

USING SEMANTIC LAYER PROJECTION FOR ENHANCING MUSIC MOOD PREDICTION WITH AUDIO FEATURES

Pasi Saari and Tuomas Eerola

Finnish Centre of Excellence in Interdisciplinary Music Research
University of Jyväskylä, Finland
firstname.lastname@jyu.fi

György Fazekas and Mark Sandler

Centre for Digital Music
Queen Mary University of London
firstname.lastname@eecs.qmul.ac.uk

ABSTRACT

We propose a novel technique called Semantic Layer Projection (SLP) for predicting moods expressed by music based on audio features. In SLP, the predictive models are formed by a two-stage mapping from audio features to listener ratings of mood via a semantic mood layer. SLP differs from conventional techniques that produce a direct mapping from audio features to mood ratings. In this work, large social tag data from the Last.fm music service was analysed to produce a semantic layer that represents mood-related information in a low number of dimensions. The method is compared to baseline techniques at predicting the expressed Valence and Arousal in 600 popular music tracks. SLP clearly outperformed the baseline techniques at predicting Valence ($R^2 = 0.334$ vs. 0.245), and produced roughly equivalent performance in predicting Arousal ($R^2 = 0.782$ vs. 0.770). The difficulty of modelling Valence was highlighted by generally lower performance compared to Arousal. The improved prediction of Valence, and the increasingly abundant sources of social tags related to digital music make SLP a highly promising technique for future developments in modelling mood in music.

1. INTRODUCTION

The modern age of digital music consumption has brought new challenges in organising and searching rapidly expanding music collections. The popular appeal of music is often attributed to its striking ability to elicit or convey emotion. Therefore, managing large music collections in terms of mood has significant advantages that complement conventional genre-based organisation.

Social music services such as Last.fm¹ play an important role in connecting digital music to crowd-sourced semantic information. A prime advantage of using Last.fm data is in the large number of users worldwide applying semantic tags, i.e., free-form labels, to elements of the music domain, e.g. tracks, artists and albums. Tags are used in order to communicate users' music listening preferences that are also used for improving the service. The data is available to

researchers through a dedicated API, which makes it possible to apply semantic computing to tags related to millions of tracks. Semantic computation of Last.fm tags has been found effective in characterising music information related to genre, mood, and instrumentation [1]. Parallel to analysing crowd-sourced tags, a tag set dedicated to music research purposes has also been collected in [2]. The importance of mood tags has been highlighted in several studies, including [3], claiming that mood tags account for 5% of the most commonly used tags. Applying semantic computation to tags can therefore yield effective mood-related semantic models for music.

The prominence of mood in music is reflected by the large number of studies modelling expressed or induced emotion. To this end, two prevalent techniques emerged: *i*) the dimensional model of Valence, Arousal and Tension; and *ii*) the categorical model of basic emotions such as happiness, sadness and tenderness. On one hand, these models have been found mutually inclusive to a large degree [4]. On the other hand, more general models of emotion have also been proposed, and refined using a taxonomy specifically designed for musically induced emotion [5].

These types of representations have been widely used in computational systems for predicting mood from audio. Feature extraction methods have been developed, for instance, in [6] and [7], providing a good basis for modelling and predicting perceived moods, genres and other characteristics of musical audio. The typical approach in most previous studies involves the use of computational algorithms, such as supervised machine learning, to predict perceived moods directly from audio features. For a more detailed overview of the advances of mood modelling and recognition, see e.g. [8].

Achieving high efficiency of these models, however, relies heavily on good quality ground-truth data. Due to the expense of human annotation, ground-truth is laborious to collect, and therefore typical data sets are limited to a few hundred tracks. This leads to challenges in mood prediction emerging from the high dimensionality of audio feature data and from the need for complex model parameter optimisation, often resulting in the lack of generalizability of the predictions to novel tracks [9]. One way of overcoming these challenges and increasing the efficiency of mood prediction is to utilise audio content related to a large number of tracks and associated crowd-sourced semantic tags.

In this work, we use multivariate techniques in a novel way to predict listener ratings of mood in 600 popular mu-

¹ Last.fm: <http://www.last.fm/>

sic tracks, using an intermediate semantic layer created from tag data related to a substantially large collection of tracks. This demonstrates how a large collection of tracks and associated mood tags can be used to improve prediction quality. The new technique involves mapping audio features (audio level) to a semantic mood space (semantic layer) first, and then mapping the semantic mood space to listener ratings (perceptual level). This differs from conventional methods that map audio directly to the perceptual level. Instead, we use direct mapping as baseline to assess the efficiency of the proposed technique.

2. RELATED WORK

This section summarises past research on connecting audio, as well as semantic and perceptual levels to represent music. Figure 1 illustrates how previous studies relate to the approach presented here.

2.1 Mapping from Audio Features to Semantic Layer

The challenge of auto-tagging music tracks can be considered analogous to our task. Gaussian Mixture Modelling (GMM) was used in [10], whereas [11] employed Support Vector Machines (SVM) for this purpose. Bertin-Mahieux et al. [12] proposed a boosting-based technique. This provided higher precision (0.312) and overall F-score (0.205) with somewhat lower recall (0.153) compared to hierarchical GMMs proposed in [10], when a set of general tag words were considered. In the context of mood tags, the authors reported 0.449, 0.176, 0.253 precision, recall and F-score, respectively, noting that, due to the specific experimental conditions, the results are bounded at a value lower than one. Miotto and Lanckriet [13] found that using semantic modelling of music tags improves auto-tagging compared to the conventional approach of treating each tag individually without any tag similarity information. The proposed Dirichlet mixture model (DMM) captured the broader context of tags and provided an improved peak precision (0.475) and F-score (0.285) compared to previous results using the same data set, when combining DMM with different machine learning techniques.

2.2 Mapping from Audio Features to Perceived Mood

Yang et al. [14] modelled moods represented in the Arousal-Valence (AV) plane using Support Vector Regression (SVR) with LIBSVM implementation [15] trained on audio features. Reported performance was lower for Valence ($R^2 = 0.281$) than for Arousal ($R^2 = 0.583$). Eerola et al. [16] compared various linear regression models at predicting multidimensional emotion ratings with acoustical features. A set of film soundtrack excerpts collected in [4] were used in this experiment. The best models based on Partial Least Squares Regression (PLS) showed high performance at predicting listener ratings of Valence, Arousal, and Tension ($R^2 = 0.72, 0.85, 0.79$). Especially for Valence, the performance was strikingly higher than in [14]. The same soundtrack data was utilised in classification of music to four basic emotion categories in [9], showing the maximum accuracy of 56.5%. Audio features related to tonality

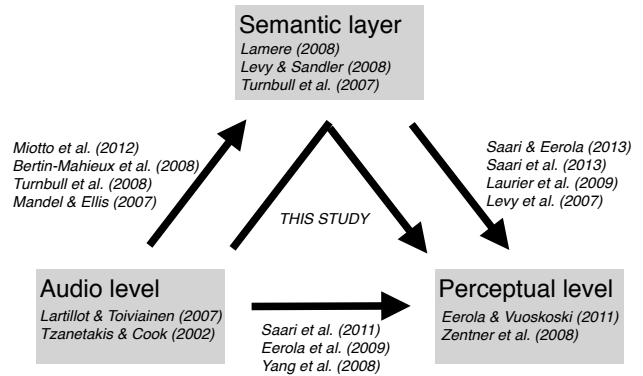


Figure 1. The difference of the present and past studies in mapping between audio features, semantic layer, and perceptual level. Selected past research is cited for each sub-task.

(average majorness of the mode and key clarity), as well as to the average slope of the onset attacks were found to be the most effective predictors of the perceived mood. SVM has been particularly popular in the annual MIREX mood classification challenge² representing the state-of-the-art in the field. Moreover, SVM together with ReliefF feature selection produced competitive results [17].

2.3 Mapping from Semantic Layer to Perceived Mood

The studies of Laurier et al. [18] and Levy et al. [19] compared semantic models of mood based on social tags to emotion models proposed by research in affective sciences, as well as expert-generated mood categories used in the MIREX challenge. The accuracy of tag-based semantic models at predicting listener ratings of musical mood was assessed in [20], proposing a technique called Affective Circumplex Transformation (ACT) for the task, based on previous research in affective sciences [21, 22].

ACT was used to predict perceived mood in 600 popular music tracks. The results showed promising performance ($R \approx 0.60$) for the ratings related to the dimensional emotion model as well as separate mood terms. Similar analysis across separate sources of curated editorial annotations for production music, and crowd-sourced Last.fm tags for commercial music, was performed in [23]. The results suggested that semantic models of mood based on tags can be used interchangeably to predict perceived mood across different annotation types and track corpora.

To apply the approach taken in [20] and [23] to new track corpora, semantic annotations need to be available for the corresponding tracks. In order to predict mood in unannotated track corpora, one must rely on other type of information, such as audio features. In the present study, we show how semantic tag data that was found to be promising and relevant in previous work can be used to enhance audio-based mood prediction.

² http://www.music-ir.org/mirex/wiki/MIREX_HOME

Valence (“negative” / “positive”), Arousal (“calm” / “energetic”), Tension (“relaxed” / “tense”), Atmospheric, Happy, Dark, Sad, Angry, Sensual and Sentimental. The excerpts were sampled from full tracks corresponding to positions in the Last.fm previews. SET600 consists of 15s clips using 320kB/s mp3 format.

3.3 Audio Feature Extraction

Audio features describing dynamics, rhythm, pitch, harmony, timbre and structure were extracted from SET10K and SET600 using the MIRtoolbox [6]. Statistical means and standard deviations over features extracted from various short 50% overlapping time frames were computed to obtain song-level descriptors. The resulting set of 128 features is presented in Table 2. For the features describing rhythmic repetition (127-128) and zero crossing rate (43-44), we used long frame length of 2s, whereas for chromagram-based features such as the repetition of register (125-126), key clarity (19-20), centroid (17-18), mode (21-22), HCDF (23-24), and roughness (25-26) we used a frame length of 100ms. For other features the frame length was 46.4ms except for low-energy ratio (3), which was extracted directly from the full extent of the signal.

Features from SET10K were normalised using the z-score transform. All feature values more than 5 standard deviations from zero were considered outliers and truncated to the extremes $[-5, 5]$ (0.1% and 1.3% of the values in SET10K and SET600 respectively). SET600 was then normalised according to the means and standard deviations of SET10K. In particular, we discovered a slight discrepancy in mean RMS energy (1) between SET10K and SET600. The energy was generally higher in SET600, perhaps due to the use of different MP3 encoders. However, this was ignored in our study for simplicity.

3.4 Regression Techniques and Model Evaluation

3.4.1 Semantic Layer Projection

We propose a novel technique for mood prediction in music termed Semantic Layer Projection (SLP). The technique involves mapping audio features to perceived mood in two stages using the semantic mood level as a middle layer, instead of the conventional way of mapping audio features directly to the perceived mood. SLP may be implemented with several potential mapping techniques. We choose to use PLS for the first mapping, due to its higher performance demonstrated in previous research, and linear regression for the second.

First, we apply PLS to the SET10K to produce a mapping from audio features to the 10-dimensional semantic mood representation obtained using ACT. We compare two variants of the semantic mood layer: (SLP_{10D}) track projections in all 10 dimensions of the mood space, and (SLP_{1D}) track projections in separate dimensions corresponding to Valence (1st dim.), and Arousal (2nd dim.). To map from audio features to the semantic layer, we apply PLS to each dimension separately. Then, we project the audio features of SET600 to the semantic layer using the obtained mappings. Finally, we apply linear regression between the 10-

Table 2. Extracted feature set. Feature statistics (m = mean, d = standard deviation) are computed across sample frames.

Category	No.	Feature	Stat.
Dynamics	1-2	RMS energy	m, d
	3	Low-energy ratio	—
	4-5	Attack time	m, d
	6-7	Attack slope	m
Rhythm	8-9	Fluctuation (pos., mag.)	m
	10	Event density	m
	11-12	Pulse clarity	m, d
	13-14	Tempo	m, d
Pitch	15-16	Pitch	m, d
	17-18	Chromagram (unwr.) centr.	m, d
Harmony	19-20	Key clarity	m, d
	21-22	Key mode (majorness)	m, d
	23-24	HCDF	m, d
	25-26	Roughness	m, d
Timbre	27-28	Brightness (cutoff 110 Hz)	m, d
	29-30	Centroid	m, d
	31-32	Flatness (< 5000 Hz)	m, d
	33-34	Irregularity	m, d
	35-36	Skewness (< 5000 Hz)	m, d
	37-38	Spectr. entropy (<5000 Hz)	m, d
	39-40	Spectr. flux	m, d
	41-42	Spread	m, d
	43-44	Zerocross	m, d
MFCC	45-46	1st MFCC	m, d
	⋮	⋮	⋮
	69-70	13th MFCC	m, d
	71-96	1st-13th Δ MFCC	m, d
	97-122	1st-13th $\Delta(\Delta)$ MFCC	m, d
Structure	123-124	Repetition (spectrum)	m, d
	125-126	Repetition (register)	m, d
	127-128	Repetition (rhythm)	m, d

dimensional (SLP_{10D}) and 1-dimensional (SLP_{1D}) layer representations and the listener ratings.

We optimise the number of components used in the PLS mappings using 50×2 -fold cross-validation. In each fold, we divide SET10K into training and test sets, and estimate how well the PLS mapping based on train set fits the test set. To decide on the number of components, we apply (50, 100)-fold cross-indexing proposed in [9]. Cross-indexing is a technique developed to tackle model over-fitting in choosing the optimal model parameterisation from several candidates. Finally, we use the selected number of components to form a model based on the whole SET10K.

3.4.2 Baseline Techniques

In this study, two baseline techniques – PLS and Support Vector Regression (SVR) – were compared with SLP. These techniques were chosen since they represent regression methods that were already found efficient in previous MIR studies. Baseline techniques were applied in the usual way, mapping audio features of SET600 directly to the ratings of perceived mood.

We use PLS in a conventional way with 2 components as in [16]. In SVR, we use the Radial Basis Function (RBF) kernel and apply grid search to optimise the cost ($C = 2^l$, $l \in [-3, \dots, 3]$) and gamma ($\gamma = 2^l$, $l \in [-13, \dots, 8]$) model parameters. Moreover, we optimise the set of audio features used in SVR by feature subset selection. To

this end, we apply the ReliefF [25] feature selection algorithm adapted for regression problems. ReliefF produces relevance weights $\tau \in [-1, 1]$ for the individual features by taking into account their prediction potential and redundancy. To choose a subset of the features, we use a relevance weight threshold $\tau_0 = 0$ and include all features with $\tau > \tau_0$.

3.4.3 Cross-Validation Procedure

For validating the performance of the techniques, we use 50×2 -fold cross-validation corresponding to 2-fold cross-validation run 50 times, and report the mean and standard deviation over the 100 performance estimates for each technique. All model optimisation and feature selection is based solely on the training set at each run.

4. RESULTS AND DISCUSSION

In SLP_{10D} and SLP_{1D} we use the rank $k = 16$ for SVD computation. This choice of k was found effective in [20], while other values had no consistent effect on the performance and did not improve the results.

Fig. 3 shows the performance of each technique at predicting the ratings for Valence and Arousal. For Valence, it is evident that SLP outperformed the baseline techniques. SLP_{10D} gave the highest performance ($R^2 = 0.334 \pm 0.035$), outperforming SLP_{1D} ($R^2 = 0.252 \pm 0.032$). SLP_{10D} performed at significantly higher level ($t(99) = 17.994, p = 5.63 \times 10^{-33}$)³ than SVR ($R^2 = 0.245 \pm 0.048$), while the difference between SLP_{1D} and SVR was not significant. Conventional PLS was the least efficient with a performance of $R^2 = 0.152 \pm 0.045$.

Cross-indexing to optimise the number of PLS components in mapping from audio features to the semantic space yielded 7 components for SLP_{10D} and 13 components for SLP_{1D} . The number of components for SLP_{10D} is the average across 10 dimensions, while the latter relates to the first dimension of SLP_{10D} . The regression model used in the second-stage mapping of SLP_{10D} relied heavily on the first semantic dimension related to Valence: the first dimension showed an average significance of $p \approx 10^{-4}$ across cv-folds. SLP_{10D} model therefore bears a strong similarity to the SLP_{1D} . ReliefF feature selection to optimise the set of audio features used in SVR yielded on average 43 features ($SD = 11$).

In general, the fact that SLP_{1D} outperformed SVR shows the efficiency of SLP. In SLP_{1D} tracks are explicitly projected to Valence already in the first-stage mapping from the audio features to the semantic layer. Therefore minimal learning is required within SET600 for the second-stage mapping to perceived mood. This contrasts to the extensive adaptation to SET600 in SVR, which involves feature selection, cost and gamma optimisation, as well as support vector optimisation.

The overall performance for predicting Valence was at a significantly lower level than the performance of $R^2 = 0.72$ reported in [16]. Most notably, the PLS technique that was successful in [16] did not give convincing performance

here. Since the set of audio features used in these studies is similar, the difference in performance is possibly due to the variety of genres covered by SET600. This is in contrast with the previous study using only film soundtracks. Film music is composed to mediate powerful emotional cues [4], which may provide higher variance in feature values so that better representations can be learnt. However, the performance in the present study is in line with other past research such as [14] ($R^2 = 0.281$).

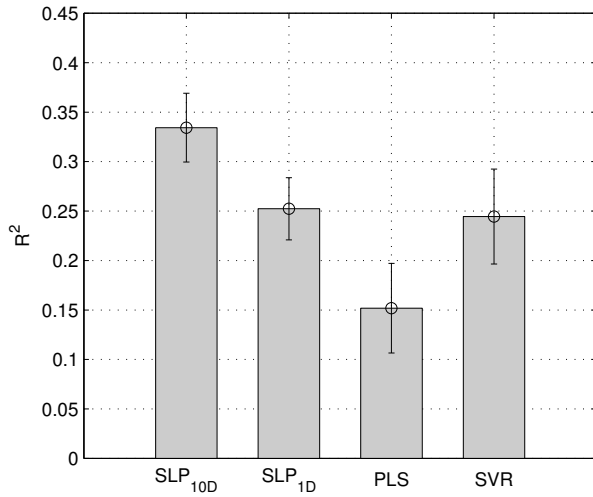
All techniques gave notably higher performance for Arousal than for Valence. In this case, SLP_{10D} again yielded the highest values ($R^2 = 0.782 \pm 0.020$), but outperformed SVR ($R^2 = 0.770 \pm 0.028$) only marginally. PLS gave the third highest performance ($R^2 = 0.751 \pm 0.027$) outperforming SLP_{1D} ($R^2 = 0.745 \pm 0.019$). For Arousal, SLP_{1D} used five PLS components, while the performance of SVR was obtained with 37 features on average ($SD = 9$). Again, the second-stage regression model in SLP_{10D} relied mainly on the 2nd dimension ($p \approx 2 \times 10^{-9}$) related to the Arousal dimension used in SLP_{1D} . Despite more complex training within SET600, SLP_{10D} gave only slight, although highly significant ($t(99) = 5.437, p = 5.4 \times 10^{-7}$) performance gain over SVR. In fact, all techniques performed better than $R^2 = 0.7$, which corroborates past findings that audio features provide a robust basis for modelling perceived Arousal in music.

Similar patterns in the general performance levels between techniques were found in modelling ratings in the other seven scales related to individual mood terms. In general, moods that are characterised by high or low arousal, such as Angry and Atmospheric, performed at similar, yet slightly lower level than Arousal, whereas moods such as Happy and Sad – characterised by positive and negative valence – produced performance similar to Valence.

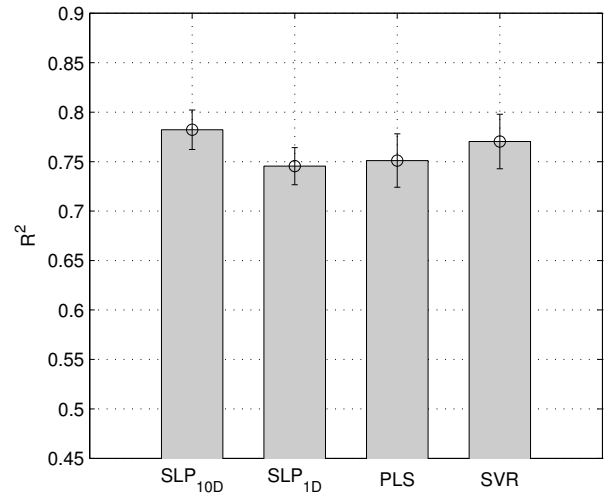
Since SLP_{10D} produced clearly the highest performance for Valence, while outperformed SVR by a more modest margin for Arousal, it is worth to compare the potential of these techniques in future approaches to mood prediction. SVR represents a sophisticated state-of-the-art technique that is efficient in learning characteristics of the training data relevant to the target mood, but requires complex optimisation of multitude of model parameters. Robust learning of SVR, and any method that could be used as baseline is solely dependent on high quality training data, which is typically laborious to collect. This also means that generalizability of these models to unknown music tracks, and possibly to new music genres, can not be guaranteed, as found in [26]. On the other hand, the efficiency of SLP is primarily based the first-stage mapping from audio to the semantic layer, and require only minimal adaptation to test data. This is suggested by the promising results of SLP_{1D} that produced explicit mood estimates already at the first-stage.

Semantic data required to built the semantic layer can be collected from online services by crowd-sourcing. Some services already make available data related to millions of tracks. Therefore, the cost of collecting training data for SLP is related mostly to obtaining the audio representation of the training set. Larger data for the semantic layer

³ Pairwise Student's t-test across cv-folds.



(a) Valence.



(a) Arousal.

Figure 3. Performance ($R^2 \pm sd$) for each technique in predicting the perceived mood.

enables more delicate learning and would presumably increase the model performance. We therefore claim that the potential of SLP in future mood prediction approaches is higher than that of SVR. Note, however, that as SLP in general can be implemented with any prediction model, SVR can in fact be implemented in the future as the mapping technique within SLP.

Finally, we seek to gain understanding of what audio features are the most useful for modelling Valence and Arousal. We apply SLP_{10D} using each audio feature category described in Table 2 separately. Table 3 shows the results. Eight harmony-related features including Mode and Key clarity were found to be the most useful in predicting Valence ($R^2 = 0.186$), and in fact, the model using only these 8 features would have outperformed PLS using all features. Features describing timbre, structure, and MFCC showed modest potential for predicting Valence ($R^2 > .10$), whereas rhythm features were largely redundant in this particular task. Prediction of Arousal was on the other hand highly efficient with most feature categories. Timbre ($R^2 = 0.687$) and MFCC ($R^2 = 0.649$) features performed the best. Prediction with harmony-related features was also competitive ($R^2 = 0.653$), while even the four pitch-related features could predict Arousal at moderate level ($R^2 = 0.471$).

In general, these results support previous findings that harmony-related features are useful in mood prediction [9], and that timbre-related features are more useful for predicting Arousal. The results also highlight the need to either optimise existing harmony-related features, or to uncover and investigate a wider variety of audio descriptors for Valence prediction.

5. CONCLUSIONS

In this study we developed a novel approach to predict the perceived mood in music called Semantic Layer Projection (SLP). By introducing a two-stage mapping from

Table 3. Performance ($R^2 \pm sd$) of SLP_{10D} using different audio feature categories. Number of features in each category are presented in brackets.

	Valence	Arousal
Dynamics (7)	0.092 ± 0.031	0.536 ± 0.034
Rhythm (7)	0.056 ± 0.044	0.583 ± 0.028
Pitch (4)	0.074 ± 0.034	0.471 ± 0.031
Harmony (8)	0.186 ± 0.035	0.653 ± 0.030
Timbre (18)	0.141 ± 0.037	0.687 ± 0.027
MFCC (78)	0.123 ± 0.030	0.649 ± 0.026
Structure (6)	0.127 ± 0.043	0.547 ± 0.025

audio features to semantic layer and finally to mood ratings, SLP provides a way to exploit semantic information about mood learnt from large music collections. It also facilitates building predictive models for disparate music collections. The proposed technique outperformed SVR, a sophisticated predictive model on the Valence dimension, and produced prediction performance roughly at the same level on the Arousal dimension.

The results highlight the difficulty of modelling the Valence dimension in music. However, SLP provides clear advantage compared to baseline techniques specifically in this task, which signifies its high potential that can be developed further in more general audio and semantics-based mood recognition models.

Future direction of the present study includes using more efficient collection of tracks to represent the semantic layer, and improving the prediction of Valence via an extension of the audio feature set. Moreover, a version of the proposed technique that takes musical genre into account – possibly by introducing a genre layer – will be developed to further generalise our model to many different types of music collections.

6. REFERENCES

- [1] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.
- [2] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," in *Proceedings of the 30th international ACM SIGIR conference on information retrieval*, 2007, pp. 439–446.
- [3] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.
- [4] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychol. Music*, vol. 39, no. 1, pp. 18–49, 2011.
- [5] M. Zentner, D. Grandjean, and K. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.
- [6] O. Lartillot and P. Toivainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects, Bordeaux, France*, September 2007.
- [7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
- [8] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models," in *Proc. of the 9th Int. Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012, pp. 492–507.
- [9] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 6, pp. 1802–1812, aug. 2011.
- [10] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.
- [11] M. I. Mandel and D. P. Ellis, "Multiple-instance learning for music information retrieval," in *Proceedings of 9th International Conference of Music Information Retrieval (ISMIR)*, 2008, pp. 577–582.
- [12] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *Journal of New Music Research*, vol. 37, no. 2, pp. 115–135, 2008.
- [13] R. Miotto and G. Lanckriet, "A generative context model for semantic music annotation and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1096–1108, 2012.
- [14] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.
- [15] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [16] T. Eerola, O. Lartillot, and P. Toivainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *9th International Conference on Music Information Retrieval*, 2009, pp. 621–626.
- [17] R. Panda and R. P. Paiva, "Music emotion classification: Dataset acquisition and comparative analysis," in *n 15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.
- [18] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags," in *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 381–86.
- [19] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.
- [20] P. Saari and T. Eerola, "Semantic computing of moods based on tags in social media of music," *IEEE Transactions on Knowledge and Data Engineering*, manuscript submitted for publication available at <http://arxiv.org/>, 2013.
- [21] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [22] K. R. Scherer, *Emotion as a multicomponent process: A model and some cross-cultural data*. Beverly Hills: CA: Sage, 1984, pp. 37–63.
- [23] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of mood expressed by music: Comparison between crowd-sourced and curated editorial annotations," in *IEEE International Conference on Multimedia and Expo (ICME 2013): International Workshop on Affective Analysis in Multimedia (AAM)*, In press 2013.
- [24] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.
- [25] M. Robnik-Sikonja and I. Kononenko, "An adaptation of relief for attribute estimation in regression," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ser. ICML '97. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1997, pp. 296–304.
- [26] T. Eerola, "Are the emotions expressed in music genre-specific? an audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journal of New Music Research*, vol. 40, no. 4, pp. 349–366, 2011.