The PMEmo Dataset for Music Emotion Recognition

Kejun Zhang
State Key Lab of CAD&CG, Zhejiang University
38, Zheda Road
Hangzhou, China 310027
zhangkejun@zju.edu.cn

Hui Zhang
State Key Lab of CAD&CG, Zhejiang University
38, Zheda Road
Hangzhou, China 310027
huiz@zju.edu.cn

Simeng Li
State Key Lab of CAD&CG, Zhejiang University
38, Zheda Road
Hangzhou, China 310027
li.simeng@zju.edu.cn

Changyuan Yang
International User Experience Business Unit, Alibaba Group
Hangzhou, China 310000
changyuan.yangcy@alibaba-inc.com

Lingyun Sun
State Key Lab of CAD&CG, Zhejiang University
Hangzhou, China 310027
sunly@zju.edu.cn

ABSTRACT

Music Emotion Recognition (MER) has recently received considerable attention. To support the MER research which requires large music content libraries, we present the PMEmo[1] dataset containing emotion annotations of 794 songs as well as the simultaneous electrodermal activity (EDA) signals. A Music Emotion Experiment was well-designed for collecting the affective-annotated music corpus of high quality, which recruited 457 subjects.

The dataset is publically available to the research community, which is foremost intended for benchmarking in music emotion retrieval and recognition. To straightforwardly evaluate the methodologies for music affective analysis, it also involves pre-computed audio feature sets. In addition to that, manually selected chorus excerpts (compressed in MP3) of songs are provided to facilitate the development of chorus-related research.

In this article, We describe in detail the resource acquisition, subject selection, experiment design and annotation collection procedures, as well as the dataset content and data reliability analysis. We also illustrate its usage in some simple music emotion recognition tasks which testified the PMEmo dataset's competence for the MER work. Compared to other homogeneous datasets, PMEmo is novel in the organization and management of the recruited annotators, and it is also characterized by its large amount of music with simultaneous physiological signals.

CCS CONCEPTS

• **Applied computing → Sound and music computing**;

KEYWORDS

Dataset; Music Emotion Recognition; Experiment; EDA

ACM Reference format:
Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The PMEmo Dataset for Music Emotion Recognition. In *Proceedings of 2018 International Conference on Multimedia Retrieval, Yokohama, Japan, June 11–14, 2018 (ICMR'18),* 8 pages.
https://doi.org/10.1145/3206025.3206037

1 INTRODUCTION

Emotion plays an important role in music, which means considerable requirements for organizing music by emotional association. Music emotion recognition (MER) is concerned with automatically extracting affective information from music and predicting the perceived emotion of it. MER has became a hot topic in music information retrieval (MIR) and received increased attention in wide-range fields, e.g., Computer Science, Psychology, Human-Computer Interaction.

Most researchers working on MER adopt methods in supervised machine learning to implement music emotion prediction [23], which usually need a large number of songs with emotion labels provided by listeners to train the models. However, it is obviously time-consuming and also labour-consuming to acquire sufficient emotion feedback especially in controlled experimental setting. As a result, there is a scarcity of music datasets containing emotion annotations though a number of widely used datasets do exist for music information research.

To support the development of MER, we present a new dataset and refer it as PMEmo. The dataset contains 794 music clips annotated by 457 subjects, which is created to develop and evaluate MER models. Specially, to eliminate the effects of cultural and educational background [8], we invite subjects from various countries and majors.

Moreover, it is worth concerning that all the music clips in PMEmo are chorus parts of their full songs, manually selected by students majoring in music. This makes the dataset a useful corpus for research on chorus-section detection. Meanwhile, the electrodermal activity (EDA) of subjects when listening to these music are also recorded, making it possible to analyse emotion states in multiple modes.

[1] PMEmo means popular music with emotional annotations

ICMR'18, June 11–14, 2018, Yokohama, Japan

ACM ISBN 978-1-4503-5046-4/18/06…$15.00
https://doi.org/10.1145/3206025.3206037

Table 1: Some Existing Music Datasets with Emotion Annotations

Name	Stimulus	Data	with Audio
Emotify[2]	400 excerpts	induced emotion	yes
MoodSwings[3]	240 excerpts (30s)	arousal and valence	no
Amg1608[4]	1608 excerpts (30s)	valence and arousal	no
emoMusic[5]	744 excerpts (45s)	arousal and valence	yes
DEAM[6]	1802 excerpts	valence and arousal	yes
SoundTracks[7]	360 +110 excerpts	valence, energy, tension and mood	yes
GMD[8]	1400 songs	genre, valence and arousal	downloadable
DEAPDataset[9]	120 music video excerpts	valence, arousal, dominance and physiological data	no
PMEmo	794 music chorus	valence, arousal and physiological data	yes

To briefly summarize, the PMEmo dataset provides:

- song metadata (song title, artists, beginning and ending timestamps of chorus section),
- manually selected music chorus clips (MP3),
- pre-computed audio features for use in MER tasks,
- manually annotated emotion labels: static labels for the whole clips (i.e., overall labels), and dynamic labels for each 0.5-second segment (i.e., continuous labels over time),
- Corresponding EDA physiological signals.

In what follows, we first review related datasets for music emotion recognition (Section 2). Subsequently, we describe the details of the dataset in Section 3, including the song selection, experiment design, the dataset's structure and content, basic data statistics and annotation reliability. In Section 4, we further implement baseline MER method to assess the quality of the dataset empirically and then discuss the results and findings. Finally, we conclude the paper and present some possible extensions (Section 5).

2 RELATED WORK

Since the early years of the MER, there have been numerous efforts to build datasets with emotional annotations facilitating the development and evaluation of music emotion recognition. Table 1 shows some well-known work on that.

Furthermore, the relationship between music and emotion has been studied by psychologists for a long period, who often divide emotion into three categories: expressed emotion, perceived emotion, and felt (or evoked) emotion [22]. As shown in Table 1, a large proportion of datasets in MER focuses on the second one: perceived emotion (i.e., emotions that are perceived by listeners). There are various approaches used to describe perceived emotion and report listeners' responses, most of which belong to one of the following methods: the categorical method and the dimensional method.

For categorical method, emotions are considered as categories that are distinct from each other. Emotify [1], which consists of 400 song excerpts (1 minute long) in 4 genres (rock, classical, pop, electronic), applied the GEMS scale (Geneva Emotional Music Scales) [24] for collecting annotations and accepted the concept of nine basic factors for music-induced emotions (i.e., wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension and sadness).

Although the categorical method is an understandable approach to emotion conceptualization, there exits a major drawback on it. The key to a categorical model is its basic emotion classes, the number of which is usually too small compared to the richness of emotion perceived by humans [22]. Considering this, dimensional models of emotion have gained support among MER researchers.

The two-dimensional, valence-arousal model (VA model) proposed by Russell [12] is one of the most popular dimensional emotion approaches, which is adopted by the major of datasets mentioned in Table 1. The MoodSwings dataset [18] contains per-second VA ratings for 240 clips of US pop songs and the AMG1608 dataset [4] comprises the VA annotations for 1,608 Western music of different genres. Yi-Hsuan et al. developed the emoMusic dataset [17] in 2013, which consists entirely of CC (Creative Commons) music from the Free Music Archive (744 songs). Then in 2016, he aggregated these data into the DEAM dataset [2], which was provided as a baseline dataset in the "Emotion in Music" task at MediaEval Multimedia Evaluation Campaign. Considering the demand for large amount of human labour, these datasets have turned to crowdsourcing through Amazon Mechanical Turk with their own verification steps to remove unreliable data.

Different from the datasets mentioned before, SoundTracks [5] and GMD [11] did not utilize Western music as emotion stimuli. the SoundTracks dataset selected film soundtracks (approx. 15 second) as stimulus sets while the GMD datasets collected 1400 Greek tracks to induce emotion.

In addition to the subjective label annotated by listeners, recent advances in emotion recognition (e.g., Chen [3] and Gao [7]'s recent work of emotion recognition from electroencephalogram (EEG) signals) have motivated the creation of novel datasets containing emotional expressions in different modalities, including physiological responses which could represent the felt emotion of listeners. the DEAPDataset [10] recorded the emotion rating and corresponding EEG signals from 32 volunteers, among whom 22

[2]http://www.projects.science.uu.nl/memotion/emotifydata/
[3]http://music.ece.drexel.edu/research/emotion/moodswingsturk
[4]https://amg1608.blogspot.ca/
[5]http://cvml.unige.ch/databases/emoMusic/
[6]http://cvml.unige.ch/databases/DEAM/
[7]https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/past-projects/coe/materials/emotion/soundtracks/Index
[8]https://hilab.di.ionio.gr/en/music-information-research/
[9]http://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html

participants' frontal face videos were also recorded. The PMEmo dataset provides VA annotations for 794 music choruses, each of which was labelled by at least 10 people, and all the EDA records of 457 subjects are also included.

3 THE PMEMO DATASET

The whole PMEmo dataset of approximately 1.3 GB can be downloaded from www.happiness.zju.edu.cn. Considering copyright restrictions, full songs are not included in this dataset but the chorus excerpts are available in PMEmo. To make the data accessible and compatible for a wide range of platforms, the metadata of songs, the acoustic features and the annotations are all stored in CSV files (delimited by comma).

In this section, we outline the song acquisition and chorus selection procedure, describe the information of subjects who participant in this work, introduce the experiment designed to collect annotation, list in detail the dataset's components and analyze basic statistical properties of the dataset, which could be used to evaluate annotation reliability.

3.1 Song Acquisition and Chorus Selection

Our target music is pop song, therefore, we search some well-known music charts for gathering songs popular all around the world: the Billboard Hot 100[10], the iTunes Top 100 Songs (USA)[11] and the UK Top 40 Singles Chart[12]. In order to fetch sufficient music resources, we downloaded the songs available on these three charts from 2016 to 2017 and collected 1000 songs all together. However, we later discovered a set of duplicates (though they do have slightly different song titles or artists from the original ones). After filtering out the reduplicative music, we obtain a full song set of 794 pop songs. Table 2 illustrates some basic statistical data of the music sources.

Table 2: The Sources of Music

Charts	Start Time	End Time	Songs
Billboard Hot 100	19^{th} week, 2016	23^{rd} week, 2017	487
iTunes Top 100	15^{th} week, 2016	21^{st} week, 2017	616
UK Top 40 Singles	37^{th} week, 2016	21^{st} week, 2017	226

To reduce the heavy cognitive load and improve the stability of emotion annotation, almost every dataset in MER utilize music segments as emotion-inducing materials and the PMEmo does this as well. Specifically, each clip in this dataset is manually selected as one of the chorus parts of each song, which is implemented by university students in music major. But what distinguishes the PMEmo from most of the other MER datasets is that the clips are of various length which are exact the duration of the chorus parts. This enables the PMEmo dataset useful in chorus detection research.

3.2 Subject Selection

A total of 457 subjects (236 females and 221 males) are recruited to participant in this work. Among them, 366 are Chinese university students who are in non-music major while 44 are majoring in music recruited to ensure high quality labelling. Meanwhile, to weaken the impact of cultural background (almost all the songs are in English), 47 English speakers are invited to annotate the datasets. Each song receives a total of at least 10 emotion annotations including one by music-majoring student and one by English speaker. Since the limitation of EDA signal acquisition device, we are unable to utilize crowdsourcing method, which means subjects can not annotate the dataset directly via the Internet. Considering this, we design an experiment to acquire the data we want and describe that in the following.

3.3 Experiment Design and Annotation Collection

We use MP150 Biopac system (Biopac Systems, Inc., Goleta, CA) to monitor the the electrodermal activity continuously at a sampling rate of 50 Hz and export the signals from AcqKnowledge software. A desktop application was developed for annotating the songs (shown in Figure 1). The annotation was done with the sliding area collecting dynamic annotations, on a scale from 1 to 9, at a sampling rate of 2 Hz and annotators should make a static annotation for the whole music excerpt on nine-point scale after finishing dynamic labelling. Furthermore, the annotators were asked to listen to the same music twice to annotate on valence and arousal separately.

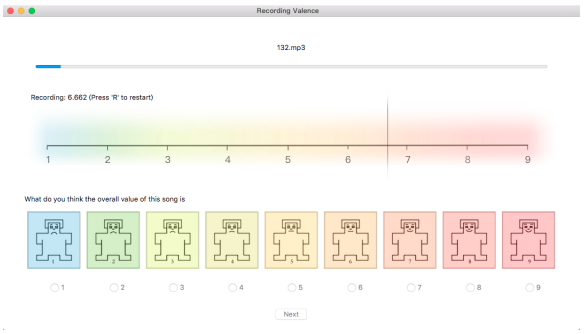

Figure 1: The Annotation Interface

Firgure 2 is a flow diagram of this experiment, for which each subject spent 50 minutes on average. At the very beginning, subjects were shown a document used to inform consent of the experiment, by which it was ensured that all the participants were aware of the purpose, potential risk and other necessary information about the experiment. After signing that paper, we displayed a video to the subjects and introduce the procedure and details in the annotation task to them. We also had a training session conducted to ensure their precise understanding of this task, during which process 4 clips representing extreme emotion (extremely large and small value in valence and arousal respectively) were played and subjects were enabled to get familiar with operating the annotation interface. Only after completely trained can subjects start the formal experiment.

It is a known issue that multiple cognitive processes will interact on each other, in other words, subjects need some buffer time to relax and remove the interference from the previous stimulus. Therefore, we set a relaxing procedure before starting annotating

[10]https://y.qq.com/n/yqq/toplist/108.html
[11]https://y.qq.com/n/yqq/toplist/123.html
[12]https://y.qq.com/n/yqq/toplist/107.html

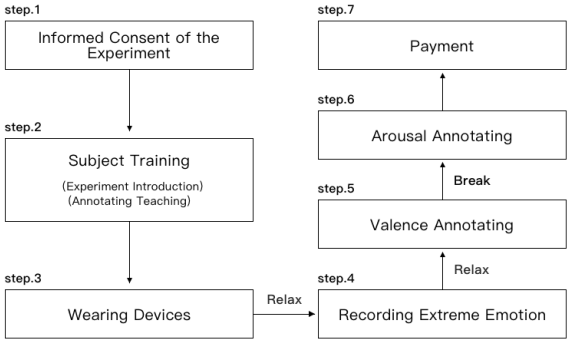

Figure 2: Experimental Procedures

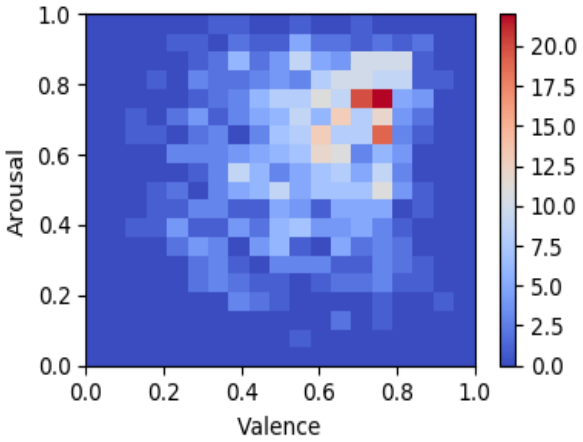

Figure 3: The Distribution of Static Annotations

and a short break (10 seconds) between each song-listening. At the relaxing procedure, subjects were required to calm down with the help of *In My Life*, a song by Kevin Kern, whose light music were proved to be in favour of relieving people's emotion [9]. EDA finger transducer (BSL-SS3LA) were equipped to subjects' fingers of the non-dominant hand before this step so that we can check the subjects' response with assistance of the skin conductance recorded by the device. Particularly, subjects' EDA responses to 4 clips with extreme emotion were also stored in order to study individual difference in the subsequent research work.

After finishing all these preparation procedures, subjects were asked to conduct annotation work. Specifically, each subject listened to 20 excerpts and one of those was duplicated to guarantee the high quality data as what was done in [4] (the annotations from this subject were accepted only if the bias between duplicated clips were within 0.25 in the Valence-Arousal space), but we did not inform the subject of the existence of identical songs. Completing the experiment, subjects were paid 50 yuan per person.

In total, 457 subjects have participated in this experiment and we finally received 401 person-time valid annotations (87.7 percent). Each music clip was annotated by at least 10 annotators including English speakers and semi-experts from music academy. As we can see from Figure 3 that a large proportion of music chorus clips have overall annotations falling in the first quadrant. Moreover, Figure 4 enumerates several emotion changing curves of various music excerpts (5 randomly chosen clips from the PMEmo dataset), which illustrates the significant difference of emotion between songs and comparatively stable tendency for the same song.

3.4 Data Reliability

Initial orientation time (IOT) is a known concept in MER that annotators need some preliminary time before they can give meaningful and reliable annotations. Schubert [14] found that the median IOT for valence was 8 seconds and for arousal that was 12 seconds. Yi-Hsuan's work in 2013 [16] showed that the annotations began converging around the 10^{th} second and after that, he obtained a similar result in 2017 [2]. Taking into account the previous work, we discard the first 15-second dynamic annotations from the data.

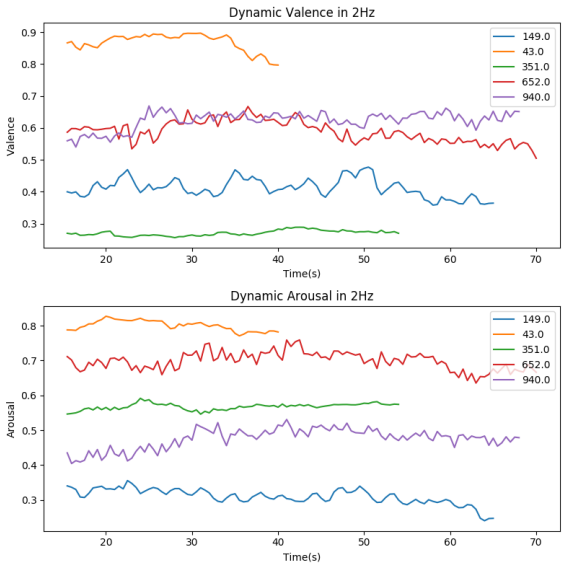

Figure 4: Emotion curves of song NO. 43, 149, 351, 652 and 940 (the first 15-second labels are removed).

We use the Cronbach's α [13] to evaluate annotation consistency. Cronbach's α is often used in psychometric test to estimate the reliability that represents the degree to which a set of items measures a single unidimensional latent construct. In this article, we compute the Cronbach's α on the sequences of annotations for each song.

Annotations are sampled in 2Hz and as what is usually done to normalize the continuous annotations [2], we process the annotations by

$$a_{j,i} = a_{j,i} + (\overline{A_j} - \overline{A}) \quad (1)$$

where $a_{j,i}$ is the label annotated by subject j at time i, and $\overline{A}$ is the mean of all the labels for this song by all subjects while $\overline{A_j}$ is the mean of dynamic labels by subject j.

Table 3 shows the mean (averaged across songs) and the standard deviation of Cronbach's α for the annotations in the PMEmo dataset. The quite high results for both valence and arousal indirectly indicate excellent internal consistency of the collected annotations.

Table 3: Cronbach's α

Dimension	The Mean Value	The Standard Deviation
Valence	0.998	0.005
Arousal	0.998	0.008

4 MUSIC EMOTION RECOGNITION EXPERIMENTS USING PMEMO

It is necessary to have a large-scale dataset for music emotion recognition, which has become an important task recently. While the PMEmo dataset is not restricted to this task, we illustrate its use for implementing and evaluating a music emotion recognition system that aims to predict the static and dynamic emotion of a song automately. In the following, we firstly introduce the features and methodologies used in the assessment. Then we present the setting and performance in each experiment.

4.1 MER Feature Set

Feature extraction is a critical step for MER model, which could affect the recognition performance to a large extent. However, for generic MER there has been no attempt made at defining a "standard" feature set. This is probably resulted from the rather young and multi-faceted characteristic of music emotion computing field. What is fortunate to researchers in MER is that the INTERSPEECH 2013 Computational Paralinguistics Evaluation (ComParE) Campaign [15] has proposed a well-evolved feature set for detecting paralinguistic phenomena automately. In this set, the key factors are frame-wise acoustic low-level descriptors (LLDs, briefly shown in Table 4) including MFCCs, energy, etc. These descriptors are very relevant for both Music Information Retrieval and other general sound analysis. The ComParE 2013 has also applied quite a number of statistical functions, such as mean and moments, to LLD contours and the calculated results constitute a large set of suprasegmental features. All of these obtained features form the ComParE 2013 baseline feature set with a 6373-dimension scale, which has been exhaustively listed at Weninger's work [20] and proved to work effectively for the task of music mood regression.

In particular, PMEmo dataset provides all the 6373-dimension features in song level for the sake of static emotion task. In comparison to the considerable size for overall mood recognition, we only extract the core 260-dimension features in segment level (the average and the standard deviation of the 65 LLDs as well as their first-order derivatives, calculated in 1s window with 0.5s overlap) for dynamic recognition task to properly reduce the computing load. Extraction of these features is done with the open-source toolkit openSMILE [6]. Furthermore, no feature selection procedure is implemented in order to fully demonstrate PMEmo's capability to MER task.

Table 4: A Simple Profile for Low-Level Descriptors Provided in ComParE Acoustic Feature Set

Category	Dimension	Content
Energy related	4	Auditory weighted frequency bands and their sum
Spectral	55	Centroid, roll-of point, skewness, sharpness, and spectral flux
Voicing related	6	fundamental frequency and harmonics-to-noise ratio

4.2 MER Methodology

In this article, we adopt two well-accepted methods, Multivariate Linear Regression (MLR) and Support Vector Regression (SVR), as the base classifiers to model emotions in valence and arousal.

For the whole song, we train and test the classifiers using the 6373-dimension features $x_1, x_2, ..., x_{6373}$ and separate static labels of valence and arousal $y_{valence}, y_{arousal}$ respectively.

$$\{X_1, X_2, ..., X_m\} \overset{static\ model}{\longrightarrow} \{e_1, e_2, ..., e_m\} \tag{2}$$

where m is the number of songs, $X_i = \{x_1, x_2, ..., x_6373\}$ is the feature set of the ith song and e_i is the value of valence or arousal for this song.

With respect to continuous mood of a song, it is natural to consider the multi-scale structure in music. Xianyu et al. proposed a double-scale SVR [21], which decoupled dynamic emotion into two scales and then recognized them separately. We are inspired by the outstanding idea and organize the dynamic MER model in a similar way: building a global classifier with 6373 song-level global features to fit the mean value of dynamic emotion, meanwhile utilizing the 260 segment-level local features to reveal the fluctuation over time.

$$\begin{aligned} &\{X_1, X_2, ..., X_m\} \overset{global\ model}{\longrightarrow} \{\overline{L_1}, \overline{L_2}, ..., \overline{L_m}\} \\ &\{Y_1^{t_1}, Y_2^{t_2}, ..., Y_m^{t_3}\} \overset{local\ model}{\longrightarrow} \{D_1^{t_1}, D_2^{t_2}, ..., D_m^{t_m}\} \\ &\text{Dynamic emotion} : L_i = \overline{L_i} + D_i^{t_i} \end{aligned} \tag{3}$$

where m is the number of songs and for the ith song, X_i is the global feature set of it while $Y_i^{t_i}$ is a matrix of 260 columns and t_i rows (t_i is the number of timestamps in the ith song), which represents the local features of this song. $\overline{L_i}$ is the mean of dynamic emotion and $D_i^{t_i}$ is the fluctuation at each timestamp.

Before building the regression models, we resized all the annotations (both of static annotations and dynamic ones) into $[0, 1]$ for a more intuitive understanding of the performance in the following experiments.

4.3 Performance of Static MER

Static task is to predict the overall emotion of a whole song, which is represented by single valence value and arousal value. To train and test the static MER regression models, the data in PMEmo was divided into 11 folds. 10 folds of that constituted the training set while the remaining one was used to test the trained models. A

10-fold cross-validation procedure was carried out on the training set for parameter optimization. In this article, Root Mean Square Error (*RMSE*) and Pearson Correlation Coefficient (r) were calculated individually for valence and arousal as the evaluation metrics. Table 5 presents 2 running results on static emotion.

- Run 1: Based on MLR, trained on the 10 training folds;
- Run 2: Based on SVR with Radial Basis kernel function, $C = 1.0$, trained on the 10 training folds.

Table 5: Evaluation Results on Static Emotion

Dimension	Run	*RMSE*	r
Valence	Run 1	0.136	0.546
	Run 2	0.124	0.638
Arousal	Run 1	0.111	0.719
	Run 2	0.102	0.764

In this evaluation procedure, Run 1 (MLR) and Run 2 (SVR) had similar results, both of which achieved good performance and revealed the bright prospect of the PMEmo dataset in the field of static MER.

4.4 Performance of Dynamic MER

As mentioned before, a hierarchical regression model aiming to recognize the global trend as well as local variation was built. It is worth notice that the lengths of music clips are not fixed in the PMEmo dataset, and resulted from that, for global scale and local scale we did the following operation respectively:

- **Global-scale operation** For each song, extracting only one global feature and mapping it into one global emotion value.
- **Local-scale operation** For each song, dividing it into 1-second segments with 50% overlap (leading to 2 Hz fluctuation frequency), then extracting the local features from these fragments and project them onto mood space.

The strategy of splitting pre-existing data into training and testing set here is the same as the previous experiment. However, there is slight distinction existing in the evaluation calculation due to the difference of predicted music units. In global scale, *RMSE* and r were calculated in the whole testing set (in song unit). While in local scale, these metrics were computed for each song and then averaged across songs. Table 6 reports the evaluation metrics for dynamic MER.

The *RMSE* results of dynamic MER also performed well. Furthermore, it can be seen from Table 6 that the major error arose in global scale (much larger than the local-scale error), presenting consistency with the result of [21]. However, the r results was not desirable in local scale, which obviously limited the dynamic emotion recognition performance. This may result from the difficulty of instantaneous perception. In other words, annotators could not perceive and report emotion changes well in very short time, affecting the description of emotion fluctuation in real time. Altogether, these two experiments indicated the advantage of PMEmo in terms of establishing MER system.

Table 6: Evaluation Results on Dynamic Emotion

Dimension	Run	Scale	*RMSE*	r
Valence	Run 1	global	0.103	0.673
		local	0.016	0.047
		fusion	0.089	0.047
	Run 2	global	0.106	0.675
		local	0.016	0.095
		fusion	0.088	0.095
Arousal	Run 1	global	0.113	0.816
		local	0.020	0.103
		fusion	0.093	0.103
	Run 2	global	0.101	0.844
		local	0.019	0.115
		fusion	0.085	0.115

4.5 Predict Music Emotion using EDA

Considering the requirements of multi-modal information research on MER, the PMEmo dataset has recorded the electrodermal activity of subjects simultaneously when listening to music. In this part, we demonstrate one of the potential approaches to taking advantage of these EDA signals and detecting corresponding music mood. Specifically, we only performed dynamic MER arbitrarily as both dynamic annotations and EDA vary over time.

Firstly, a low-pass filter of 0.6 Hz was employed to diminish the noise from motion and artifacts [19]. Then the filtered skin electric conductance data was scaled in z-scores (z-score = $\frac{X-\mu}{\sigma}$, where μ is the mean of vector X and σ is the standard variation of it) due to the great individual distinctions, by which we could obtain the variation of EDA as the music emotion varying. Because of the different acquisition rates of EDA and continuous emotion (50 Hz and 2 Hz, respectively), we resampled the EDA signals into 2 Hz. Finally, we trained and tested MLR and SVR models with pre-processed EDA data as well as the dynamic affective value.

Table 7: Evaluation Results using EDA Data

Dimension	Run	Scale	*RMSE*	r
Valence	Run 1	global	0.139	0.063
		local	0.016	0.060
		fusion	0.118	0.059
	Run 2	global	0.141	0.017
		local	0.016	0.059
		fusion	0.119	0.048
Arousal	Run 1	global	0.186	0.011
		local	0.019	0.097
		fusion	0.157	0.106
	Run 2	global	0.194	0.040
		local	0.019	0.099
		fusion	0.160	0.110

Table 7 shows the prediction performance using EDA data, which represented the feasibility of indicating music emotion from listeners' physiological data. The comparatively low value of *RMSE* revealed the correlation between perceived emotion and evoked emotion. The PMEmo dataset enables researchers to study MER in other respects besides acoustic feature space.

5 CONCLUSION AND FUTURE WORK

We have created a brand new dataset, PMEmo, that enables music retrieval and emotion recognition research in multiple modalities. It has been released to public and are available for free download. The PMEmo dataset contains information on the level of songs, manually selected chorus excerpts and their features in common MER use. In addition to the standard valence-arousal annotations appearing in other homogeneous datasets as well, PMEmo is characterized by its large amount of music with simultaneous physiological signals, and it is also novel in the components of annotators recruited. We set several procedures to ensure the quality of labelling and all the annotations as well as physiological data were acquired under the strictly controlled laboratory conditions.

Despite the good performance of the PMEmo dataset at music retrieval and emotion recognition tasks in multiple modalities, we would like to consider further extensions in the future. In particular, we are thinking about enriching the cultural background of songs since it does affect the perception of valence to large extent. Beyond that, we intend to add text-based features that allow analysis of lyrics, which is also an important composition of music content but ignored in our current work. Next to the music descriptors, abundant categories of physiological information are required to distinguish listeners' emotion state more accurately. Finally, other popular emotion dimensions such as dominance may be included to increase the comprehensive ability of the dataset in various applications.

To support these further research work in the future, we have developed a crowdsourcing platform customized for annotating multimedia content, especially for music emotion labelling. Researchers are allowed to use multimedia resources on this platform and release their experiments easily. It is a practical network application which can help experimenters quickly gather and assess multimedia affective data. We will publish this platform soon and believe it can actually bring convenience to Affective Computing research community.

ACKNOWLEDGMENTS

This project is supported by the National Key Research and Development Program of China (2016YFC0200700), National Natural Science Foundation of China (No. 61672451 and No. 61402141), National Basic Research Program of China (973 Program, No. 2015CB352503), and Alibaba-Zhejiang University Joint Institute of Frontier Technologies. It would not have been possible without the help of numerous students, in particular, Lumin Huang, Mengqi Jin, Ke Duan, Yingping Cao, Jun Zheng, Xiaoyi Huang, and Kaili Zhu from Happy Lab, Zhejiang University. Finally, we would like to thank Prof. Li for providing an experimental research facility, which ensured the smooth progress of the experiment.

REFERENCES

[1] Anna Aljanaki, Frans Wiering, and Remco C Veltkamp. 2016. Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management* 52, 1 (2016), 115–128.

[2] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PloS one* 12, 3 (2017), e0173392.

[3] Tanfang Chen, Shangfei Wang, Zhen Gao, and Chongliang Wu. 2016. Emotion Recognition from EEG Signals Enhanced by User's Profile. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 277–280.

[4] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. 2015. The AMG1608 dataset for music emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 693–697.

[5] Tuomas Eerola and Jonna K Vuoskoski. 2011. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39, 1 (2011), 18–49.

[6] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[7] Zhen Gao and Shangfei Wang. 2015. Emotion recognition from eeg signals using hierarchical bayesian network with privileged information. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 579–582.

[8] Xiao Hu and Yi-Hsuan Yang. 2017. Cross-dataset and cross-cultural music mood prediction: A case on Western and Chinese Pop songs. *IEEE Transactions on Affective Computing* 8, 2 (2017), 228–240.

[9] Chia-Hui Ko, Yi-Yu Chen, Kuan-Ta Wu, Shu-Chi Wang, Jeng-Fu Yang, Yu-Yin Lin, Chia-I Lin, Hsiang-Ju Kuo, Chia-Yen Dai, and Meng-Hsuan Hsieh. 2017. Effect of music on level of anxiety in patients undergoing colonoscopy without sedation. *Journal of the Chinese Medical Association* 80, 3 (2017), 154–160.

[10] Sander Koelstra, C Muhl, M Soleymani, JS Lee, A Yazdani, T Ebrahimi, T Pun, A Nijholt, and I Patras. [n. d.]. DEAP dataset (2012). *URL http://www. eecs. qmul. ac. uk/mmv/datasets/deap/index. html. BibliographyA* 6 ([n. d.]).

[11] Dimos Makris, Ioannis Karydis, and Spyros Sioutas. 2015. The greek music dataset. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*. ACM, 22.

[12] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[13] J Reynaldo A Santos. 1999. Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of extension* 37, 2 (1999), 1–5.

[14] Emery Schubert. 2013. Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. *Psychology of music* 41, 3 (2013), 350–371.

[15] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

[16] Mohammad Soleymani, Anna Aljanaki, Yi-Hsuan Yang, Michael N Caro, Florian Eyben, Konstantin Markov, Björn W Schuller, Remco Veltkamp, Felix Weninger, and Frans Wiering. 2014. Emotional analysis of music: A comparison of methods. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1161–1164.

[17] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 2013. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 1–6.

[18] Jacquelin A Speck, Erik M Schmidt, Brandon G Morton, and Youngmoo E Kim. 2011. A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation.. In *ISMIR*. 549–554.

[19] Iris Van Den Bosch, Valorie N Salimpoor, and Robert J Zatorre. 2013. Familiarity mediates the relationship between emotional arousal and pleasure during music listening. *Frontiers in human neuroscience* 7 (2013).

[20] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer. 2013. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology* 4 (2013).

[21] Haishu Xianyu, Xinxing Li, Wenxiao Chen, Fanhang Meng, Jiashen Tian, Mingxing Xu, and Lianhong Cai. 2016. SVR based double-scale regression for dynamic emotion prediction in music. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 549–553.

[22] Yi-Hsuan Yang and Homer H. Chen. 2011. *Music Emotion Recognition* (1st ed.). CRC Press, Inc., Boca Raton, FL, USA.

[23] Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 40.

[24] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494.