RANKING-BASED EMOTION RECOGNITION FOR EXPERIMENTAL MUSIC

Jianyu Fan, Kıvanç Tatar, Miles Thorogood, Philippe Pasquier
Simon Fraser University
Vancouver, Canada
jianyuf, ktatar, mthorogo, pasquier@sfu.ca

ABSTRACT

Emotion recognition is an open problem in Affective Computing the field. Music emotion recognition (MER) has challenges including variability of musical content across genres, the cultural background of listeners, reliability of ground truth data, and the modeling human hearing in computational domains. In this study, we focus on experimental music emotion recognition. First, we present a music corpus that contains 100 experimental music clips and 40 music clips from 8 musical genres. The dataset (the music clips and annotations) is publicly available at: http://metacreation.net/project/emusic/. Then, we present a crowdsourcing method that we use to collect ground truth via ranking the valence and arousal of music clips. Next, we propose a smoothed RankSVM (SRSVM) method. The evaluation has shown that the SRSVM outperforms four other ranking algorithms. Finally, we analyze the distribution of perceived emotion of experimental music against other genres to demonstrate the difference between genres.

1. INTRODUCTION

The research in MER proposes computational approaches to recognize the emotion of music. The increasing numbers of MER studies in recent years have been focusing on particular musical genres, such as classical music, pop, rock, jazz, and blues [41]. So far, to our knowledge, MER in experimental music has yet to be explored.

The definition and use of the term experimental music have been an ongoing discussion within the last century. John Cage [15] clarifies the action of experimentalism as "the outcome of which is not foreseen". Demers [17] defined experimental as "anything that has departed significantly from norms of the time…" [p.7] and continues by the two assumptions of "…that experimental music is distinct from and superior to a mainstream-culture industry and that culture and history determine aesthetic experience" [p.139]. Experimental music does not only rely on harmony and melody [6]. Experimental music explores the continuum between rhythm, pitch, and noise; the notion of organized sound; the expansion of temporal field; and the morphologies of sound. In this study, our definition of experimental music encompasses experimental electronic music such as acousmatic music, electroacoustic music, noise music, soundscape compositions as well as experimental music with acoustic instruments such as free improvisation or improvised music. We also include Contemporary Art practices that use sound as a medium in our definition of experimental music.

There are many applications in which a computational model of MER for experimental music would be beneficial. MER computational models can be used in the system architecture of Musical Metacreation (MuMe) systems for experimental music. MuMe is the partial or complete automation of musical tasks [34]. A variety of MuMe systems apply machine listening. Machine listening is the computational modeling of the human hearing. In that sense, a computational model for MER in experimental music can be useful to design a machine listening algorithm for a MuMe system. Moreover, we can use computational MER models in the analysis of experimental music works. Also, we can design mood enabled recommendation systems for experimental music albums using a MER model for experimental music.

Still, MER has several challenges. First, music perception can be dramatically different if listeners are from different regions of the world and have various unique cultural backgrounds [5,18]. Second, it is difficult for researchers to collect ground truth data to cover a wide range of population that well distributed in different parts of the world [5]. Third, in the previous studies, researchers designed listening tests that asked participants to annotate the music pieces by rating their emotion perception of the music pieces [41,49]. However, the cognitive load of rating emotion is heavy for participants [9]. This causes the low-reliability of the annotations [19,44]. Fourth, the level of participant's agreement on the emotion of a music clip varies because the perception of music is subjective. Even for one individual, the ratings can change during a day [49]. Fifth, in the case of experimental music emotion recognition, there is no annotated dataset available. The current MIREX MER task is the case of pop music emotion recognition.

To overcome these difficulties, we designed a ranking-based experiment to collect ground truth annotations based on a crowdsourcing method. Crowdsourcing method is to elicit a large amount of data from a large group of people from online communities [8]. Our ground truth annotations were gathered from 823 annotators from 66 countries, which covers diverse cultural backgrounds. Then, to reduce the cognitive load, we used a ranking-based method to ask participants to do pairwise comparisons between experimental music clips. The ranking

based approach only needs relative comparisons instead of absolute ratings. This improves the objectiveness of the ground truth data. We applied the Quicksort algorithm to select comparisons during the data collection stage to reduce the workload (see Section 4.1). Then, we proposed a SRSVM method and compared it with other ranking algorithms. The results show that SRSVM is better than four other ranking algorithms regarding experimental music emotion recognition.

The database, containing the 140 music clips and the annotations, can be freely downloaded at http://metacreation.net/project/emusic/. We believe that public release of such a dataset will foster research in the field and benefit MER communities. The main contributions of this paper are thus four-fold:

- We provide a music corpus, EMusic. The corpus includes 100 experimental music clips and 40 mainstream music clips.
- We use a crowdsourcing method to collect the pairwise ranking data for experimental music clips, and share an annotated experimental music dataset.
- We proposed the SRSVM method for experimental music emotion recognition and compared our approach with other ranking algorithms.
- We compared the annotations of experimental music with that of other music genres.

2. RELATED WORKS

The Music Information Research Evaluation eXchange (MIREX) community evaluates systems for Audio Music Mood Classification every year. Studies have been classified into two major categories based on the model of emotion: categorical and dimensional approaches.

2.1 Categorical Approaches in MER

Categorical MER approaches use discrete affect models to estimate emotion. Discrete affect models propose that we can describe all emotions using a set of basic emotions. These basic emotion categories are happiness, sadness, fear, anger and disgust [22, 33], shame, embarrassment, contempt and guilt [3], as well as exuberance, anxious/frantic and contentment [32]. There is still no consensus on the discrete emotion categories of music [32].

In the previous studies with categorical MER approaches, researchers conducted experiments to collect the ground truth annotations. Then, researchers used the audio features of music clips with classification methods to model the relationship between audio features and emotion categories [23, 45, 46].

2.2 Dimensional Approaches in MER

Dimensional affect models use a Cartesian space with continuous dimensions to represent emotions [7,14,40,48]. The simplest dimensional affect model has two dimensions: valence and arousal. Other dimensional affect models with additional dimensional such as tension, potency, and dominance have also been proposed in the literature [32]. MER studies use dimensional affect models to compute continuous values that represent the emotion of audio samples. These studies focus on continuous machine learning models such as regression models. Researchers gather the ground truth data by conducting an evaluation experiment in which the participants label the emotion music clips on a dimensional affect grid.

2.3 Rating or Ranking

Affective ratings instruments have been used for collecting affective annotations. Researchers have used such tools in video emotion recognition [27, 30], music emotion recognition [11], speech emotion recognition [35], soundscape emotion recognition [20] and movement emotion recognition [43]. However, recent studies show that rating based experiments have limitations and fundamental flaws [13]. Rating-based experiments neglect the existence of interpersonal differences on the rating process. In addition, rating emotion in a continuum is difficult because annotators tend to score the samples based on the previous ratings instead of their non-biased feelings [44]. Yang and Lee indicated that the rating-based approach imposes a heavy cognitive load on the subjects [48]. Moreover, the contextual situation of annotators can affect the consistency of ratings [12].

Ranking has been an alternative approach for eliciting responses from subjects [9, 39, 48]. Metallinou and Narayanan found that there is a higher Inter-annotator reliability when people were asked to describe emotions in relative terms rather than in absolute terms [2]. Yannakakis et al. also showed that the inter-rater agreement of the ordinal data is significantly higher than that of the nominal data [12].

Yang and Chen designed a ranking-based experiment to collect ground truth data and build a ranking model recognize the perceived emotion of pop music [9]. The result showed that the ranking-based approach simplifies the annotation process and enhances the Inter-annotator reliability. Hence, we designed a ranking-based method to for experimental music emotion recognition, where annotators made pairwise comparisons between two audio clips based on valence and arousal.

2.4 Emotion Taxonomy

According to previous studies [1, 24], two types of emotions are at play when listening to music.

- Perceived emotion: Emotions that are communicated by the source.
- Induced emotion: Emotional reaction that the source provokes in listeners.

The perceived emotion is more abstract and objective. It is the emotion the source conveys. The perceived emotion of happy songs is always “happy”. However, the induced emotion is more subjective. The same happy music may not necessarily induce happiness in the listener. In this study, we focus on the perceived emotion of music clips because it is more objective.

3. DATA COLLECTION

To build a MER system for experimental music, we first built an experimental music corpus: EMusic. Then, we collected emotion annotations using a crowdsourcing method.

3.1 Corpus Construction

In EMusic corpus, there are 100 experimental music clips and 40 music clips from 8 musical genres, including blues, classical, country, electronic, folk, jazz, pop and rock. The 100 experiment music clips are extracted from 29 experimental music pieces, which are high quality works of Electroacoustic music. The 40 music clips are selected from 1000 songs database [29]. We segmented these compositions using multi-granular novelty segmentation [31] provided in the MIRToolbox [32]. Using this automatic segmentation method, we ensure that each segment is consistent. Then, we manually chose novel clips to create a homogeneous and consistent corpus that would not disturb the listeners. A 0.1 seconds fade in/out effect has been added to each audio clip.

Music clips are converted to a format in wav (44100 Hz sampling frequency, 32 bits precision and mono channel). All the audio samples are normalized. Regarding the duration, Xiao et al. [50] showed that the use of six to eight seconds is good for presenting stable mood for classical music segments. Fan et al. [19] indicated that the duration of six seconds is long enough for soundscape emotion recognition. Following the previous study, we aimed for the average duration of 6 seconds in this experiment (Mean: 6.20s, Std: 1.55s). The duration of clips varies because of the automatic segmentation by novelty.

3.2 Select Comparisons

To create a robust set of annotations, we need multiple annotations per pairwise comparison of audio clips. Baveyes et al. [44] found that collecting three annotations per comparison is a good compromise between the cost and the accuracy of the experiment. Therefore, we follow this approach for its feasibility within our experiment.

To efficiently create pairwise comparisons presented to the listeners, we use a Quicksort algorithm [44]. For the first iteration of the algorithm, we select one audio sample as the pivot. All remaining clips are to be compared with the pivot so that the algorithm generates 139 comparisons. We then collect three annotations for each comparison and determine the result to be the one that provided by at least two annotators. In the case that we did not select a pivot that has the lowest or the highest valence or arousal, we end up with two separate sets after the first iteration. Therefore we repeatedly select a new pivot in each set until each audio clip received a rank of valence and a rank of arousal from 1 to 140. The computational complexity of the Quicksort algorithm is O(*NlogN*).

3.3 Online Experiment

We conduct an online experiment to annotate our corpus of experimental music clips with affective labels. We used the CrowdFlower[1] platform to crowd source annotations from people online. To sort the 140 music clips based on valence and arousal independently, we launched one task for valence and another task for arousal.

[1] https://www.crowdflower.com/

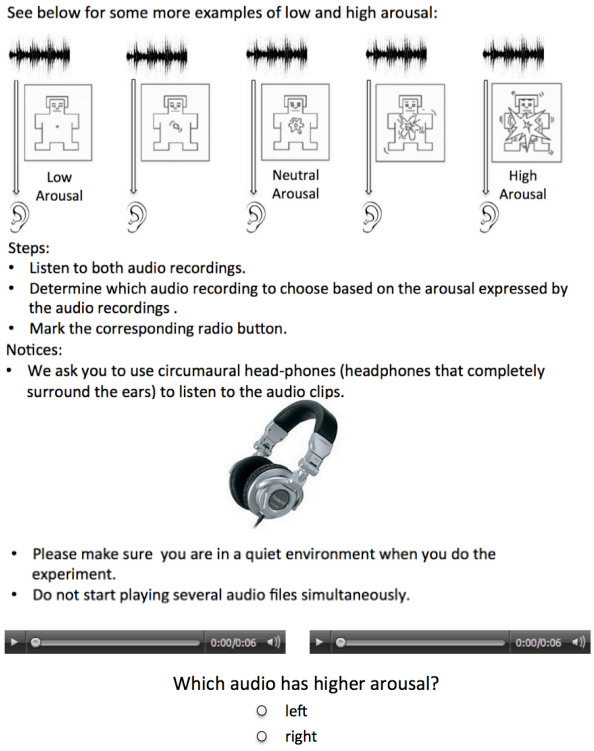

Figure 1. The interface of crowdsourcing study.

At the beginning of the annotation process, subjects are provided with the terminology of arousal and v dalence. In our experiment, we used valence to describe perceived
pleasantness of the sound. We provided subjects with the Self-Assessment Manikin [28] at the beginning of the task to make sure the task was understood. The Self-Assessment Manikin is a pictorial system used in experiments to represent emotional valence and arousal axes. Its non-verbal design makes it easy to use regardless of age, educational or cultural background. We modified the pictorial system by adding arrows to inform annotators that we were collecting perceived emotion.

We requested annotators to follow a tutorial to get familiar with the annotation interface. Annotators were notified that they were required to use headphones to listen to the audio clips. We asked them to turn the volume up to a comfortable level given a test signal. Annotators were then presented with a quiz, where 5 gold standard comparisons were provided. These comparisons were easily comparable regarding valence and arousal, which were carefully selected by experts. The annotators could continue to the task only if they achieve an 80% of accuracy in the quiz.

To ensure the quality of the annotations, we tracked annotators' performance by inserting gold standard comparisons throughout the tasks. Similar to the comparisons in the quiz, these 5 comparisons were easily comparable regarding valence and arousal. If their answers were not the same as the default answer, they would be noticed by a pop out window. If they had strong reason to explain their answer, they could message the reason to us. This also affects annotators' reputation on the CrowdFlower.

Annotators could listen repeatedly to an audio clip. After an annotator had listened to both audio clips, the option to enter the response was presented in the form of an input button. For easing the fatigue that increases naturally during manual data annotation [2], they could pause the annotation process at any time and continue at a later stage. The volume control bar was disabled so that annotators could not adjust the individual volumes themselves. An annotator had to rank 5 pairs of clips before being paid US$0.05 and was able to exit the task at any time.

3.4 Annotation Results

A total of 823 annotators performed the task from 66 different countries. Most of the workers are Venezuelans (31.71%), Brazilian (6.93%), Serbian (6.44%), Russian (5.95%) and Bosnians (5.10%). The annotators were from the world population and it is unlikely they have a background in experimental music. This avoids the potential bias brought by experts.

Each pair was displayed to annotators until three annotations are collected for this pair. 823 annotators provided 2817 comparisons for arousal and 2445 comparisons for valence. The 823 trusted annotators had an average accuracy of 91.81% in the quiz. Annotators took approximately 13s to perform a task. This also proves that annotators carefully listened to both music clips.

Categories	Arousal	Valence
Percent Agreement	0.839	0.801
Krippendorff's α	0.360	0.222

Table 1. Inter-annotator reliability.

We evaluate the Inter-annotator reliability based on percent agreement and Krippendorff's α. Percent agreement calculates the ratio between the number of annotations that are in agreement and the total number of annotations. However, percent agreement overestimates inter-annotator reliability because it does not consider the agreement expected by chance. Krippendorff's α is more flexible and allows missing data (comparisons can be annotated by any number of workers). Thus, no comparisons are discarded to compute this measure. Their values can range from 0 to 1 for Percent agreement and from -1 to 1 for Krippendorff's alpha.

In Table 1, the inter-annotator reliability is similar to other emotion studies [30, 44]. The percent agreement indicates that annotators agreed on 83.9% and 80.1% of comparisons. The value of Krippendorff's α is between 0.21 to 0.40, which indicates a fair level of agreement.

4. LEARN TO RANK

4.1 Standard Ranking Algorithms

The state-of-the-art ranking algorithms can be three categories: the pointwise approach [42], the pairwise approach [36] and the listwise approach [10]. The pointwise approach learns the score of the samples directly. The pointwise approach takes one train sample at a time and trains a classifier/regressor based on the loss of the single sample. The pairwise approach solves the ranking problems by using a pair of samples to train and provides an optimal ordering for the pair. Listwise methods try to minimize the listwise loss by evaluating the whole ranking list. Each ranking algorithm assigns a ranking score to each sample, and rank the sample based on the score.

In the following, we introduce five ranking algorithms: ListNet, Coordinate Ascent, RankNet, RankBoost and RankSVM. ListNet is a listwise ranking algorithm [10], which uses neural networks to predict the ranking score. The algorithm calculates the probability of the sample ranking within top-k, and computes the difference between the probability distribution of predicted ranks and ground truth data based on cross entropy. Coordinate Ascent algorithm is a gradient-based listwise method for multi-variate optimization [16]. It directly optimizes the mean of the average precision scores for each ranking. RankNet is a pairwise ranking algorithm, which predicts the ranking probability of a pair of samples <A, B>. If sample A receives a higher ranking score than that of sample B, then the object probability $\bar{P}_{AB}$ equals 1, otherwise, $\bar{P}_{AB}$ equals 0. The loss function of RankNet is the cross-entropy between the predicted probability and the object probability. RankBoost is another pairwise ranking algorithm [47]. It replaces training samples with pairs of samples to learn the association between samples. RankSVM is a common pairwise method extended from support vector machines [36]. The difference between features vectors of a pair of training samples can be transformed to a new feature vector to represent the pair. RankSVM converts a ranking task to a classification task.

4.2 Searching Strategies

Given a test sample, a ranking model provides a ranking score regarding valence/arousal. A ranking score is a real number. To obtain the predicted rank of the test sample based on the ranking score, we used two search strategies: one-by-one search and smoothed binary search.

4.2.1 One-by-One Search

First, we obtain predicted ranking scores of the entire training set and the test sample. Then, we sorted all clips by ranking score to obtain the predicted ranking of the test sample. Ties are unlikely to happen since we set the value of the score retains 6 digits after the decimal point.

4.2.2 Smoothed Binary Search

Smoothed binary search compares the ranking score of a test sample with the ranking scores of pivots selected from the training set to find the rankings of a test sample along the valence/arousal axis. We add a smoothed window to traditional binary by selecting a group of pivots instead of one pivot. Following is the description of the smoothed binary search:

- Given a test sample, pick an odd number of clips from the training set that are consecutive on the valence/arousal axis as pivots. The odd number of clips avoids the ties. The group of pivots has the medium value of valence/arousal among the subset.
- Predict the ranking score for the group of pivots and the test sample, and compare their ranking score. The test sample with a score of less than half of the pivots comes before the pivots, while the test

sample with a score greater than half of the pivots comes after pivots.

- Recursively apply the above steps until the size of subsets is 2. The average ranking of these two training samples is the predicted rankings.

4.3 SRSVM

We propose the SRSVM for experimental music emotion recognition. The training of SRSVM is the same as standard RankSVM. During the testing/ranking stage, SRSVM finds the predicted ranking of the test sample based on the smoothed binary search.

5. PERFORMANCE ANALYSIS

5.1 Features Selection

We began with a feature set including rms, brightness, loudness, spectral slope, spectral flux, spectral rolloff, attack leap, regularity, pulse clarity, hcdf, inharmonicity, perceptual sharpness, pitch, key, tempo, and 12 MFCCs. We used 23-ms analysis windows and calculated the mean and standard deviation to represents signals as the long-term statistical distribution of local spectral features, which ended up with a 56-dimension feature vector [21]. We used MIRToolbox [32] and YAAFE [4] libraries to extract audio features.

Selected Features
Mean of Root Mean Square
Standard deviation of Root Mean Square
Standard deviation of Brightness
Mean of MFCC 1
Standard deviation of MFCC 2
Standard deviation of MFCC 8
Mean of MFCC 12
Mean of Hcdf
Mean of Loudness
Standard deviation of Loudness
Mean of Regularity

Table 2. Selected features for predicting valence/arousal

Before training the model, we build a feature selector that removes all low-variance features over the entire corpus to select a subset of discriminative features. The threshold of variance is 0.02, which is chosen as a heuristic value. This step kept 43 features out of 56 features. Then, we used a random forests method, which has ten randomized decision trees to evaluate the importance of features based on the Gini impurity index. We ended up having an 11-dimensional feature vector (see Table. 2). Because our dataset includes 100 experimental music clips and 40 clips belong to other genres, we tested the ranking algorithms using the whole dataset and the subset of experiment music separately.

5.2 Comparing with Ranking Algorithms

We evaluate the ranking algorithms of experimental MER using Goodman-Kruskal gamma (G). Goodman-Kruskal gamma measures the association between the predicted rankings and the ground truth annotations [37, 38]. G depends on two measures: the number of pairs of cases ranked in the same order on both variables (number of concordant, N_S) and the number of pairs of cases ranked in reversed order on both variables (number of discordant, N_D). G ignores ties. In our experiment, we had no ties. G is close to 1 indicate strong agreement, -1 for total disagreement, and 0 if the rankings are independent.

$$G = \frac{N_S - N_D}{N_S + N_D} \tag{1}$$

We used the leave-one-out validation method to compare the SRSVM with ListNet, RankNet, Coordinate Ascent, and RankBoost. For a given test sample, ranking algorithms output a predicted valence/arousal score. To obtain the predicted rankings of the whole test set, we used one-by-one searching strategy and smoothed binary search strategy. Then, we measured the gamma between the predicted rankings and the ground truth annotation.

As we can see from Table 3, when we use SRSVM, we obtain the best performance when the windows size is three samples (G: 0.733, $p < 0.001$). When the window size is 1, the test sample will be compared with one pivot iteratively until it falls into a small interval. This becomes a standard binary search. After adding a smoothed window, the test sample is compared with a group of pivots. This increases the accuracy of predicting whether the test sample is larger or smaller than the pivots.

Algorithm	One-by-One Search	Smoothed Binary Search (Number of samples)		
		1	3	5
ListNet	0.044	0.088	0.057	0.022
RankNet	0.096	0.386	0.269	0.255
Coordinate Ascent	0.191	0.436	0.387	0.486
Rank-Boost	0.619	0.679	0.697	0.717
RankSVM	0.398	0.690	**0.733** **SRSVM**	0.697 **SRSVM**

Table 3. Goodman-Kruskal gamma of ranking algorithms for arousal recognition using the whole dataset

Method	One-by-One Search	Smoothed Binary Search (Number of samples)		
		1	3	5
ListNet	0.015	0.002	0.049	0.002
RankNet	0.063	0.155	0.055	0.260
Coordinate Ascent	0.016	0.130	0.195	0.254
Rank-Boost	0.438	0.467	0.345	0.440
RankSVM	0.333	0.490	**0.573** **SRSVM**	0.556 **SRSVM**

Table 4. Goodman-Kruskal gamma of ranking algorithms for valence recognition using the whole dataset

When using the whole dataset, the valence recognition is harder than arousal recognition. However, the SRSVM still obtains the best performance (G: 0.573, $p < 0.001$).

Method	One-by-One Search	Smoothed Binary Search (Number of samples)		
		1	3	5
ListNet	0.001	0.037	-0.013	0.013
RankNet	0.110	0.096	0.242	0.299
Coordinate Ascent	0.237	0.515	0.519	0.556
Rank-Boost	0.698	0.741	0.740	0.748
RankSVM	0.300	0.776	**0.801** SRSVM	0.776 SRSVM

Table 5. Goodman-Kruskal gamma of ranking algorithms for arousal recognition using the subset that only contains experimental music clips.

As Table 5 shows, when we only consider experimental music, the Gamma statistic of SRSVM for arousal recognition has the best result (G: 0.801, $p < 0.001$). The results of the experimental music case are better than the results of the case including clips of all genres.

Method	One-by-One Search	Smoothed Binary Search (Number of samples)		
		1	3	5
ListNet	0.115	0.037	-0.012	0.036
RankNet	0.058	0.116	0.246	0.277
Coordinate Ascent	0.067	0.100	0.131	0.106
Rank-Boost	0.167	0.236	0.279	0.346
RankSVM	0.434	0.570	**0.795** SRSVM	0.628 SRSVM

Table 6. Goodman-Kruskal gamma of ranking algorithms for valence recognition using the subset that only contains experimental music clips.

Table 6 shows that when we only consider experimental music, the Gamma statistic of SRSVM for valence recognition (G: 0.795, $p < 0.001$) is significantly higher than using the whole dataset.

From Table 3-6, we can see the best performing model is SRSVM with 3 samples as the smoothed window. The second best performing model is SRSVM with 5 samples as the smoothed window. This result implies that a good emotion-recognition can be obtained by using SRSVM.

5.3 Comparing between Experimental Music and Other Genres

We convert the rankings to ratings to visualize the distribution of the ranking data. This illustration has two assumptions. First, the distances between two successive rankings are equal. Second, the valence and arousal are in the range of [-1.0, 1.0].

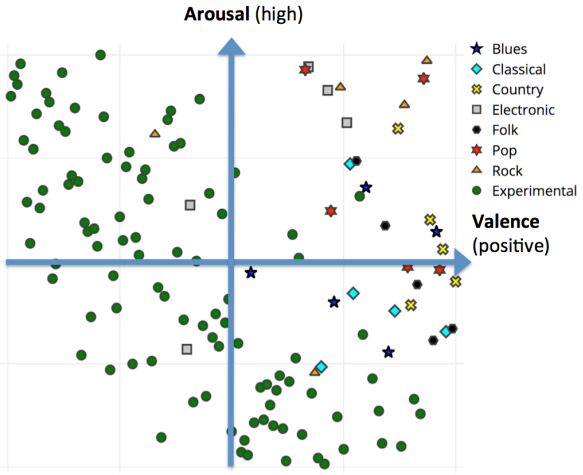

Figure 2. The distribution of the ground truth annotations, the green dots represent experimental music clips

From Figure 2, it can be observed that other genres have both higher perceived valence and arousal comparing to experimental music. Because we have only 5 samples per genre, we need to have a large ground truth dataset to prove that assumption. The figure also shows the negative correlation between valence and arousal of experimental music clips. To test this, we run a Pearson correlation test on the ground truth data. Our Pearson correlation coefficient is -0.3261, which indicates there is a weak negative correlation between the two dimensions.

6. CONCLUSIONS AND FUTURE WORKS

We present an annotated dataset for experimental music emotion recognition. 140 music clips are ranked along the valence and arousal axis through a listening experiment. It is available at http://metacreation.net/project/emusic/. We presented a SRSVM method to predict rankings of experimental music clips regarding valence/arousal and compared SRSVM with other ranking method. We also compared the valence and arousal of experimental music with that of the music of other genres, which shows other genres of music have both higher perceived valence and arousal than experimental music.

Even with the smaller number of clips, we found other genres have both higher perceived valence and arousal comparing to experimental music. In the future, we plan to compare the perceived emotion of different genres by collecting a larger dataset.

7. REFERENCES

[1] A. Kawakami, K. Furukawa, K. Katahira and K. Okanoya, “Sad music induces pleasant emotion,” Front Psychol Vol. 4, No. 311, 2013.

[2] A. Metallinou and S. Narayanan, “Annotation and processing of continuous emotional attributes: challenges and opportunities,” *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–8, 2013.

[3] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological review.* Vol. 97, No. 3, pp. 315-331, 2014.

[4] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an Easy to Use and Efficient Audio Feature Extraction Software," *Proceedings of the International Symposium on Music Information Retrieval*, pp. 441–446. 2010.

[5] C. J. Stevens, "Music perception and cognition: A review of recent cross-cultural research," *Topics in Cognitive Science*, Vol. 4, No. 4, pp. 653– 667, 2012.

[6] C. Palombini, "Pierre Schaeffer. 1953: Towards an Experimental Music," *Music and Letters*, Vol. 74, No. 4, pp. 542–57, 1993.

[7] D. Liu, L. Lu, and H.-J. Zhang, "Automatic mood detection from acoustic music data," *Proceedings of the International Symposium Music Information Retrieval*,, pp. 81–87, 2003.

[8] D. McDuff, "Crowdsourcing affective responses for predicting media effectiveness," *Ph.D. Dissertation*. Massachusetts Institute of Technology, 2014.

[9] D. Yang and W.-S. Lee, "Disambiguating music emotion using software agents," *Proceedings of the International Conference on Music Information* Retrieval, 2004.

[10] F. Xia, T.-Y. Liu, J. Wang, W.-S. Zhang, and H. Li, "Listwise approach to learning to rank: Theory and algorithm," *Proceedings of the IEEE International Conference on Machine. Learning*, pp. 1192–1199, 2008.

[11] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," *Proceedings of the IEEE International Conference Acoustics, Speech and Si*gnal Processing, 2014.

[12] G. N. Yannakakis and H. P. Matínez, "Grounding Truth via Ordinal Annotation," *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2015.

[13] G. N. Yannakakis, H. P. Mart´ınez, "Ratings are overrated!" *Frontiers on Human-Media Interaction*, Vol. 2, No. 13, 2015.

[14] J. A. Sloboda and P. N. Juslin, "Psychological perspectives on music and emotion," in *Music and Emotion: Theory and Research*, Oxford University Press, 2001.

[15] J. Cage, *Silence: Lectures and Writings*, Wesleyan, 1961.

[16] J. Chen, C. Xiong, and J. Callan, "An empirical study of learning to rank for entity search," *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016.

[17] J. Dermers, *Listening through the Noise: The Aesthetics of Experimental Electronic Music*, Oxford University Press, 2010.

[18] J. Fan and M. Casey, "Study of Chinese and UK hit songs prediction," *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, pp. 640–652, 2013.

[19] J. Fan, M. Thorogood, P. Pasquier, "Automatic Soundscape Affect Recognition Using A Dimensional Approach," *Journal of the Audio Engineering Society*, Vol. 64, No. 9, pp. 646–653, 2016.

[20] J. Fan, M. Thorogood, and P. Pasquier, "Automatic Recognition of Eventfulness and Pleasantness of Soundscape," *Proceedings of the 10th Audio Mostly*, 2015.

[21] J. J. Aucouturier and B. Defreville, "Sounds like a park: A computational technique to recognize soundscapes holistically, without source identification," *Proceedings of the International Congress on Acoustics*, pp. 621–626, 2009.

[22] J. Panksepp: *Affective Neuroscience: The Foundation of Human and Animal Emotions*, Oxford University Press, 1998.

[23] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification—a hybrid approach," *Proceedings of the International Conference on Music Information Retrieval*, pp. 657-662, 2009.

[24] K. Kallinen and N. Ravaja, N, "Emotion perceived and emotion felt: Same and different," *Musicae Scientiae*, Vol. 5, No. 1, pp. 123-147, 2006.

[25] K. Svore, L. Vanderwende, and C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources," *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 448–457, 2007.

[26] L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*. Vol. 49, No. 268, pp-732–764, 1954.

[27] L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches," *Proceedings of the International conference on Language Resources and Evaluation*, 2006.

[28] M. M. Bradley and P. J. Lang, "Measuring emotion: the self- assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, Vol. 25, No. 1, pp. 49–59, 1994.

[29] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pp. 1–6, 2013.

[30] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," *Proceedings of the IEEE International Conference Acoustics, Speech and Signal Processing*, pp. 2376–2379, 2011

[31] O. Lartillot, D. Cereghetti, K. Eliard, and D. Grandjean, "A simple, high-yield method for assessing structural novelty," *Proceedings of the 3rd International Conference on Music & Emotion,* 2013

[32] O. Lartillot, P. Toiviainen, and T. Eerola, "A Matlab Toolbox for Music Information Retrieval," in: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, P. D. R. Decker, (Eds), *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization,* pp. 261–268, Springer, Berlin, Heidelberg, 2008.

[33] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*. Vol. 6, No. 3, pp.169–200, 1992.

[34] P. Pasquier, A. Eigenfeldt, O. Bown, and S. Dubnov, "An Introduction to Musical Metacreation," *ACM Computers In Entertainment, Special Issue: Musical Metacreation*, Vol. 14, No. 2, 2016.

[35] R. Elbarougy and M. Akagi, "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," *Proceedings of the Signal & Information Processing Association Annual Summit and Conference*, 2012.

[36] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," *Proceedings of International Conference on. Artificial* Neural *Network*, 1999.

[37] R. Morris, "Crowdsourcing workshop: The emergence of affective crowdsourcing," *Proceedings of the Annual Conference Extended Abstracts on Human Factors in Computing Systems*, 2011.

[38] R. Morris and D. McDuff, "Crowdsourcing techniques for affective computing," in R.A. Calvo, S.K. DMello, J. Gratch and A. Kappas (Eds). *Handbook of Affective Computing*, Oxford University Press, 2014.

[39] S. Ovadia, "Ratings and rankings: Reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, Vol. 7, No. 5, pp. 403–414, 2004.

[40] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," *Proceedings of the International Symposium Music Information Retrieval*, pp. 621–626, 2009.

[41] T. Eerola, Tuomas, and J. K. Vuoskoski. "A Review of Music and Emotion Studies: Approaches, Emotion Models, and Stimuli," *Music Perception: An Interdisciplinary Journal*, Vol. 30, No. 3, pp. 307–340, 2013.

[42] T. Y. Liu, "The Pointwise Approach," in *Learning to rank for information retrieval*. Berlin: Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2011.

[43] W. Li, P. Pasquier, "Automatic Affect Classification of Human Motion Capture Sequences in the Valence-Arousal Model," *Proceedings of the International Symposium on Movement and Computing*, 2016.

[44] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, Vol. 6, No. 1, pp. 43–55, 2015.

[45] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," *Proceedings of the International Conference on Information Retrieval*, pp. 375–376, 2013.

[46] 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," *Proceedings of the International Conference on Music Information Retrieval*, 2010.

[47] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Proceedings of the International Conference on Machine Learning*, pp. 170–178, 1998.

[48] Y.-H. Yang and H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 762– 774, 2011.

[49] Y.-H. Yang and H.-H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intel. Systems & Technology*, Vol. 3, No. 3, 2012.

[50] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "What is the best segment duration for music mood analysis?" *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, pp. 17–24, 2008.