THE AMG1608 DATASET FOR MUSIC EMOTION RECOGNITION

Yu-An Chen,[1] Yi-Hsuan Yang,[2] Ju-Chiang Wang,[2] and Homer Chen[1]

[1] National Taiwan University, Taiwan
[2] Academia Sinica, Taiwan
Emails: b96901042@ntu.edu.tw, {yang, asriver}@iis.sinica.edu.tw, homer@ntu.edu.tw

ABSTRACT

Automated recognition of musical emotion from audio signals has received considerable attention recently. To construct an accurate model for music emotion prediction, the emotion-annotated music corpus has to be of high quality. It is desirable to have a large number of songs annotated by numerous subjects to characterize the general emotional response to a song. Due to the need for personalization of the music emotion prediction model to address the subjective nature of emotion perception, it is also important to have a large number of annotations per subject for training and evaluating a personalization method. In this paper, we discuss the deficiency of existing datasets and present a new one. The new dataset, which is publically available to the research community, is composed of 1608 30-second music clips annotated by 665 subjects. Furthermore, 46 subjects annotated more than 150 songs, making this dataset the largest of its kind to date.

Index Terms—Music emotion recognition, personalization, crowdsourcing

1. INTRODUCTION

Music emotion recognition (MER) is concerned with the development of computational models to predict the perceived emotion of music [1], [2]. It has recently received considerable attention, as it has the potential to break new ground for intelligent, human-centric access of music content [3]. Most existing work on MER adopts supervised machine learning to train models for music emotion prediction using emotion labels of a number of songs provided by listeners [1].

To develop and evaluate MER models, a dataset containing sufficient emotion labels is required. Specifically, the dataset should be large in the number of songs, subjects, and annotations per subject. A large number of emotion annotations of songs are often required for machine learning to produce accurate emotion prediction models [4]. Furthermore, each song needs to be annotated by a large number of subjects so that the general emotional responses to the song can be obtained [5], [6]. Finally, a large number of annotations per subject are needed for developing personalization methods to address subjectivity issue of emotion perception and to improve the performance of a practical MER system [7].

TABLE I
STATISTICS OF MUSIC EMOTION RECOGNITION DATASETS

Name	#songs	#subjects	#annotations per Song	#annotations per Subject
MER60 [7]	60	99	40	15–60
DEAP [8]	120	32	14–48	40
MoodSwings [9]	240	546	7–23	1–8
AMG1608	1608	665	15–32	12–924

However, as the first three rows of Table I show, existing MER datasets rarely meet these requirements. Each song in the MER60 dataset [7] was annotated by 40 subjects, but only ten subjects annotated more than 30 songs. The DEAP dataset [8] has more annotations per subject, but the number of songs and subjects are limited. The MoodSwings dataset [9] has more subjects, but again it has small number of annotations per subject. Emotion annotation is known to be time consuming and labor intensive; however, the deficiency of dataset may hinder the progress of MER [7].

We present a new dataset and refer to it as AMG1608. The dataset contains 1608 30-second music clips annotated by 665 subjects. It is created to facilitate the development and evaluation of MER. This dataset includes the annotations of 46 subjects, and each of them annotated more than 150 music clips, making the dataset a useful corpus for research on MER personalization. As shown in Table I, the dataset is larger than the previous ones in many measures. It is made publically accessible to the research community (check out the blog of the dataset http://goo.gl/8vnMCH).

A number of web resources are utilized to construct the dataset. To collect the emotion annotations for this large dataset, we follow the footstep of previous work [9], [11] and exploit the online crowdsourcing service Amazon Mechanical Turk (AMT) [10]. We also obtain the mood category information from the online music guide service website All Music Guide (AMG) to make the emotion distribution of the songs well-balanced. The 30-second clips are audio previews provided by the music service company 7digital.

In what follows, we provide in Section 2 the details of the dataset, including the song selection, subject selection,

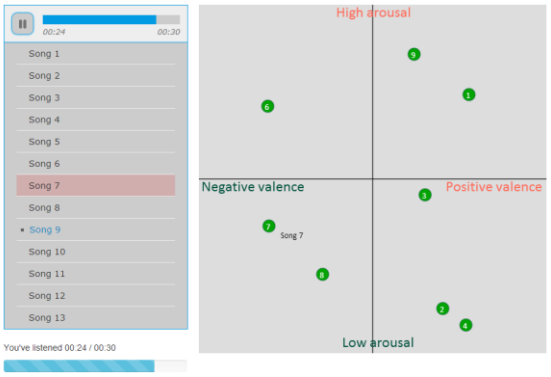


Fig. 1. The graphical interface used in the emotion annotation tasks. The square panel on the right represents the VA plane with valence and arousal as the horizontal and vertical axes, respectively. Upon a subject clicks on the panel and labels the VA value of a song, a green dot showing the song ID is displayed at the specified VA coordinates. The subject can drag and drop the green dots to edit the annotations. The panel on the left shows the song list and provides a playback interface.

and annotation processes. Then, in Section 3 we empirically assess the quality of the dataset through a baseline MER method and a baseline personalization method. Finally, we conclude the paper in Section 4.

2. DATA COLLECTION

2.1. Emotion representation

The emotion representation in an MER model can be either categorical or dimensional. The categorical approach describes the music emotion using a set of discrete affective terms such as exciting, cheerful, calm, bittersweet, and sad [12]. In contrast, the dimensional approach represents the emotion in the continuous space using, for example, valence and arousal (VA) as coordinates [13]. Here, valence represents how pleasant the emotion is, and arousal represents the degree of excitement of the emotion. In either approach, the emotion representation can be static (e.g. the entire song is assigned with one emotion label) or dynamic (e.g. the output is a second-by-second emotion trajectory) [5], [14]. We opt to use the dimensional emotion representation and ask subjects to annotate perceived emotion of an entire song as a point in the VA space.

2.2. Song selection

We obtain a list of contemporary Western music from AMG, which has 34 distinct mood categories defined by music editors.[1] Then, a 30-second audio preview of each song is downloaded from 7digital. Besides ensuring sufficient songs in the dataset, it is preferable to have the songs uniformly distributed in the emotion space. Therefore, we use the VA value of each mood category generated by the tag2VA algorithm [15] as the initial value of each song. This marks every song in the VA space. Next, a number of songs are randomly selected from each quadrant of the VA space. The selection is made such that the songs are evenly distributed in the four quadrants. In the selection process, a song is discarded if it is associated with multiple categories. Finally, 1608 pieces of songs with sufficiently diverse affective contents are selected for annotation.

TABLE II
LIST OF DATA COLLECTED

Category	Item
Demographic data	Gender and age.
Music experience	Years of music education and frequency of music listening.
Music preference	Five-point rating for the preference of each music genre: rock, metal, country, jazz, pop, classical, blues, hip-hop, electronic, reggae, dance.

2.3. Subject selection

A total of 665 subjects are recruited to annotate the dataset via the Internet. Among them, 22 are recruited from the campus and 643 from the AMT. To ensure the AMT subjects come from the same country and provide high quality annotations, they are required to be residents of the United States and have over 90% tasks done on the AMT been accepted. The 22 subjects from the campus are asked to annotate a 240-song subset of AMG1608. The songs annotated by the AMT subjects are part of the entire dataset, and the actual number of songs annotated by each subject varies. Each song receives a total of 15 emotion annotations from the AMT subjects.

2.4. Annotation collection

For annotation quality, we divide the 1608 songs into 134 annotation tasks (referred to as human intelligence tasks, or HITs, in AMT). A subject is asked to annotate 13 songs in each HIT. The 13 songs in each HIT are generated as follows. First, three songs are randomly selected from each quadrant. Then, one of the 12 songs is duplicated and used to assess the annotation quality. A HIT is abandoned if the VA values of the two identical songs are not close enough.

At the beginning of each HIT, a subject is given the definition of VA representation of emotion and instructions of how to annotate emotion label. Moreover, the subject is informed of the existence of identical songs, but not which songs they are. Next, the subject is provided with a square interface panel, as shown in Fig. 1, on a PC monitor to interact with. The value of each dimension of the panel ranges from

[1] http://www.allmusic.com/moods

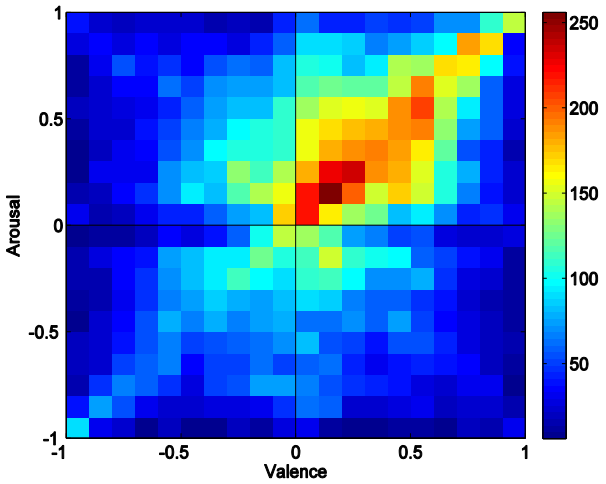


Fig. 2. Histogram of the emotion annotations of the 1608 songs.

−1 to 1. The subject can annotate a song by placing the cursor on the panel to indicate the location of the perceived VA value of the song. In addition to the music emotion annotations, we also collect demographic information, music experience and music preference of the subjects through a questionnaire. Table II lists the information we have collected.

2.5. Dataset analysis

Among the 665 subjects, 345 are male and 320 are female. The average age of them is 32.0±11.4. On average, each subject spends 455 seconds to complete one HIT. To measure the inter-subject agreement of the annotations for each song, we calculate the Krippendorff's α [16]. The average Krippendorff's α across the songs is 0.31 for valence and 0.46 for arousal, which are both in the range of fair agreement. Finally, one may see from Fig. 2 that a large portion of the annotations falls in the first quadrant.

3. EXPERIMENTS

We show the necessity of having a large-scale dataset for MER and for personalization through two experiments. In the first experiment, we evaluate the performance of an MER model trained with various numbers of training data to demonstrate that sufficient training data lead to an accurate MER model. In the second experiment, we personalize an MER model for each of the selected subjects using annotations from the subject to show that the personalization improves the accuracy of MER. We briefly introduce the MER model and the feature sets below. Then, we present the setting and the results of each experiment in Sections 3.1 and 3.2.

The acoustic emotion Gaussian (AEG) approach is adopted for MER modeling as it has been shown to be effective in previous work [17]. This approach uses a Gaussian mixture model for MER modeling and personalizes it using personal annotations from the user [17], [18]. In particular, we adopt the maximum a posteriori linear regression method for personalization [19]. Two toolboxes, MIRtoolbox [20] and YAAFE [21], are used to extract the acoustic features listed in Table III from the audio signals. All features are concatenated to form a long feature vector as previous work shows this technique improves accuracy of MER [2], [22].

TABLE III
LIST OF ACOUSTIC FEATURES

Feature	#	Description
MFCC	40	20 Mel-frequency cepstral coefficients and their first-order time differences [24].
Tonal	17	Octave band signal intensity using a triangular octave filter bank and the ratio of these intensity values [21].
Spectral	11	Linear predictor coefficients that capture the spectral envelope of the audio signal [25]. Spectral flux [21] and spectral shape descriptors [26].
Temporal	4	Shape statistics (centroid, spread, skewness, and kurtosis) of the audio signal [27].

3.1. Performance of general MER

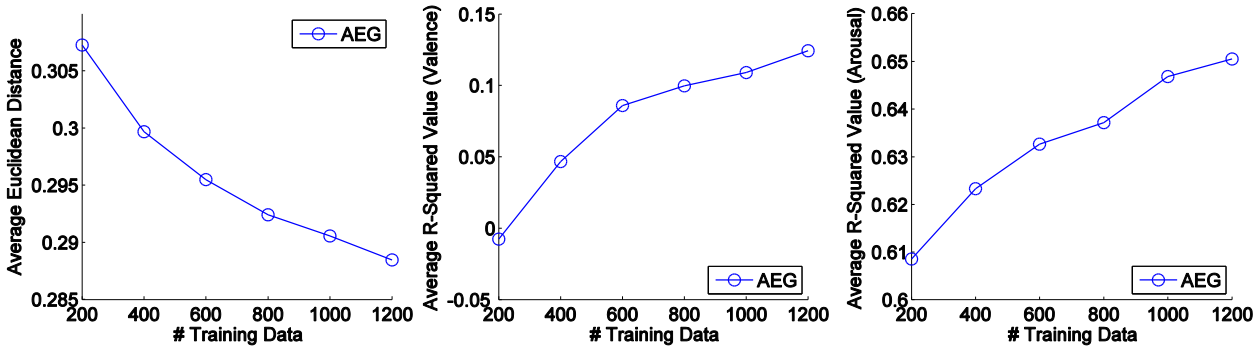
To show the advantage of a large dataset, we train an MER model using different amounts of training data and evaluate its performance using a set of testing data. We use 200, 400, 600, 800, 1000, and 1200 songs for training and 400 songs for testing. The training data and testing data do not overlap with each other. We repeat this process ten times and randomly select the training and the testing data in each time.

We measure the accuracy of MER using Euclidean distance and the R^2 statistics [23]. The former measures the prediction error, with lower value indicating better performance. In contrast, R^2 indicates the fitness of a regression model, so a high R^2 value is preferred. Moreover, R^2 is computed separately for valence and arousal.

Fig. 3 shows the performance of the MER model constructed with different numbers of training data. Two observations can be made from the results. First, the performance of the MER model is comparable to the one reported in previous work [11], [22], showing the quality of this dataset. Second, it is clear that the MER model becomes more accurate as the number of training data grows, demonstrating that a large MER dataset is desirable to develop accurate MER models.

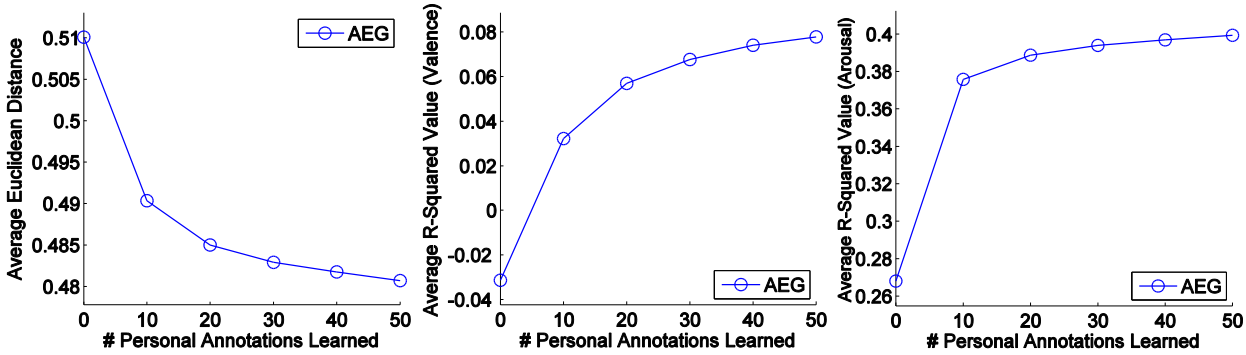
3.2. Personalization of MER

We select 46 users that annotated more than 150 songs as test users for the evaluation of MER personalization, which is carried out in an online fashion. For each test user, a general MER model is trained with 600 randomly selected songs that are not annotated by the test user under consideration. Then, the general model is incrementally personalized five times using songs annotated by the user. The songs are expanded



(a) (b) (c)

Fig. 3. The performance of MER model trained with various number of data, in terms of (a) the average Euclidean distance, (b) the R^2 for valence prediction, and (c) the R^2 for arousal prediction.



(a) (b) (c)

Fig. 4. The performance of personalized MER models, in terms of (a) the average Euclidean distance, (b) the R^2 for valence prediction, and (c) the R^2 for arousal prediction.

by 10 each time. That is, the number of songs is 10, 20, 30, 40, and 50 in temporal order, and the preceding songs are a subset of the current songs each time. The accuracy of the personalized model is evaluated using another 100 songs annotated by the user. The same 100 songs are used for evaluating the performance of the general model and the personalized models. This entire process is repeated five times for each user. Each time, the training data for the general model of a user, the training data for personalization, and the test data for performance evaluation are randomly selected and are mutually exclusive.

Three observations can be made from the experimental results shown in Fig. 4. First, the performance of the general MER model is much worse than the one shown in Fig. 3. This implies that a general MER model is not effective when it is used as the MER model of an individual listener. In other words, the subjectivity of music emotion perception has to be addressed if an MER system is to be tailored to an individual user. Second, as the results clearly show that the performance of a general model for an individual user improves after personalization, the need for personalization is well justified. Finally, we can see that the effect of personalization improves along with the number of personal annotations that are available. It seems that we need more personal annotations to model the personal perception of valence comparing to arousal. With this new dataset, we envision that more advanced personalization methods can be developed to further improve the performance of MER personalization.

4. CONCLUSION

We have created a new dataset for MER research and released it to the public. The dataset is large in the number of songs, the number of subjects recruited, and the number of annotations provided by each subject. We employed crowdsourcing to collect this dataset and used several techniques to ensure the quality of the annotations. Furthermore, we provide empirical evidences to show the need for a large dataset for MER and personalization study.

5. ACKNOWLEDGEMENT

This work was supported by the Ministry of Science and Technology of Taiwan under Contract 103-2221-E-002-166-MY3.

6. REFERENCES

[1] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intelligent Syst. Technology,* vol. 3, no. 3, pp. 40:1–40:30, May 2012.

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

[3] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE,* vol. 96, no. 4, pp. 668–696, Apr. 2008.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2000.

[5] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, pp. 465–470, 2010.

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

[7] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen, "Music emotion recognition: The role of individuality," in *Proc. ACM Int. Workshop Human-Centered Multimedia*, pp. 13–22, 2007.

[8] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Comput.,* vol. 1, no. 3, pp. 18–31, Apr. 2012.

[9] E. M. Schmidt and Y. E. Kim, "Modeling musical emotion dynamics with conditional random fields," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, pp. 777–782, 2011.

[10] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment Decision Making,* vol. 5, no. 5, pp. 411–419, Aug. 2010.

[11] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, "1000 songs for emotional analysis of music," in *Proc. ACM Int. Workshop Crowdsourcing Multimedia*, pp. 1–6, 2013.

[12] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology Music,* vol. 39, no. 1, pp. 18–49, Aug. 2010.

[13] J. A. Russell, "A circumplex model of affect," *J. Personality Social Science,* vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[14] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5412–5416, 2014.

[15] J.-C. Wang, Y.-H. Yang, K. Chang, and H.-M. Wang, "Exploring the relationship between categorical and dimensional emotion semantics of music," in *Proc. ACM Workshop Music Inform. Retrieval User-Centered Multimodal Strategies*, pp. 63–68, 2012.

[16] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, Sage, 2012.

[17] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "The acoustic emotion Gaussians model for emotion-based music annotation and retrieval," in *Proc. ACM Multimedia*, pp. 89–98, 2012.

[18] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Personalized music emotion recognition via model adaptation," in *Proc. Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf.*, pp. 1–7, 2012.

[19] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen, "Linear regression-based adaptation of music emotion recognition models for personalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 2149–2153, 2014.

[20] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. Int. Conf. Digital Audio Effects*, pp. 237–244, 2007.

[21] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an easy to use and efficient audio feature extraction software," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, pp. 441–446, 2010.

[22] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech, Language Process.,* vol. 16, no. 2, pp. 448–457, Feb. 2008.

[23] A. Sen and M. S. Srivastava, *Regression Analysis: Theory, Methods, and Applications*, Springer Science & Business Media, 1990.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 28, no. 4, pp. 357–366, Jan. 1980.

[25] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE,* vol. 63, no. 4, pp. 561–580, Apr. 1975.

[26] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Technical Report, IRCAM, Paris, France, 2004.

[27] O. Gillet and G. Richard, "Automatic transcription of drum loops," in *Proc. IEEE Int. Conf. Acoutst., Speech, Signal Process.*, pp. 269–272, 2004.