the two approaches are reported in Section 5. Section 6 concludes the paper.

2. RELATED WORK

Most previous works on MER [5]–[9] categorize emotions into a number of emotion classes and apply the standard pattern recognition procedure to train a classification model. Typically, the emotion classes are defined in terms of arousal (how exciting or calming) and valence (how positive or negative). For example, the emotion classes can be divided into the four quadrants in Thayer's arousal-valence emotion plane [16], see Fig. 1. However, because of the subjectivity problem, an MER system that simply assigns one emotion class to each song in a deterministic manner will not perform well in practice. Below we review previous works that have discussed this problem.

In [9], the subjectivity problem is bypassed by limiting the genre of dataset to western classical music, whose perceived emotion is much easier to gain major agreement within the western culture. However, since the purpose of MER is to facilitate music retrieval and management, and since it is the popular music that dominates the everyday music listening, disregarding the subjectivity problem by using western classical music is inappropriate.

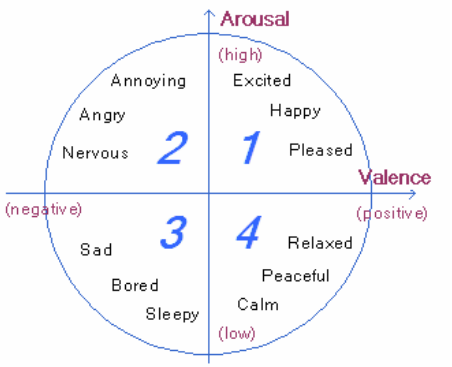

Fig. 1. Thayer's arousal-valence emotion plane.

In [10], the subjectivity problem is addressed by the employment of the fuzzy classifiers, which assign a *fuzzy vector* for a song to indicate the relative strength of each class. For example, $(0.1\ 0.0\ 0.8\ 0.1)^t$ represents a fuzzy vector with the strongest emotion strength for class 3, while $(0.1\ 0.4\ 0.4\ 0.1)^t$ shows an ambiguity between class 2 and 3. The ambiguity that fuzzy vectors carry is important since it reflects how likely individuals may perceive different emotions to a same song.

In [11], the multi-label classifiers, which allow assigning more than one emotion class to the same song, are adopted in response to the fact that human perception of music emotion is not uniform. Principally the underlining ideas of [10] and [11] are the same: to provide emotion intensity measurement for each emotion class.

However, formulating MER as a classification problem as most previous works [5]–[11] is not practical from a user's point of view because of the following two reasons. First, unlike speech emotion recognition [17] where it is only necessary to recognize basic prototypical emotions such as happy, sad or angry, MER requires a finer-grained definition of emotion for effective music retrieval and management. A categorical definition of emotion may be too coarse for a user to retrieve music easily. Second, the adjectives used to describe emotion classes may be ambiguous, and the use of adjectives for the same emotion can vary from person to person.

An alternative is to view the emotion plane as a continuous space and recognize each point of the plane as an emotion state. Specifically, we first compute the arousal and valence values (AV values) of each music sample and represent the music sample as a point in the emotion plane. Then the user can retrieve music by specifying a point in the emotion plane according to his/her emotion state, and the system would return the music samples whose locations are closest to the specified point.

In [12], where the continuous perspective is elaborated, MER is formulated as a regression problem and SVR is

employed to predict the AV values. With this regression approach, the problems inherent to categorical approaches are avoided. In addition, because there is more freedom in describing a song, the subjectivity problem is also alleviated. For instance, besides the quadrant to which the song belongs, one can further know the emotion intensity the song expresses by examining its AV values.

Despite that it has more freedom in describing a song; the regression approach may fail to exactly resolve the subjectivity problem since the regressors are still trained based upon the average opinions of the subjects, and since individual differences in the perception of popular music may be too high. None previous works has explicitly taken individuality into consideration in developing the MER system and none has quantitatively evaluated the effect of individuality on the performance of MER.

3. PROPOSED APPROACHES

In light of the above observations, we propose the GWMER and PMER approaches to study the role of individuality. Both approaches use SVR to predict AV values.

3.1 GWMER

To train a regressor, typically a subjective test that asks participants to annotate the training data and forms the ground truth by averaging participants' opinions is performed. For GWMER, the ground truth is set in an alternative way. To evaluate the importance of individuality, a number of *user groups* are defined based on individual factors include sex, academic background, music experience and personality, and *group-wise regressors* are trained for each user group based on the average opinion of the participants *belonging to each user group*.

The prediction accuracy of each group-wise regressor is interpreted as the importance of the associated individual factor to MER. The rational for this argument is based on the following two observations. First, if individual difference is large for the specific individual factor (e.g. sex), the individual difference should decrease as we group users according to this individual factor. Second, if individual difference has remarkable impact on the performance of MER, the reduction of individual difference should improve the prediction accuracy of MER.

The importance of utilizing personal information for music information retrieval has also been advocated in [18], but no quantitative evaluation is reported. For GWMER, the following individual factors are considered and evaluated:

1) Demographic property: Because the participants of our subjective test are mainly college students, it is relatively less informative to consider age and the level of education. Instead, we group participants by sex and the academic background, which is defined as follows: the first academic group includes colleges of liberal arts, social science, management, and law, and the second academic group includes colleges of engineering, science, life science, medicine, and computer science.
2) Music experience: The perception of music can be influenced by music experiences include music expertise, musicianship, taste and familiarity with the music [19]. We represent music experience in terms of the habit of music listening (rare to often), ability to play an instrument (can or cannot), and the like to listen to music with the following prototypical emotions: happy, exciting, angry, sad, sleepy, or relaxing.
3) Personality: To describe personality, the Big Five personality traits [20] are used. The Big Five are five broad factors or dimensions of personality discovered through empirical psychological research. These factors are extraversion (introversion), agreeableness (antagonism), conscientiousness (lack of direction), neuroticism (emotional stability), and openness to experience (closeness). Typically personality is measured using self report questionnaires, in which the solution to each question contributes to a rating to one of the five personality factors. However, to prevent participants from being exhausted by a lengthy questionnaire, the personality is measured by self report personality inventory [21] instead. We ask participants to rate a list of adjectives related to personality: if the adjective describes his/her personality well, rate +1; if the opposite adjective does, rate –1; otherwise, 0. For example, sociable, adventurous and open-minded are used for measuring extraversion. If a participant rates all the three adjectives +1, he/her gets a rating of +3 for extraversion. Table 2 lists the adjectives we used for self report personality inventory.

A total of 15 individual factors are considered. They are listed in Table 1 along with possible values, which take a binary form. For the five personality factors, we take positive ratings as high (one), and negative as low (zero).

The system diagram of GWMER is shown in Fig. 2. Each group-wise regressor is trained based on the average opinion of the participants belong to a user group. In practice, a collection of group-wise regressors are pre-trained, and the most suitable one will be chosen to respond to a particular user according to the user's personal information.

Note GWMER represents a compromise between a general MER system and a personalized one. By grouping users based on individual factors, GWMER reduces individual differences without resorting to too much user burden (a user only needs to fill in personal information). Moreover, GWMER is quite generic; individual factors can be added to or removed from the system easily. For example, age can be included to consider the generational difference.

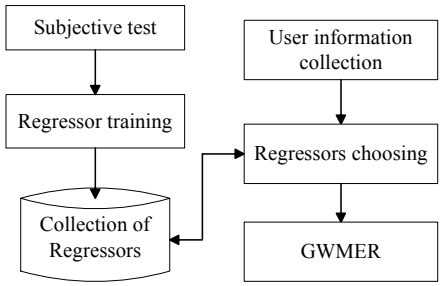

Fig. 2. System diagram of the GWMER approach.

Table 1. The individual factors considered in GWMER

	Type	Name	Value (0/1)
1	Demographic property	Sex	Male / female
2		Academic group	First / second
3	Music experience	Music listening	Rare / often
4		Instrument play	Cannot / can
5		Love happy songs	No / yes
6		Love exciting songs	No / yes
7		Love angry songs	No / yes
8		Love sad songs	No / yes
9		Love sleepy songs	No / yes
10		Love relaxing songs	No / yes
11	Personality (The Big Five)	Extraversion	Low/ high
12		Agreeableness	Low/ high
13		Conscientiousness	Low/ high
14		Neuroticism	Low/ high
15		Openness to experience	Low/ high

Table 2. The adjectives for self report personality inventory

Personality trait	Adjective
Extraversion	Sociable, adventurous, open-minded.
Agreeableness	Forgiving, trusting, cooperative.
Conscientiousness	Organized, careful, self-disciplined.
Neuroticism	Worried, irritable, discontented.
Openness to experience	Imaginative, wide-interested, original.

3.2 PMER

Though it is wonderful to develop a general MER system that performs equally well in practice for each individual, it might be unnecessary. As [2] points out, it is only necessary that one's personal computer be able to recognize *his*/*her* emotion. Therefore, it may be beneficial to personalize the MER system. However, though intuitively correct, for MER no quantitative performance study has been done to evaluate whether the prediction accuracy for each individual is significantly improved by personalization.

A critical issue for personalization relates to the additional user burden. Typically personalization is made based on listening history [14] or relevance feedbacks [15], whose user burdens are believed to be small. For MER, one may exploit the listening history by assuming the AV values for songs listened in the same time period are near, or use relevance feedback to know whether the user is satisfied with the prediction result. However, both methods fail to acquire the exact AV values a user feels the songs are, and thus of little help to improve the prediction accuracy.

PMER personalizes the MER system by asking users to pre-annotate the AV values for a limited number of songs in advance, and training *personalized regressors* based on these annotations. The weighted combination of the predicted AV values of a general regressor and a personalized regressor are then used to predict the remaining songs. Note to prevent too much user burden, the number of songs annotated by a user should be kept reasonably small.

Since the personalized regressor is trained for each user, its performance for the particular user should be higher than the general regressor, if individuality indeed plays such an important role.

4. SYSTEM DESCRIPTION

Our MER system is based on the regression approach proposed in [12], yet differs in the regression training part. The system diagram of our MER system is shown in Fig. 3, and the details are described below.

4.1 Data Collection

To study the role of individuality, each song should be annotated by a sufficient number of participants. Therefore, we collect 60 famous popular songs from English albums, and make each song annotated by 40 participants. The collected music samples are trimmed to 25 seconds by manually trimming the chorus part and converted to a uniform format: 22,050 Hz, 16 bits, and mono channel PCM WAV. The emotions of these songs distribute roughly uniformly in each quadrant of the emotion plane.

Note the scale of our dataset is relatively small because labels are hard to obtain, especially for perceived emotion. To increase the number of annotations per song, we cannot help but reduce the number of songs. However, a dataset of scale similar to the CAL500 data set [22] is indeed called for to make the evaluation more statistically reliable.

4.2 Subjective Test

Participants are recruited form the campus (all non-experts); each rewarded $3 to perform the subjective test. They are asked to annotate 15 of the music samples in a computer lab. Most of the 99 participants feel the annotation easy and pleasant, and many of them volunteer to annotate more songs. 6 participants annotate the entire dataset. In sum, we get 2400 annotations: each of the 60 song is annotated by 40 participants.

We design a user interface called 'AnnoEmo' for the subjective test using the Java language. A participant annotates the AV values for each song by using a mouse to click a point in the emotion plane displayed by computer. The emotion plane is defined as a coordinate space spanned by arousal and valence, where each value is confined within [–1, 1]. After clicking a point, a rectangle is formed on the specified point. The participant can then click on the rectangle to listen to the associate music, or drag and drop the rectangle to other places since after listening to other songs the participant may want to modify the annotation of previous songs. Once formed, the rectangle would exist throughout the subjective test. Therefore, as the participant annotates more songs, more rectangles are presented on the emotion plane, making it easy for the participant to compare the annotations of different music samples. The participants are allowed to listen to the music samples multiple times (by clicking on the rectangle or on a song list) and no limitation is given to the total duration of the annotation process. Typically the annotation process can be finished within 15 minutes for 15 songs. Fig. 4 shows a snapshot of AnnoEmo for annotating the AV values.

Besides, for GWMER, we need to collect personal information of each participant of the subjective test. A snapshot of AnnoEmo for collecting personal information is shown in Fig. 5. We have made the dataset (features and annotations) and the software 'AnnoEmo' available on our website [23].

4.3 Feature Extraction

For feature extraction, two free toolkits PsySound [24] and Marsyas [25] are applied. A total of 45 features are extracted to represent each music sample in the feature space. Each feature is normalized linearly to [0, 1] before the regressor training. To compare the performance of each regressor fairly, no feature selection algorithm is applied.

PsySound extracts features based on a range of psychoacoustical models [24], so it can generate features that are more relevant to emotion perception. The effectiveness of the PsySound features for MER has been demonstrated in [10], and 15 of these features are found particularly related to emotion perception. These features include spectral centroid, loudness, sharpness, timbral width, volume, spectral dissonance, tonal dissonance, multiplicity, tonality, and chord.

Marsyas is a generic software framework for rapid development and evaluation of computer audition applications [25]. It generates 19 timbral texture features (spectral centroid, spectral rolloff, spectral flux, time domain zero-crossing and MFCC), 6 rhythmic content features (by beat and tempo detection) and 5 pitch content features (by multi-pitch detection). Marsyas features have been found important for music signal processing, especially for music genre classification.

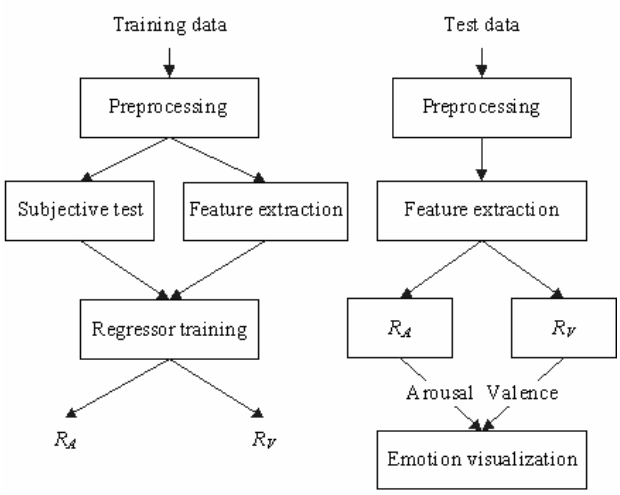

Fig. 3. System diagram of the MER system. Left: training phase; right: testing phase.

Fig. 4. A snapshot of AnnoEmo for annotating the AV values.

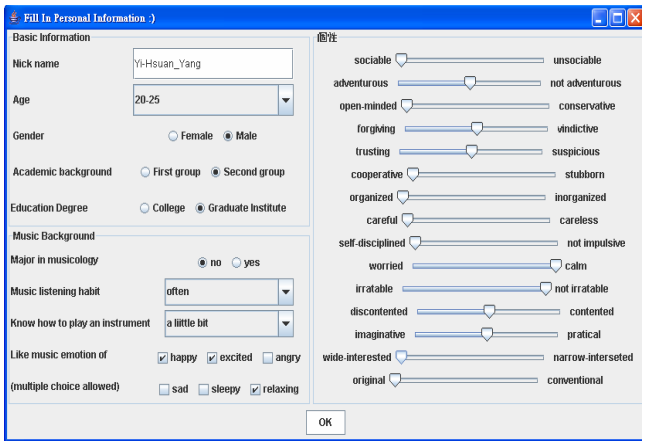

Fig. 5. A snapshot of AnnoEmo for collecting the personal information of a participant.

4.4 Regressor Training

Regression theory [26] aims at predicting a real value from some observed variables (features). It has a sound theoretical foundation and allows easy performance analysis. Given N inputs (x_i, y_i), $1 \leq i \leq N$, where N is the total number of inputs, x_i is the feature vector for the ith input instance, and $y_i \in \mathbb{R}$ ($\mathbb{R}$ denotes a set of real values) is the real value to be predicted for the ith instance, the regression system trains a regression algorithm (regressor) $R(\cdot)$ such that the mean squared error ε expressed below is minimized:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}(y_i - R(x_i))^2 , \quad (1)$$

where $R(x_i)$ is the prediction result for the ith instance by $R(\cdot)$. Since the AV values are viewed upon as real values from the continuous perspective, the regression theory can be well applied to directly predict arousal and valence.

Support vector machines (SVM) has been found in many cases superior to existing machine learning methods; therefore, we adopt SVR, which is an extension of SVM to regression problems, to predict the AV values. Since we want to predict both arousal and valence, two regressors are required and are denoted as R_A and R_V. Our implementation of SVR is based on the library LIBSVM [27]. A grid parameter search is applied to find the best parameters for SVR.

For general MER, a general (baseline) regressor is trained based on the average opinion of all the participants in the subjective test. For GWMER, a number of group-wise regressors are trained for each user group based on the average opinion of participants belongs to each specific user group. For PMER, each user has a personalized regressor that is trained particular for him/her.

4.5 Emotion Visualization

Associated with the AV values, each music sample is visualized as a point in the emotion plane, and the similarity between music samples can be estimated by computing the Euclidean distance in the emotion plane. A user interface that supports music retrieval/recommendation by specifying a point in the emotion plane can be realized in a form similar to 'AnnoEmo'. Such a user interface is of great use in managing large scale music databases.

5. EXPERIMENTAL RESULTS

The prediction accuracy of a regressor can be evaluated in terms of ε and the R^2 statistics defined below:

$$\text{the } R^2 \text{ statistics} = 1 - \frac{N \times \varepsilon}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \quad (2)$$

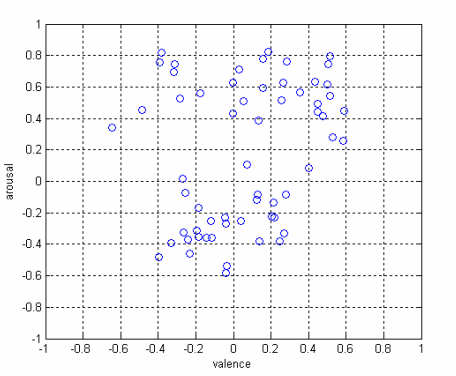

Fig. 6. Distribution of the ground truth data. While many neutral values are assigned to valence, there are two evident clusters along the arousal axis, implying the fact that arousal is easier to be predicted than valence.

where $\bar{y}$ is the mean of the ground truth. It can be observed that the R^2 statistics is actually defined as one minus the total squared error ($N \times \varepsilon$) normalized by the variance of the ground truth. The R^2 statistics is a standard performance measure for regression algorithms [26], and is often interpreted as the proportion of response variation described by the regressors in the model. A negative value of the R^2 statistics means the prediction model is worse than simply taking the sample mean. For our dataset, the variance for valence and arousal is 0.0952 and 0.2055.

The generalization performances of the proposed approaches are evaluated by the leave-one-out (LOO) cross validation technique [28], which is known to provide an almost unbiased estimate of the generalization error even with a small dataset. As the name suggests, LOO uses one data instance of the dataset as the testing data, and uses the remaining instances as training data to train the regressor. This procedure is repeated until each instance is used once as the testing data. The R^2 statistics for valence and arousal are computed separately.

5.1 General Approach

For the baseline general SVR regressor, the R^2 statistics reaches 17% for valence and 79.9% for arousal using LOO. The lower accuracy for valence is consistent to the result of previous MER works. It has been found that generally valence is much more difficult to be predicted than arousal. Two reasons may account for this phenomenon. First, while there are a number of features related to arousal such as loudness (loud/soft), tempo (fast/slow), and pitch (high/low), few salient features have been found for valence. Second, individual difference for valence is larger than that for arousal; there is a good chance that two persons perceive opposite valence toward the same song. Fig. 6 shows the distribution of the ground truth data, from which we can observe while participants can easily distinguish high-arousal songs from low-arousal songs

(there are two evident clusters along the arousal axis), much more neutral values are assigned to valence. Most of the neutral labels come from averaging: half of the participants feel the song positive, while others feel it negative.

To gain more insights to the prediction accuracy, we define three *performance levels* in terms of the mean absolute error η, which is defined in a similar way as ε except the square operation in Eq. (1) is replaced by taking absolute value. As illustrated in Fig. 7, a performance level of p represents the absolute error of the AV values for each song is lower than $0.1\times p$ (within each associated red circle) on average. The corresponding value of the R^2 statistics for each performance level is also depicted. We can observe the prediction accuracy of the general regressor reaches performance level 2 for arousal and level 3 for valence.

5.2 GWMER

In our experiment, individual factors are evaluated one by one. For each individual factor, we partition the participants into two groups 0 or 1 (as the 'Value' column shown in Table 1), and train regressors for each user group. For each individual factor, the prediction accuracy for either user group is evaluated separately. If the accuracies for the two groups are better than the baseline, it implies human perception of music has distinguishable difference between the two user groups, and the reduction of this individual difference improves the prediction accuracy.

Experimental results are tabulated in Table 3. The first column specifies which kind of participant is categorized to group 1 for the individual factor, and the number in the parenthesis indicates how many participants (among the total 99 participants) are categorized to group 1. The resulting R^2 statistics indicates:

1) Individual difference is not removed by each single individual factor; only few group-wise regressors have better prediction accuracy than the baseline regressor, and the improvement is not significant.
2) While the R^2 statistics for arousal is nearly uniform, the R^2 statistics for valence varies a lot. This finding implies again that personal difference in perceiving valence is much larger than in perceiving arousal.
3) For some groups of users such as people who rarely listen to music, the prediction of valence becomes even more difficult.

If the removal of individual difference can indeed improve the prediction accuracy, the failure to make significant improvement by the GWMER approach leads to the following conclusion: individuality is too subtle to be captured by each single individual factor considered in our experiment. We may need to investigate the combination of individual factors or exploit other individual factors to adequately describe individuality.

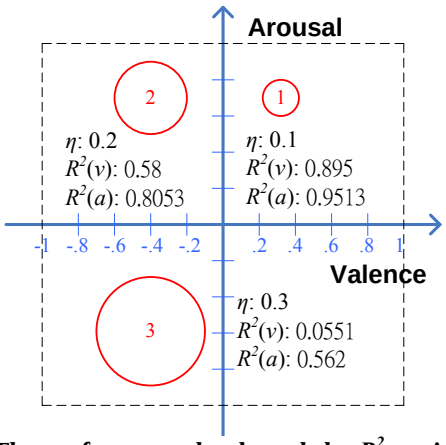

Fig. 7. The performance levels and the R^2 statistics for valence (v) and arousal (a) corresponding to three levels of mean absolute error η.

Table 3. The resulting R^2 statistics (in %) of each individual factor for the GWMER approach

Value of individual factor	Group 1		Group 0	
	$R^2(v)$	$R^2(a)$	$R^2(v)$	$R^2(a)$
Baseline	17.0	80.2	17.0	80.2
Female (53)	16.0	76.1	11.1	80.8
Second academic group (42)	4.6	77.9	19.3	79.7
Often listen to music (67)	20.6	77.9	6.5	77.6
Can play an instrument (44)	8.7	77.7	14.6	74.7
Love happy songs (68)	14.1	79.9	17.4	77.4
Love exciting songs (38)	19.8	73.8	12.2	80.0
Love angry songs (6)	15.7	79.8	13.1	78.7
Love sad songs (29)	6.7	79.9	17.9	79.5
Love sleepy songs (11)	-1.5	72.8	17.2	80.4
Love relaxing songs (63)	15.7	80.0	7.3	76.7
High extraversion (33)	6.1	76.0	13.2	80.0
High agreeableness (36)	13.5	75.2	11.9	77.8
High conscientiousness (38)	12.9	77.3	9.6	80.5
High neuroticism (11)	-2.2	71.6	17.4	79.5
High openness to experience (39)	7.4	77.3	17.4	78.6

5.3 PMER

To evaluate whether the removal of individual difference indeed improves the prediction accuracy, we train a personalized regressor for each user. Here we assume the user burden is not an issue, and exploit the maximal available number of pre-annotated songs. Therefore, LOO is used to evaluate the performance of each personalized regressor for the 6 participants who have annotated all the 60 songs. Specifically, for each of the 6 participants, we train a personalized regressor using his/her annotation of 59 songs, and use the weighted combination of the prediction AV values of the general regressor and personalized regressor to predict the remaining one. This procedure is repeated until each song is used once for testing.

A weight $\omega \in [0,1]$ is introduced to compensate the degree of personalization as shown by the following equation:

$$R_{PMER}(x_i) = (1-\omega)R_{general}(x_i) + \omega R_{personalized}(x_i), \quad (3)$$

where $R_{general}(x_i)$ and $R_{personalized}(x_i)$ denote the prediction result of the ith input instance by the general regressor and personalized regressor. Clearly as ω approaches 1, PMER becomes more fully personalized; when $\omega = 0$, it is namely the general MER.

We evaluate the R^2 statistics with ω = 0, 0.5, and 1. Experimental results tabulated in Table 4 indicate:

1) Compared to the baseline method (ω = 0), personalization generally produces better prediction accuracy for arousal and valence. Significant improvement for valence is observed.
2) Best performance is obtained by setting ω = 0.5, where partial personalization is performed.
3) Although the general regressor achieves a R^2 statistics of 17% for valence and 79.9% for arousal (Section 5.1), its performance drops to 8.8% for valence and 66.7% for arousal on average for predicting the 6 participants. This finding supports the argument that individuality, and personalization, plays an important role to make an MER system effective in practice.
4) The success of personalization (namely the removal of individual difference) validates the conclusion made in the end of Section 5.2.

This experimental result demonstrates the need of personalization. However, one should note that actually the prediction accuracy for valence is not improved a lot. The average prediction accuracy of PMER to a particular user (19.4%) is only slightly better than that of a general MER to general users (17.0% as mentioned in Section 5.1). Moreover, as shown in Fig. 8, the prediction accuracy degrades when the number of pre-annotated songs by the user decreases, especially for valence. For PMER to be more feasible in practice, the number of pre-annotated songs should keep low to relief user burden, and the prediction accuracy for valence should be further enhanced. A reasonable goal is to have the prediction accuracy of valence reaches performance level 2. Personalization is useful, but more advanced works are still needed.

6. CONCLUSIONS

In this paper, the role of individuality in MER is investigated and two approaches are proposed. The first approach GWMER groups users by 15 individual factors and trains regressors for each user group in an effort to reduce individual difference by grouping, whereas the second approach PMER personalizes the MER system by asking users to pre-annotate a number of songs. To our best knowledge, this work represents one of the first attempts to quantitatively evaluate the effect of individuality on MER. The methods described in this paper can be easily applied to other researches related to emotion recognition.

Table 4. The resulting R^2 statistics (in %) for the 6 participants for the PMER approach

Participant		$\omega = 1$	$\omega = 0.5$	$\omega = 0$
1	$R^2(v)$	2.7	4.8	-20.2
	$R^2(a)$	74.2	75.0	68.1
2	$R^2(v)$	26.1	28.4	21.1
	$R^2(a)$	73.9	72.7	68.9
3	$R^2(v)$	22.4	23.5	18.3
	$R^2(a)$	81.5	81.5	78.7
4	$R^2(v)$	7.4	15.6	9.7
	$R^2(a)$	62.8	64.1	57.1
5	$R^2(v)$	22.8	23.7	12.4
	$R^2(a)$	66.9	68.4	65.8
6	$R^2(v)$	21.6	20.5	11.5
	$R^2(a)$	66.0	67.4	61.9
avg	$R^2(v)$	17.2	19.4	8.8
	$R^2(a)$	70.9	71.5	66.7

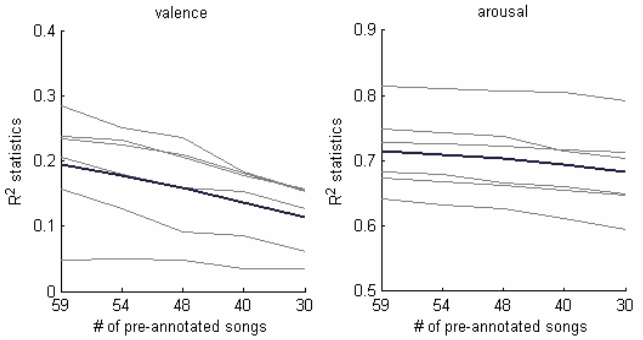

Fig. 8. Degradation of the R^2 statistics as the number of pre-annotated songs by the user decreases. The light line represents the R^2 statistics for each user, and the heavy line represents the average of the 6 users. Left: valence; right: arousal.

Experiments are conducted to evaluate GWMER and PMER. We adopt the regression approach proposed in [12] and use SVR to train regressors. A user interface 'AnnoEmo' is developed for the subjective test. 99 participants are invited to annotate 60 songs, each song being annotated by 40 participants.

Though sex, academic background, music experience and personality have been considered, grouping users by each of these individual factors does not significantly improve the prediction accuracy. This finding implies the subtlety of individuality and suggests the need to explore other methods to describe individuality.

The second experiment demonstrates the effect of personalization. PMER outperforms the baseline method by a great margin, especially for valence prediction. Best performance is achieved by partial personalization. Despite of the improvement made by personalization, the prediction accuracy for valence is sill not as satisfactory as that for arousal. To enhance the prediction accuracy of valence and keep user burden low, more work is needed.

7. ACKNOWLEDGMENT

This work was supported by a grant from the National Science Council of Taiwan under the contract NSC 95-2752-E-002-006-PAE.

8. REFERENCES

[1] A. Jaimes, N. Sebe, and D. Gatica-Perez, "Human-centered computing: A multimedia perspective," *Proc. ACM MM*, pp. 855–864, 2006.

[2] R. W. Picard, *Affective Computing*, the MIT Press, 1997.

[3] S. Reddy and J. Mascia, "Lifetrak: music in tune with your life," *Proc. Human-centered multimedia*, pp. 25–34, 2006.

[4] S. Dornbush, K. Fisher, K. Mckay, A. Prikhodko, and Z. Segall, "XPOD – A human activity and emotion aware mobile music player," *Proc. Int. Conf. Mobile Technology, Applications and Systems*, 2005.

[5] E. Schubert, "Measurement and time series analysis of emotion in music," Ph.D. dissertation, School of Music & Music Education, Univ. New South Wales, Sydney, Australia, 1999.

[6] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," *Proc. ACM SIGIR*, pp. 375–376, 2003.

[7] D. Yang and W. Lee, "Disambiguating music emotion using software agents," *Proc. Int. Conf. Music Information Retrieval*, pp. 52–58, 2004.

[8] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," *Proc. IEEE Int. Conf. Acoustic, Speech, and Signal Processing*, pp. 17–21, 2006.

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

[10] Y.-H. Yang, C.-C Liu, and H.-H. Chen, "Music emotion classification: A fuzzy approach," *Proc. ACM MM*, Santa Barbara, USA, pp. 81–84, 2006.

[11] P. Synak and A, Wieczorkowska, "Some issues on detecting emotions in music," *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Springer, pp. 314–322, 2005.

[12] Y.-H. Yang, Y.-C. Lin, Y.-F. Su and H.-H. Chen, "Music emotion classification: A regression approach," *Proc. IEEE Int. Conf. Multimedia and Expo.*, pp. 208–211, 2007.

[13] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, 2004.

[14] A. Andric and G. Haus, "Automatic playlist generation based on tracking user's listening habits," *Multimedia Tools and Applications*, vol. 29, pp. 127–151, Springer, 2006.

[15] K. Hoashi, K. Matsumoto, and N. Inoue, "Personalization of user profiles for content-based music retrieval based on relevance feedback," *Proc. ACM MM*, pp. 110–119, 2003.

[16] R. E. Thayer, *The Biopsychology of Mood and Arousal*, New York, Oxford University Press, 1989.

[17] C.-M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[18] W. Chai and B. Vercoe, "Using user models in music information retrieval systems", *Proc. Int. Symp. Music Information Retrieval*, 2000.

[19] M. Lesaffre, M. Leman, and J.-P. Martens, "A user-oriented approach to music information retrieval," *Content-Based Retrieval*, Dagstuhl Seminar Proceedings, 2006.

[20] L. R. Goldberg, "The structure of phenotypic personality traits," *American Psychologist*, vol. 48 no. 1, pp. 26–34, 1993.

[21] L. R. Aiken, *Psychological Testing and Assessment*, New York, Allyn & Bacon, 2002.

[22] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," *Proc. ACM SIGIR*, 2007.

[23] http://mpac.ee.ntu.edu.tw/~yihsuan/hcm07/.

[24] D. Cabrera, "PSYSOUND: A computer program for psychoacoustical analysis," *Proc. Australian Acoustic Society Conf.*, pp. 47–54, 1999.

[25] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no.5, pp. 293–302, 2002.

[26] A. Sen and M. Srivastava, *Regression Analysis: Theory, Methods, and Applications*, Springer, 1990.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001.
Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition*, John Wiley & Sons, Inc., 2000.