Emotional Analysis of Music: A Comparison of Methods*

Mohammad Soleymani
University of Geneva
Carouge (GE), Switzerland
mohammad.soleymani@unige.ch

Anna Aljanaki
Utrecht University
Utrecht, the Netherlands
a.aljanaki@uu.nl

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

ABSTRACT

Music as a form of art is intentionally composed to be emotionally expressive. The emotional features of music are invaluable for music indexing and recommendation. In this paper we present a cross-comparison of automatic emotional analysis of music. We created a public dataset of Creative Commons licensed songs. Using valence and arousal model, the songs were annotated both in terms of the emotions that were expressed by the whole excerpt and dynamically with 1 Hz temporal resolution. Each song received 10 annotations on Amazon Mechanical Turk and the annotations were averaged to form a ground truth. Four different systems from three teams and the organizers were employed to tackle this problem in an open challenge. We compare their performances and discuss the best practices. While the effect of a larger feature set was not very apparent in the static emotion estimation, the combination of a comprehensive feature set and a recurrent neural network that models temporal dependencies has largely outperformed the other proposed methods for dynamic music emotion estimation.

Categories and Subject Descriptors

H5.5 [**Information storage and retrieval**]: Content Analysis and Indexing

Keywords

Music, emotion, crowdsourcing, audio features, music emotion recognition, performance evaluation

1. INTRODUCTION

Music as an art is inherently emotionally expressive. The emotional characteristics of music are invaluable for music indexing and recommendation. There are however a number of challenges in identifying the emotion expressed by music. As such, a considerable amount of work has been dedicated to the development of automatic music emotion recognition (MER) systems [2, 10]. This paper presents a cross-comparison of the four different systems proposed to perform this task. The valence-arousal (V-A) model of emotion is used in this work. This is a dimensional model of emotion, in which valence ranges from unpleasant to pleasant emotions, and arousal indicates emotional intensity from passive to activated.

The *Emotion in Music* task was presented in the open challenge[1] MediaEval 2013 benchmarking initiative for multimedia evaluation. The only other evaluation task for MER is the audio mood classification (AMC) task of the annual music information retrieval evaluation exchange (MIREX),[2] which offers the possibility of analyzing the annotated audio files (totaling 600 excerpts) on their own restricted access server. However, AMC describes emotions using five discrete emotion clusters instead of affect dimensions, which do not have origins in psychological research, and some have noted semantic or acoustic overlap between clusters [4]. Furthermore, the dataset only applies a singular static rating per audio excerpt, which belies the time-varying nature of music.

Our new benchmarking corpus employs music licensed under Creative Commons (CC),[3] enabling us to redistribute the content, from the Free Music Archive (FMA),[4] an online library of high-quality music. A 45 seconds excerpt was extracted from a random point of each song to reduce the annotation load and make a uniform dataset. To collect annotations, we have turned to crowdsourcing using Amazon Mechanical Turk (MTurk),[5] which was successfully used by others to label large libraries [7]. 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. Each excerpt is annotated by a minimum of 10 workers, which is substantially larger than any existing music dataset with dynamic annotations of emotion.

2. DATASET AND TASK DESCRIPTION

The task, presented at MediaEval 2013, was composed of two subtasks. In the first task, the dynamic emotion characterization task, arousal and valence were estimated for the given song dynamically in time, with temporal resolu-

*The work of Aljanaki, Veltkamp and Wiering is supported by the FES project "COMMIT/". See the last section for the rest of the authors.

MM'14, November 03 - 07 2014, Orlando, FL, USA.

ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2655019.

[1] http://www.multimediaeval.org
[2] http://www.music-ir.org/mirex/wiki
[3] http://www.creativecommons.org
[4] http://www.freemusicarchive.org
[5] http://www.mturk.com

tion being one second. The second task, the static emotion characterization task, required participants to deploy multimodal features to automatically detect the overall valence and arousal for each song. We initially developed a dataset of 1,000 songs. However, a set of duplicates were later discovered and removed, which reduced the size of the dataset to 744 songs. The dataset was split between the development set (619 songs) and the evaluation set (125 songs).

The dynamic annotations were collected using a web-interface on a scale from -1 to 1, where the Mechanical Turk workers could dynamically annotate the songs on valence and arousal dimensions separately while the song was being played. The static annotations were made on nine-point scale on valence and arousal for the whole 45 seconds excerpts after the dynamic annotations. We also collected data on other factors that may affect annotations. To estimate annotator's current mood, we use an implicit mood assessment method [5], asking a worker to choose to which extent an artificial noword, e.g., "smon", "twus", or "bimp", expresses a mood. The mood words were "energetic," "aggressive," "helpless," "nervous," "passive," "pleased" and "relaxed," with the four possible answers ranging from "not at all" to "to a great extent." In addition, we automatically collected the time of day in order to study its effect on emotional annotation. For a detailed description of the crowdsourcing techniques employed and dataset statistics we refer the reader to [6]. The database is freely available on the internet[6].

2.1 Data analysis

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 excerpts were 0.54 for valence and 0.55 for arousal, which are in the range of moderate agreement. For the dynamic annotations, we used Kendall's coefficient of concordance (Kendall's W) with corrected tied ranks to measure 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 0.17 ± 0.18 for arousal and 0.21 ± 0.22 for valence. The observed agreement was statistically significant (p-value<0.05) for arousal in 70.4% of songs and for valence in 75.5% of songs. Kendall's W showed that agreement among arousal annotations compared to the valence annotations is almost equal for the static annotations and lower for the dynamic annotations. This shows that the workers were more consistent in following the valence trends dynamically.

We have discarded the first 5 samples of dynamic annotations, taking into account the reaction time of the raters. However, this number was chosen arbitrarily, based on our own experience. In order to investigate how long it takes for the raters to get a stable understanding of songs' emotional expressions, we calculated the Krippendorff's alpha on interval scale for all the songs and for all the time samples for both valence and arousal. The results are shown in Figure 1, which shows that valence inter-annotator agreement stabilizes around 10 seconds whereas the arousal dynamic annotations take longer to stabilize. This might be

[6]http://cvml.unige.ch/databases/emoMusic/

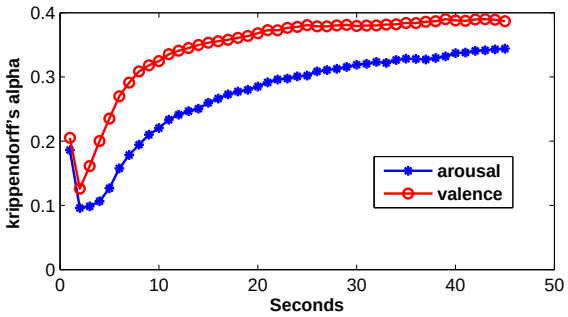

Figure 1: Krippendorff's alpha of dynamic annotations averaged over all dynamic samples.

one of the reasons for higher inter-annotator agreement for dynamic valence annotations. We hence conclude that we should have discarded the first 10 to 20 seconds to have more reliable dynamic labels.

A mixed linear model was utilized to study the effect of different independent variables on the annotations, once for the static arousal scores and once for the static valence scores. In this model, the independent variables were considered as fixed effect. The independent variables were time of the day, as a nominal variable in hours, and mood scores, as an ordinal variable on four-point scale, assigned to nonwords, e.g., aggressive and energetic. The effect of songs and workers were considered as random effects in the model. We only considered the intercepts to change by the random effects and not the slopes. An analysis of variance (ANOVA) test on the coefficients found some of them to be significantly different from zero and therefore have significant effect on arousal, namely, time of day ($F(23) = 2.35$, $p < 3 \times 10^{-4}$), and energetic ($F(3) = 3.31$, $p < 0.02$) mood score. Mood and time of day did not have any effect on valence.

The averaged dynamic annotations were strongly correlated with the static scores ($\rho_{arousal} = 0.95$ and $\rho_{valence} = 0.94$). Given that the dynamic annotations were done before the static ones this shows that the overall impression of the annotators stayed stable throughout the annotation session.

3. METHODS

The ground truth for the first dynamic task consists of 40 values corresponding to the last 40 seconds of the excerpts, that were averaged across workers. For the static task, we averaged ratings given by the workers to the whole excerpts on both valence and arousal.

To evaluate the estimation models from content features the R^2 statistics and root-mean-square error (RMSE) are reported for static estimation and averaged correlation ($\bar{\rho}$) and RMSE are reported for dynamic estimation. Averaged correlation is a measure of the similarity of the trends and waveforms, whereas the RMSE provides an estimate of how far off the estimations were. The reported measures on dynamic annotated data are averaged for all the excerpts. Random level results are calculated by setting the target to the average score in the development set.

3.1 Baseline method

A simple baseline system was developed by the task organizers to provide a baseline to the participants to beat and compare how well their models are performing. In the following, in conjunction with acoustic feature selection, multivariate linear regression (MLR) was used as the baseline

algorithm because of its relatively low computational complexity and its effectiveness. The MLR was trained on the development set and evaluated on the evaluation set.

For the Baseline system we extracted the following features from the audio signals: Mel-frequency cepstrum coefficients, chromagram, octave-based spectral contrast and statistical spectrum descriptors, such as spectral centroid, flux, rolloff and flatness. The Echonest[7] API was used to generate additional timbre, pitch and loudness features.

3.2 TUM system

Technische Universität München (TUM) team's approach is based on supra-segmental features calculated by applying statistical functionals to the contours of frame-wise low-level descriptors (LLDs) over either one-second segments or whole songs. It has been shown in [9] that this set of affective features provides robust cross-domain assessment of emotion (continuous valence and arousal) in speech, music and acoustic events. Despite its rather 'brute-force' nature, it outperformed a hand-crafted set of musically motivated features for MER [9].

The set contains 6,373 features. LLDs include auditory weighted frequency bands, their sum (corresponding to loudness), spectral measures such as centroid, skewness and sharpness. Furthermore, voicing related LLDs such as fundamental frequency (corresponding to 'main melody') and harmonics-to-noise ratio (corresponding to 'percussiveness') are added. Delta regression coefficients are added to capture time dynamics. Statistical functionals include for example mean, moments, quartiles, as well as contour related measurements such as rise and fall times, amplitudes of local maxima, and linear and quadratic regression coefficients. An exhaustive list and a detailed analysis of feature relevance for MER can be found in [9].

TUM used support vector regression (SVR) for song-level regression and bidirectional long short-term memory recurrent neural networks (BLSTM-RNNs) for dynamic regression. In addition, to improve modeling of the dynamic emotion profile, TUM investigated adding delta regression coefficients of the valence and arousal targets as additional regression tasks. The complexity constant for SVR training was varied from 10^{-4} to 10^{-1}. BLSTM-RNNs with two hidden layers (128 LSTM units per layer and direction) were used. Therefore, the first layer performs information reduction to a 128-dimensional feature set. The segments of each song were processed in order, forming sequences. Gradient descent with 25 sequences per weight update was used for training. An early stopping strategy was used, using a held out validation set in each fold. To alleviate over-fitting to the high dimensional input feature set, Gaussian noise with zero mean and standard deviation 0.6 was added to the input activations, and sequences were presented in random order during training. Ten BLSTM-RNN were trained on the ten training folds of the development set; segment level predictions were averaged across networks.

3.3 UAizu system

University of Aizu team (UAizu) proposed the following approach. Features were extracted from excerpts downsampled to 22,050 kHz. UAizu tried various standard features such as MFCC, line spectral pairs, chromagram, timbre features such as spectral centroid, flux and zero crossing rate), spectral crest factor and spectral flatness measure. All feature vectors were calculated using the Marsyas toolbox [8] with 512-sample frames with no overlap. For the dynamic emotion estimation task, first order statistics (mean and standard deviation) of the feature vectors were calculated for a window of about 1 second giving 45 vectors per excerpt. For the static emotion estimation, the same statistics for these 45 vectors were calculated, resulting in a single high dimensional feature vector per song.

Valence and arousal were modeled by separate Gaussian process regression (GPR). UAizu used standard Gaussian likelihood function which allows exact inference to be performed. The GP mean was set to zero and the type of covariance kernel was set a composite function including the sum of squared exponential (SE) and rational quadratic (RQ) function defined as follows:

- `SE`: $k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp(-(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')/2l^2)$
- `RQ`: $k(\mathbf{x}, \mathbf{x}') = \sigma^2(1 + (\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')/2\alpha l^2)^{-\alpha}$

where σ and l are parameters learned from the training data.

3.4 UU system

Utrecht University (UU) team used two MATLAB toolboxes, the MIRtoolbox [3] and the PsySound [1], to extract 42 audio features. The features include low-level spectral features, as well as some high level harmonic and rhythmic features such as harmonic change detection function (HCDF) [3], and dynamic loudness (using the model of Chalupper and Fastl [1]). For the dynamic task, the features were extracted from every second of the audio with no overlap; for the static task, the averaged features and their standard deviation was used.

UU discovered some outliers in the data, containing speech, noise and clapping. The most influential ones were removed by calculating Cook's distance for each data point for both valence and arousal regressive models. The outlier points with Cook's $d > 0.05$ both for valence and arousal (4 files) were removed, which increased prediction accuracy.

UU used M5 algorithm to select features. M5 removes the weakest features until no improvement is observed in the estimate of the error given by the Akaike information criterion. For valence, 24 features were selected; for arousal, 27 features out of 42. Among the most important features were loudness (accounted for 44% of variance in case of arousal and 9% of variance in case of valence), spectral centroid, entropy, spectral spread and HCDF.

The dynamic valence and arousal were estimated using SVR with the RBF kernel using WEKA[8], where parameters were optimized manually.

4. RESULTS AND DISCUSSION

All the static and dynamic annotations and subsequently the predictions were scaled between $[-0.5, 0.5]$. A summary of the results is given in Table 1. Both in the static and dynamic task, the arousal estimations are far better than valence estimations. All the RMSE for dynamic estimations of valence and arousal for the three submissions are significantly lower (one-sided Wilcoxon test p-value<0.01) than the random level (averaged training targets) and than the Baseline. The simple Baseline system was able to perform better than random for arousal but fell short of performing

[7] `http://www.echonest.com/`

[8] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 1: To evaluate the estimation models from content features the R^2 statistics and root-mean-square error (RMSE) are reported for static estimation and averaged correlation ($\bar{\rho}$) and RMSE for dynamic estimation. $\bar{\tau}$ is the normalized ranking distance of the retrieved songs using emotions. For RMSE and $\bar{\tau}$, smaller is better and for R^2 and $\bar{\rho}$ larger is better, where RMSE,$R^2, \bar{\tau} \in [0, 1], \bar{\rho} \in [-1, 1]$. Acronyms: RND: random level, BSL: Baseline

(a) Static

Run	Arousal		Valence		Ranking Dist.
	RMSE	R^2	RMSE	R^2	$\bar{\tau}$
RND	0.16	0	0.15	0	0.43
BSL	0.12	0.48	0.15	0	0.40
TUM	0.10	0.59	0.11	0.42	0.34
UAizu	0.10	0.63	0.12	0.35	0.35
UU	0.10	0.59	0.12	0.31	0.35

(b) Dynamic

Run	Arousal		Valence	
	RMSE	$\bar{\rho}$	RMSE	$\bar{\rho}$
RND	0.25 ± 0.13	0.10 ± 0.33	0.23 ± 0.11	0.05 ± 0.31
BSL	0.25 ± 0.11	0.16 ± 0.36	0.23 ± 0.10	0.06 ± 0.30
TUM	0.08 ± 0.05	0.31 ± 0.37	0.08 ± 0.04	0.19 ± 0.43
UAizu	0.10 ± 0.05	0.11 ± 0.36	0.09 ± 0.05	0.06 ± 0.28
UU	0.10 ± 0.06	0.14 ± 0.28	0.12 ± 0.07	-0.01 ± 0.27

better than random for valence estimation for both static and dynamic subtasks. However, their correlations vary by submissions and emotion dimensions. For the static subtask, all submissions outperformed the provided Baseline. The dynamic subtask appeared to be more challenging and only TUM could consistently beat the baseline performance. The BLSMT-RNN takes advantage of temporal dependencies, which is not supported by MLR, SVR or GPR that was used in the Baseline, UAizu and UU systems.

Although the inter-annotator agreement of valence is higher for the dynamic task, the accuracy of arousal prediction is still higher than that of valence in both static and dynamic tasks, which is consistent with results reported in the literature [2]. The more comprehensive set of features in TUM could outperform the other systems in the static estimation of valence but not arousal. The combination of a large feature set and a recurrent neural network has largely outperformed the other proposed methods for dynamic music emotion estimation. Although it is important to have a model that is intuitive, in practice it is still advantageous to extract a large number of audio features and feed them into a sophisticated machine learning model. UU team also showed that outlier removal and feature selection can improve the performance.

In order to test whether the R^2 and RMSE metrics are representative for real use cases of music information retrieval, we calculated the averaged Kendall Tau $\bar{\tau}$ normalized ranking distance (the smaller the better) of a hypothetical retrieval system as follows. Any given song in the evaluation set was taken as a query by example given to a system that indexes songs based on valence and arousal scores. We then calculated the ranked retrieved results based on their similarity defined by the Euclidean distance of the songs to each other with static valence and arousal scores as features. The ground truth ranking to which the ranking distance was calculated was created based on the ground truth annotations whereas the second ranking was based on the estimated valence and arousal scores by regression models. The results are shown in the last column of Table 1(a). The averaged ranking distances ($\bar{\tau}$) are consistent with the regression evaluation metrics we employed.

From the experience of collecting this dataset we found it important to carefully listen to the audio files to remove outliers and duplicates. Moreover, in order to get a more stable dynamic annotation, we should discard the first few seconds of the annotations; we regret that we only experimented with the case of dropping the first 5 seconds albeit our analysis shows dropping 10 to 20 seconds might be better. Our current study only identifies a few user factors that have some influence on the behavior/preference of the annotators in emotion annotation. In future work, we will explore more factors such as the credential of an MTurk worker, personality traits, gender, age, musical expertise, active musicianship, broadness of taste and familiarity with music.

5. SUMMARY

In this paper, we have presented a comparative study of four systems for automatic music emotion recognition, which employ different feature sets and training schemes. The study is conducted on a novel dataset of substantial size of music dynamically annotated with emotion, with a detailed discussion of the implications of the results. We hope this study can contribute to the advancement of affective analysis in music and other art forms.

6. ADDITIONAL AUTHORS

Michael N. Caro, Drexel University, USA, Florian Eyben, Technische Universität München (TUM), Germany, Konstantin Markov, University of Aizu, Japan, Björn Schuller, TUM/Imperial College London, Germany/UK, Remco Veltkamp, Utrecht University, Netherlands, Felix Weninger, TUM, Germany and Frans Wiering, Utrecht University, Netherlands.

7. REFERENCES

[1] D. Cabrera. PsySound: A computer program for psychoacoustical analysis. *Proc. Australian Acoustical Society Conf.*, pages 47–54, 1999.

[2] Y. E. Kim et al. Music emotion recognition: A state of the art review. In *ISMIR*, 2010.

[3] T. P. Lartillot, O. A Matlab toolbox for musical feature extraction from audio. *Conf. Digital Audio Effects*, 2007.

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

[5] M. Quirin, M. Kazén, and J. Kuhl. When Nonsense Sounds Happy or Helpless: The Implicit Positive and Negative Affect Test (IPANAT). *J. Pers. Soc. Psych.*, 97(3):500–516, 2009.

[6] M. Soleymani et al. 1000 Songs for Emotional Analysis of Music. In *ACM MM*, CrowdMM '13, pages 1–6, 2013.

[7] J. A. Speck et al. A comparative study of collaborative vs. traditional musical mood annotation. In *ISMIR*, 2011.

[8] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech & Audio Processing*, 10(5):293–302, 2002.

[9] F. Weninger et al. On the acoustics of emotion in audio: What speech, music and sound have in common. *Frontiers in Emotion Science*, 4(Article ID 292):1–12, 2013.

[10] Y.-H. Yang and H.-H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intel. Systems & Technology*, 3(4), 2012.