1000 Songs for Emotional Analysis of Music

Mohammad Soleymani
Imperial College London
London, UK
m.soleymani@imperial.ac.uk

Michael N. Caro
Drexel University
Philadelphia, USA
mcaro1987@gmail.com

Erik M. Schmidt
Drexel University
Philadelphia, USA
eschmidt@drexel.edu

Cheng-Ya Sha
National Taiwan University
Taipei, Taiwan
ads901119@gmail.com

Yi-Hsuan Yang
Academia Sinica
Taipei, Taiwan
yang@citi.sinica.edu.tw

ABSTRACT

Music is composed to be emotionally expressive, and emotional associations provide an especially natural domain for indexing and recommendation in today's vast digital music libraries. But such libraries require powerful automated tools, and the development of systems for automatic prediction of musical emotion presents a myriad challenges. The perceptual nature of musical emotion necessitates the collection of data from human subjects. The interpretation of emotion varies between listeners thus each clip needs to be annotated by a distribution of subjects. In addition, the sharing of large music content libraries for the development of such systems, even for academic research, presents complicated legal issues which vary by country. This work presents a new publicly available dataset for music emotion recognition research and a baseline system. In addressing the difficulties of emotion annotation we have turned to crowdsourcing, using Amazon Mechanical Turk, and have developed a two-stage procedure for filtering out poor quality workers. The dataset consists entirely of creative commons music from the Free Music Archive, which as the name suggests, can be shared freely without penalty. The final dataset contains 1000 songs, each annotated by a minimum of 10 subjects, which is larger than many currently available music emotion dataset.

Categories and Subject Descriptors

H3 [**Information storage and retrieval**]: Content Analysis and Indexing

Keywords

music, emotion, crowdsourcing

CrowdMM'13, October 22, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2506364.2506365.

1. INTRODUCTION

The appeal of music lies in its ability to express emotions, and it is commonly used for mood and emotion regulation in our daily life [13]. In seeking to develop tools for navigating today's vast digital music libraries, emotional associations provide an especially natural domain for indexing and recommendation. Because there are a myriad of challenges to such a task, powerful tools are required for the development of systems that automate the prediction of emotion in music. As such, a considerable amount of work has been dedicated to the development of automatic music emotion recognition (MER) systems [8, 10, 24, 25]. Given the perceptual nature of human emotion, most existing work on MER has pursued supervised machine learning approaches [1], training MER systems using emotion labels or ratings entered by human subjects for a number of training clips.

We are presented with an especially difficult problem in seeking to collect training data for emotion recognition; Firstly, the interpretation of emotion varies between listeners, thus requiring each clip be annotated by a distribution of subjects. This makes the collection of such data especially-time consuming and labor-intensive. In addition, given the subjective nature of the data, it is difficult to identify poor annotations caused as a result of inattentive labeling, listener fatigue, or other error.

Furthermore, one of the largest difficulties in developing systems for content-based music information retrieval (Music-IR) is the sharing of music content within the research community. Despite of years of efforts, progress on MER has been hindered by these difficulties [25] (Ch. 2). Although the emotion annotations can be distributed, this is not the case for the audio files, which are usually copyright-protected. Without public, common datasets, it is virtually impossible to compare systems built upon different training data, thereby limiting the validity of the conclusions that can be drawn. To get around these issues the common approach is to share only extracted features from the audio, such as the case with the MoodSwings Turk[1] dataset. However, the audio files are needed if one wants to extract new music features relevant to emotion expression. As a result, researchers often opt to collect the training data on their own, which is a time-consuming and labor-intensive process.

The only current evaluation task for MER is the audio mood classification (AMC) task of the annual music infor-

[1] `http://music.ece.drexel.edu/research/emotion/moodswingsturk/`

mation retrieval evaluation exchange[2] (MIREX). This task has been held since 2007, aiming at promoting MER research and providing benchmark comparisons [7]. The audio files (totaling 600 clips) are available to the participants of the task, who have agreed not to distribute the files for commercial purposes. Being the only benchmark in the field of MER so far, this contest draws many participants every year. However, AMC describes emotions using five discrete emotion clusters instead of affect dimensions (e.g., valence and arousal), which do not have origins in psychology literature, and some have noted semantic or acoustic overlap between clusters [12]. Furthermore, the dataset only applies a singular static rating per audio clip, which belies the time-varying nature of music.

In this work we present a new publicly available dataset from MediaEval benchmarking campaign[3] for the development of music emotion estimation systems. Our new benchmarking corpus employs Creative Commons[4] (CC) licensed music from the Free Music Archive[5] (FMA), which enables us to redistribute the content. For annotations we have turned to crowdsourcing using Amazon Mechanical Turk (MTurk)[6], as others have found success using these tools to label large libraries [9, 23]. In addition we have developed a two-stage procedure for filtering out poor quality workers, where workers must first pass a test demonstrating a thorough understanding of the task, and an ability to produce good quality work. The final dataset spans 1000, 40-second clips, and each clip is annotated by a minimum of 10 workers. This dataset is accessible to the researchers for non-commercial purposes after signing the usage agreement from the MediaEval website[7].

The proposed dataset is unique in the following aspects. First, the audio files are distributable under the CC license and can be shared freely. Second, annotators are usually less familiar with songs in FMA because these songs are not published by music labels, and we therefore reduce potential biases introduced by familiarity with the songs [26]. Third, it contains both clip-level, static emotion annotations and second-by-second, dynamic emotion annotations. Fourth, each song received at least 10 annotations which is larger than many existing datasets on music and affect [9] (except for MER60 [24], which uses 40 annotations per song).

2. EMOTION REPRESENTATION

Numerous models have been presented throughout psychology literature for the modeling of human emotion, spanning both categorical (discrete) [6] and parametric (continuous) [18] representations. Discrete representations of emotion, and their theoretical underpinnings, were originally inspired by the representation scheme of Darwin, who considered emotion important for survival. Discrete representations presuppose the existence of a certain number of basic and universal emotions [4, 19]. Some of the most widely-known research on basic emotions was carried out by Ekman [4], whose work demonstrated the universality of facial expressions of emotion. According to Scherer [19], there is currently no answer to the question of how many different emotions there are, but most lists in use contain 6-14 different emotions.

The challenge of ensuring that emotion categories receive a consistent interpretation contributed to the motivation for the development of dimensional approaches. Recent tools that have been developed, e.g., by [2], are more recent reflexes of the effort to use dimensional approaches to minimize the effect of differences in interpretations of discrete categories and to verify inter-participant consistency, e.g., Bradley and Lang [2]. Dimensional theories of emotion represent emotions in a continuous space. From that perspective, the discrete emotions are folk-psychological concepts that can be identified with points in this space. [14].

Dimensional representations used by psychologists often represent emotions in an n-dimensional space (generally 2 or 3-dimensional). The most well-known example of such a space is the valence-arousal (V-A) representation [18], which is the model selected for use in this work. Valence indicates positive versus negative emotion, and arousal indicates emotional intensity.

An advantage of using dimensional representations of emotion is that when people are asked to describe their emotions, they are often better at positioning content in comparison to a reference point (e.g., this song was more exciting than the previous one), compared to the situation where they are asked to provide an absolute score [26]. Using a dimensional representation of emotion also makes continuous dynamic annotation of music possible (i.e., continuously annotating songs as they are played).

Considering the advantages and disadvantages of different emotional representations, we opted to use the dimensional model and limit it to two dimensions of arousal and valence.

3. DATA COLLECTION

As previously discussed, for our music corpus we employed the Free Music Archive, an online library of high-quality music which if freely accessible. Our corpus contains 1000 clips, and our goal was to collect time-varying (per second) continuous V-A ratings, as well as a single discrete (9 point) A-V ratings applied to the entire clip. Annotations were performed via crowdsourcing using Amazon's Mechanical Turk.

3.1 Song selection

The FMA is directed by WFMU[8], one of the most renowned freeform radio stations in America. In addition to being CC licensed, the audio in FMA has been handpicked by established audio curators to ensure high quality. So far, it contains over 85,000 songs spanning a variety of genres. The following information can be obtained for each song using the FMA API[9]: song title, artist name, album name, number of listens (#listens), #downloads, #comments, #starred, song length, bit rate, MPEG audio Layer III (MP3), and URL amongst others.

We downloaded the top 300 songs (ranked according to #listens) in MP3 format for each of the following eight genres: Blues, Electronic, Rock, Classical, Folk, Jazz, Country, and Pop. We did not consider other genres such as International, Novelty, Old-times, and Spoken because they are either ambiguous or contain non-music. We then excluded

[2] http://www.music-ir.org/mirex/wiki/
[3] http://www.multimediaeval.org/mediaeval2013/emotion2013/
[4] http://www.creativecommons.org
[5] http://www.freemusicarchive.org
[6] http://www.mturk.com
[7] http://www.multimediaeval.org
[8] http://www.wfmu.org
[9] http://www.freemusicarchive.org/api/

overly long (>10 minutes) and overly short (<1 minutes) songs, and picked the top 125 songs from each genre, leading to a dataset of 1,000 songs. We did not explicitly limit the number of songs contributed by each artist, but found 53-100 unique artists for each genre, providing a very good distribution across numerous recording artists.

3.2 Annotation Interface

Inspired by previous tools such as Feeltrace and its later version GTrace [3, 15], which were developed for offline annotation of emotion in videos, we have developed our own online annotation interface for music. Shown in Figure 1 is an example of the continuous annotation interface for valence, where workers slide the ball (shown in brown) from left to right indicating the current emotion. In order to maximize workers' engagement in the task, we disabled the play and pause buttons, and only allowed audio playback when their mouse cursor was inside the rectangular box. Instead of using the mouse, workers were able to play and pause using shortcut keys on their keyboard.

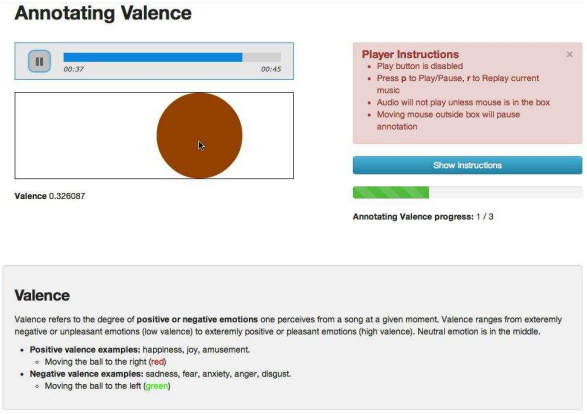

Figure 1: Dynamic annotation interface. Workers were able to move the ball to indicate the level of arousal or valence the musician was expressed while the song was being played.

After annotating the songs continuously, annotators are additionally asked to rate the level of arousal or valence for the whole clip on a 9 point scale. Self Assessment Manikins (SAM) were shown to facilitate the understanding of the scale, which have been used commonly throughout the literature [2].

We also collected data on other factors that may effect a subjects annotations. To collect data on an annotators current mood, we apply a common approach where workers were asked to choose on what extent an artificial word (i.e., a nonsense word), e.g., smon, twus, bimp, yulf , expresses a mood word [17]. The mood words were "energetic," "helpless," "nervous," "passive," "pleased," and "relaxed." The possible answers were "not at all," "very little," "somewhat," and "great extent." In addition, we automatically collected the time of day in order to study the effect of the time of the day on emotional annotation.

To ensure high-quality data, a video tutorial was also made available to the annotators which depicted the whole procedure of performing the annotation task. In addition, before allowing participants to begin each task a series of instruction boxes would pop up, showing the workers where to put the mouse, how to play the music, and reminding them of the rules. The final task consists of multiple pages and a progress bar is embedded on each page. The interface was developed with HTML5 and JavaScript using jQuery library[10]. To avoid compatibility problems, we required and verified that the workers were using Mozilla Firefox or Google Chrome browsers.

3.3 Data Collection by Crowdsourcing

Quality control is a key issue in crowdsourcing, and our strategy was designed following many current state-of-the-art crowdsourcing approaches [11, 22]. A two-step approach was taken for worker recruitment. The first step was publishing the qualification task that consisted of a single micro-task or Human Intelligence Task (HIT) involving two songs. Participants were provided with the definitions of arousal and valence and they were asked to give their demography information, including, gender, age, location. Next, they were asked to play two short music audio clips which contained highly dynamic emotion shifts; they then indicated whether arousal and valence were increasing or decreasing, ideally demonstrating an understanding of the dimensional model. In addition, they were also asked to indicate the genre of the song using multiple choice check boxes. Finally, we asked the workers to write two to three sentences describing the clips they listened to, ideally demonstrating a willingness to put reasonable effort into a task, and a basic ability to describe music (e.g., style, instrumentation, etc.).

The first HIT was used for the purpose of recruiting and screening MTurk workers as experiment participants. Workers were chosen and qualifications were granted for the main task by considering the quality of their description and the correctness of their answers. The second step was the main task described in Section 3.2 and involved a series of 334 micro-tasks. Each micro-task involved annotating 3 audio clips of 45 seconds on arousal and valence scales dynamically and statically, as a whole. Workers were paid $0.25 USD for the qualification HITs and $0.40 USD for each main HIT that they successfully completed.

3.4 Data Collection Analysis

Our qualification HIT was published for 1000 workers. We did not set any requirement such as HIT acceptance rate or the number of completed HITs so as not to shrink the pool of the workers. In total, 778 workers completed our qualification HIT. From 778 initial workers, only 287 (36.9%) succeeded in receiving the qualification required for our main HITs. We invited the qualified workers to participate in our main HITs via MTurk messaging. Out of 287 qualified workers, 100 workers (36.0%) of the invited workers, performed at least one of the main HITs. This means only 12.8% of the initial participants were qualified and performed the main HITs. From 100 workers who participated in our main HITs, 57 were male and 43 were female. Their age average was 31.7 ± 10.1. The workers on average spent 7 minutes and 40 seconds annotating three 45 seconds clips for both arousal and valence. Workers on the main HITs reported to be from 10 different countries, 72% from the USA and 18% from India and 10% from the rest of the world. In total, we spent $1,784.50 USD on collecting more than 20000 annotations. The average effective hourly rate for our HITs was $3.12 USD. On average, every worker annotated

[10] `http://www.jquery.com`

107.9 songs or 36.0 HITs; only one worker completed all the submitted HITs.

To record the ball movement in dynamic annotation we relied on mouse movement events which are not sampled regularly. The sampling frequency depends on browser, operating system status and CPU load. The average sampling interval for all the data collected was 0.23 second (4.3Hz) with the standard deviation of 0.09 second. To be conservative, we resampled the annotation time series to 1Hz (per second) sampling frequency. Once per second is significantly lower that the average sampling rate, and more than sufficient for representing emotional responses to music [16].

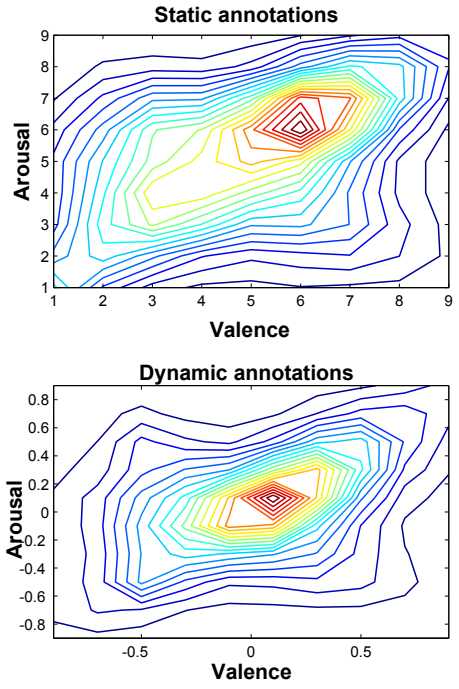

Figure 2: Contours representing the distribution of annotation in case of static (top) and dynamic (bottom) annotations.

In order to measure the inter-annotation agreement, we calculated Krippendorff's alpha on an ordinal scale for the static annotations. The Krippendorff's alpha for the static annotations on the whole clips were 0.32 for valence and 0.35 for arousal which are in the range of fair agreement. For the dynamic annotations, we used Kendall's coefficient of concordance (Kendall's W) with corrected tied ranks, for measuring inter-annotation agreement. Kendall's W is a non-parametric rank based measure and is a good indicator of the agreement between the shapes of the time series generated by dynamic annotations which is more important than the worker related constant bias. Kendall's W was calculated for each song separately after discarding the annotations of the first 5 seconds. The average W is 0.23 ± 0.16 for arousal and 0.28 ± 0.21 for valence. The observed agreement was statistically significant for arousal in 60.0% of songs and for valence in 65.8% of songs. Kendall's W showed that agreement among arousal annotations compared to the valence annotations is higher for the static annotations and lower for the dynamic annotations; the significance on agreement of dynamic annotations was tested by Wilcoxon test ($p < 5 \times 10^{-7}$). This shows that the workers are more consistent at annotating arousal in music compared to valence for the whole song whereas they are more consistent in following the valence trends dynamically. The distribution of the annotations for both static and dynamic annotations are shown in Figure 2. Dynamic and static annotations have similar distributions.

In considering the effect from the time of day, we note that most of the HITs were performed in the evening time, but there seems to be no time that workers were inactive (see Figure 3). The higher number of HITs in the evening can be explained by the workers' preference for working in the evening in addition to the presence of students with classes during the day or workers with day time jobs.

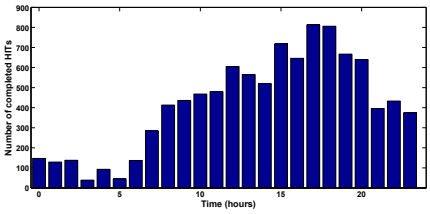

Figure 3: The average arousal and valence ratings in different energetic intervals. The higher the energetic score were workers reported higher arousal and valence scores.

Two multi-way ANalysis Of VAriance (ANOVA) test was performed once on arousal and another time on valence static ratings to test the effect of different dependent variables on the annotations. We performed two 3-way ANOVA on arousal and valence ratings as independent variables with the first two dependent variables as workers and songs. The third dependent variable was picked from the following: time of the day in minutes, mood scores given to artificial words, namely, helpless, nervous, passive, pleased and relaxed. Results of the significance and F-values are given in Table 1. The effect of time on ratings is significant for both arousal and valence. In the mood response, only "energetic" appears to have a significant effect on the ratings. We found that workers gave higher arousal and valence statics ratings when giving higher energetic scores to the artificial, nonsense words (see Figure 4).

Table 1: Multiway ANOVA test results which showed there was a significant difference between the mean values of arousal and valence ratings as a result of different variables after controlling for different workers and songs. Statistical significance was difined as $p < 0.01$.

Variable	p-value	F-score	p-value	F-score
	Arousal ratings		**Valence ratings**	
time	2×10^{-12}	1.35	4×10^{-6}	1.22
energetic	2×10^{-3}	4.92	3×10^{-3}	4.74

Even though we reviewed the performance of the workers in the qualification HITs there were still noisy annotations. The causes of the noisy annotations were two-fold. First,

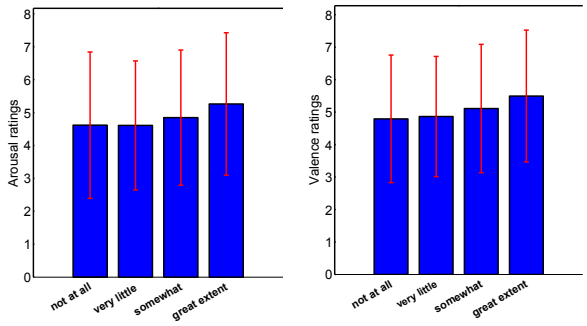

Figure 4: The average arousal and valence ratings in different energetic intervals. The higher the energetic score were workers reported higher arousal and valence scores.

some browsers failed to record the ball movements. Second, some workers failed to understand or engage in the annotation process. The first problem could have been solved by having the qualification HIT the same as the main HIT. That way we could detect the technological problems in advance and not assign the qualification. The second problem, low quality work, needs more attention. One possibility was to annotate some of the songs by experts and use them as verifiable micro-tasks. In future work, we can review the current annotations and use the high quality work as the verifiable micro-task in our design.

4. BASELINE METHOD

The static ratings given to the whole clips by the workers on both arousal and valence were averaged to serve as the ground truth. The dynamic annotation of the first 5 seconds of 45 seconds clips were discarded due to instability of their values. The arousal and valence dynamic annotation including 40 values corresponding to the last 40 seconds of the clips were averaged to generate the ground truth for the dynamic emotion estimation.

In the following, we discuss the selection of acoustic features used for music emotion recognition as well as a baseline that provides researchers with a classification algorithm and its performance statistics. In conjunction with acoustic feature selection, multivariate linear regression (MLR) was used as the baseline algorithm because of its relatively low computational complexity and effectiveness [20, 21]. This approach gives competing researchers metrics so they know how well their models are performing.

The following features have been extracted from audio signals. **MFCCs:** Mel-Frequency Cepstrum Coefficients (MFCC) were attained using Rastamat toolbox [5]. They were originally designed for speech recognition and they have been shown to be one of the most informative feature domains for music emotion recognition. **Octave-Based Spectral Contrast:** This uses 7 octave-based bands and the feature is in 14 dimensions. The seven are the spectral valley, the second are the spectral peaks. Then contrast is the difference between these two dimensions. Spectral valley sorts the values in each band in ascending order and sums the first 2% of the bandwidth. The peaks do the same, but sorting in descending order. **Statistical Spectrum Descriptors (SSDs):** This is a four dimensional feature composed of spectral centroid, spectral flux, spectral rolloff, and spectral flatness in that order. **Chromagram:** This was extracted using the Chromagram MATLAB implementation by Ellis[11] with default values. Although the chromagram appears to be one of the more intuitive representations as it provides information about the key and mode, thus far it has shown little promise towards solving this problem [21].

In addition to feature extraction via signal processing techniques in MATLAB, The Echonest API[12] was called through python to obtain features. Echonest provides beat synchronous features and therefore use varying hop times. A vector of window start times is included to help in the aggregation of these features. The effect of timbre, pitches, and loudness features on arousal-valence determination were investigated along with the aforementioned features that were extracted in MATLAB.

4.1 Results

As discussed, multivariate linear regression was selected for the baseline system because it is a simple and generalizable prediction method. 700 clips were randomly chosen as the train set and the remaining 300 clips serve as the test or evaluation set. All the annotations including for the static and dynamic ones were scaled between $[-0.5, 0.5]$. The euclidean distance between the estimated arousal and valence points as well as R^2 were calculated for the evaluation of the static results. To evaluate the dynamic results, mean distance and Kendall's Tau ranking correlation were used. The average values of arousal and valence on the training set was chosen as the random level baseline to be compared with our results. A summary of the results is given in Table 2. On the estimation of static ratings, the arousal estimations are far better than valence estimations which are in the order of chance level. Consistently, arousal estimation results are superior to valence estimation on the continuous, dynamic affect estimation task.

5. CONCLUSIONS

We introduced a new publicly available dataset for emotional analysis of music. The songs are collected from FMA under the CC license which makes them redistribute for the researchers interested in this topic. Amazon Mechanical Turk was used as a crowdsourcing platform for collecting more than 20,000 annotations on 1,000 songs. The analysis on the annotations showed there is a higher agreement in arousal ratings compared to the valence ratings. The time of the day and workers' reported "energetic" mood had a small but significant effect on the ratings. A set of baseline results were obtained using a simple linear regression and generic audio features and is reported as a reference for the future users of the dataset. The collection of both dynamic and static annotations will give the opportunity to study the effect of emotional trend on the perception of music affect as a whole.

6. ACKNOWLEDGMENT

The work of Soleymani is supported by the European Research Council under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag). The work of Yang is supported by National Science Council of Taiwan under contract NSC 101-2221-E-001-017. The work of Schmidt and Caro is supported by National Science Foundation awards IIS-0644151

[11] http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/
[12] http://www.echonest.com

Table 2: To evaluate the estimation models from content features R^2 and Euclidean distances are reported for static estimation and Kendall Tau (ρ) is reported with distance for dynamic estimation. The reported measures on dynamic annotated data are averaged for all the clips. Baseline results are calculated by setting the target to the average score in the training set. The results that are significantly better (Wilcoxon test $p < 0.01$) than the baseline (averaged training targets) are indicated with asterisk (*)(Dist: mean absolute difference and Euclidean distance for the case of two dimensions)

	Static estimation on individual clips					Dynamic estimation on 40 samples				
Features	**Arousal**		**Valence**		**Both**	**Arousal**		**Valence**		**Both**
	Dist	R^2	Dist	R^2	Dist	Dist	ρ	Dist	ρ	Dist
All	$0.10 \pm 0.07^*$	0.54	0.12 ± 0.09	0.07	0.15 ± 0.09	$0.08 \pm 0.05^*$	0.15 ± 0.22	0.09 ± 0.06	0.05 ± 0.20	$0.13 \pm 0.06^*$
MFCC	$0.11 \pm 0.08^*$	0.34	0.12 ± 0.08	0.10	0.14 ± 0.09	$0.09 \pm 0.06^*$	0.12 ± 0.22	0.09 ± 0.06	0.03 ± 0.21	$0.14 \pm 0.07^*$
Shape	$0.13 \pm 0.08^*$	0.24	0.12 ± 0.08	0.12	0.13 ± 0.09	$0.10 \pm 0.06^*$	0.10 ± 0.22	0.09 ± 0.06	0.05 ± 0.25	$0.14 \pm 0.07^*$
Contrast	$0.11 \pm 0.08^*$	0.38	0.12 ± 0.08	0.08	0.15 ± 0.09	$0.09 \pm 0.06^*$	0.10 ± 0.22	0.09 ± 0.06	0.02 ± 0.23	$0.13 \pm 0.06^*$
Chroma	0.15 ± 0.09	0.02	0.12 ± 0.08	0.04	0.12 ± 0.09	0.11 ± 0.07	0.09 ± 0.20	0.09 ± 0.06	0.02 ± 0.19	0.16 ± 0.07
Baseline	0.15 ± 0.09	-	0.13 ± 0.09	-	0.12 ± 0.10	0.12 ± 0.07	0.05 ± 0.43	0.09 ± 0.06	-0.02 ± 0.59	0.16 ± 0.08

and CNS-0960061. The authors thank Dave Rosen and Szu-Yu Chou for their contributions to the data collection.

7. REFERENCES

[1] M. Barthet, G. Fazekas, and M. Sandler. Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. In *Int'l Symp. Computer Music Modelling & Retrieval*, pages 492–507, 2012.

[2] M. M. Bradley and P. J. Lang. Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[3] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder. feeltrace': an instrument for recording perceived emotion in real time, 2000.

[4] P. Ekman. *Basic Emotions*, pages 45–60. John Wiley & Sons, Ltd, 2005.

[5] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource.

[6] K. Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, (48):246–268, 1936.

[7] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, pages 462–467, 2008.

[8] A. Huq, J. P. Bello, and R. Rowe. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, 39(3):227–244, 2010.

[9] Y. E. Kim, E. Schmidt, and L. Emelle. Moodswings: A collaborative game for music mood label collection. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, pages 231–236, 2008.

[10] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2010.

[11] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proc. annual SIGCHI Conf. Human factors in computing systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[12] C. Laurier and P. Herrera. Audio music mood classification using support vector machine. In *MIREX task on Audio Mood Classification*, 2007.

[13] A. J. Lonsdale and A. C. North. Why do we listen to music? a uses and gratifications analysis. *British Journal of Psychology*, 102:108–134, 2011.

[14] S. Marsella, J. Gratch, and P. Petta. *Computational models of emotion*, chapter 1.2, pages 21–41. Oxford University Press, Oxford, UK, 2010.

[15] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affective Computing*, 3(1):5–17, 2012.

[16] B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim. Improving music emotion labeling using human computation. In *Proc. ACM SIGKDD Workshop on Human Computation*, 2010.

[17] M. Quirin, M. Kazén, and J. Kuhl. When Nonsense Sounds Happy or Helpless: The Implicit Positive and Negative Affect Test (IPANAT). *Journal of Personality and Social Psychology*, 97(3):500–516, 2009.

[18] J. A. Russell. A circumplex model of affect. *J. Personality Social Psychology*, 39:1161–1178, 1980.

[19] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[20] E. M. Schmidt and Y. E. Kim. Prediction of time-varying musical mood distributions from audio. In *Proc. Int. Soc. Music Information Retrieval Conf.*, August 2010.

[21] E. M. Schmidt, D. Turnbull, and Y. E. Kim. Feature selection for content-based, time-varying musical emotion regression. In *Proc. ACM Int. Conf. Multimedia Information Retrieval*, Philadelphia, PA, March 2010.

[22] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010*, Geneva, Switzerland, 2010.

[23] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2011.

[24] 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*, pages 89–98, 2012.

[25] Y.-H. Yang and H. H. Chen. *Music Emotion Recognition*. CRC Press, Boca Raton, Florida, 2011.

[26] Y.-H. Yang and H. H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech & Language Processing*, 19(4):762–774, 2011.