PREDICTION OF MULTIDIMENSIONAL EMOTIONAL RATINGS IN MUSIC FROM AUDIO USING MULTIVARIATE REGRESSION MODELS

Tuomas Eerola, Olivier Lartillot, Petri Toiviainen
Finnish Centre of Excellence in Interdisciplinary Music Research,
University of Jyväskylä, Finland
firstname.lastname@jyu.fi

ABSTRACT

Content-based prediction of musical emotions and moods has a large number of exciting applications in Music Information Retrieval. However, what should be predicted, and precisely how, remain a challenge in the field. We provide an empirical comparison of two common paradigms of emotion representation in music, opposing a multidimensional space to a set of basic emotions. New ground-truth data consisting of film soundtracks was used to assess the compatibility of these models. The findings suggest that the two are highly compatible and a quantitative mapping between the two is provided. Next we propose a model predicting perceived emotions based on a set of features extracted from the audio. The feature selection and transformation is given special emphasis and three separate data reduction techniques are compared (stepwise regression, principal component analysis, and partial least squares regression). Best linear models consisting of 2-5 predictors from the data reduction process were able to account for between 58 and 85% of the variance. In general, partial least squares models performed the best and the data transformation has a significant role in building linear models.

1. INTRODUCTION

Emotional impact of music is one of the most important reasons for listening to music. A reliable content-based prediction of emotions in music would be a highly useful application of MIR, as suggested by the promising prototypes recently been put forward. It seems however that an improvement of the study would require a precise clarification of the concept under study, which is difficult due to the inherent fuzziness of the topic. Previous research defined mood as "sound and feel" of music (*AllMusicGuide*), of "feeling inspired by the music pieces" (*Last.fm*) [1]. Such broad opening of the study to a large realm of semantic expression, although interesting by itself, makes however the problem particularly difficult to tackle. Dealing with the concept of *emotion* instead, which is rooted on a large background of scientific research, would enable on the contrary a better controlled study of research.

However, robust and generalizable prediction of emotions has been difficult for several reasons, namely due to their conceptual elusiveness, their highly contextual dependencies on situation, context and musical style, and the limitations of the computational approaches utilised to date, which emphasize mainly on low-level acoustic features. The conceptual elusiveness of emotions is apparent in both the multitude of theoretical approaches taken, as well as the high individual variability in the subjective self-reports of emotional experiences. During the past decade, basic emotion model, dimensional models, and domain-specific emotion models have all received support in studies of music and emotion [2]. However, it still remains to be clarified whether models and theories designed for everyday emotions – such as the basic emotion model – can also be applied in an aesthetic context such as music. It has been argued, for example, that a few primary basic emotions seem inadequate to describe the richness of the emotional effects of music [3].

Current computational efforts of modelling polyphonic timbre seem to have reached what Aucouturier has called a 'glass-ceiling' effect, probably due to their strict reliance on low-level audio features. This ceiling appears to be around 50-60% of the variance explained [4]. Out of these three shortcomings, we aim to provide advances in two of them, namely by carrying simultaneous conceptual comparison of basic emotions and the circumplex model, and by performing the selection of relevant audio and musical features by means of multivariate methods.

2. BACKGROUND

2.1 Mood, emotion, and affect terms

Mood ontologies structure emotional adjectives and labels into a set of various mood clusters. Following purely theoretical studies [5, 6], more systematic approaches attempt to automatically infer the set of clusters based on analysis of large set of mood labels that are further reduced with the help of statistical tools: agglomerative hierarchical clustering of 179 AMG mood labels [1], consensus among a set of candidate labels used in literature [7, 8] collected through psychological experiments [9, 10], etc.

Representation of emotion in a dimensional affective space has gained support among researchers in music and emotion [2]. Instead of claiming that independent neural system exists for every basic emotion, the two-dimensional circumplex model [7] proposes that all affective states arise from two independent neurophysiological systems: one related to *valence* (a pleasure-displeasure continuum) and the other to *activity* (activation-deactivation). In contrast, Thayer [11] suggested that the two underlying dimensions of affect were two separate arousal dimensions: *energetic arousal* and *tense arousal*. However, the two-dimensional models have been criticized for their lack of differentiation when it comes to emotions that are close neighbours in the valence-activation space, such as anger and fear. It has also been discovered, that the two-dimensional model is not able to account for all the variance in music-mediated emotions [12] and three-dimensional variant containing valence, energy arousal and tension arousal has given better empirical results [13].

2.2 Ground truth collection

Extensive work has been carried out for the collection of ground truth related to mood ontology [10, 14]. Concerning the dimensional paradigm, Kim et al [15] have collected dynamic ratings expressed on the valence-activity space from thousands of songs drawn randomly from the uspop2002 database via a customized online game.

2.3 Mood and emotion prediction

Previous computational works attempt to predict mood clusters [16, 17] and emotion categories [18, 19]. Lu, Liu, and Zhang [20] studied mood detection and tracking using a variety of acoustic features related to intensity, timbre, and rhythm. Their classifier used Gaussian Mixture Models (GMMs) for Thayer's four principal mood quadrants in the valence-activity representation. The system was trained using a set of 800 classical music clips, each 20 seconds in duration, hand labeled to one of the 4 quadrants. Their system achieved an accuracy of 85% when trained on 75% of the clips and tested on the remaining 25%.

We believe that linear models are more useful than classifications for understanding emotion in music. Indeed, music is often emotionally ambiguous and listeners are not particularly certain of the emotion categories if given complex examples. Valence and activity mapping has been previously done [21, 22], but selecting the optimal set of features is more challenging, due to statistical constraints imposed by linear models.

3. NEW GROUND-TRUTH SET: SOUNDTRACKS

In the present work, both discrete and dimensional models of emotions are simultaneously investigated in order to clarify their mutual relationship and applicability to music and emotions. The three-dimensional model is used to collect data regarding the dimensional approach as it encompasses both lower dimensional models. In order to obtain a large sample of unknown yet emotionally stimulating musical examples, a selection of film soundtracks was used. Soundtracks are composed for the purpose of conveying powerful emotional cues, and may serve as a relatively 'neutral' musical material in terms of music preferences and familiarity. A three-part selection process was utilized. First, 12 experts chose 360 excerpts representing Happy, Sad, Tender, Scary and Angry emotions as well as different quadrants in the 3D affect space.

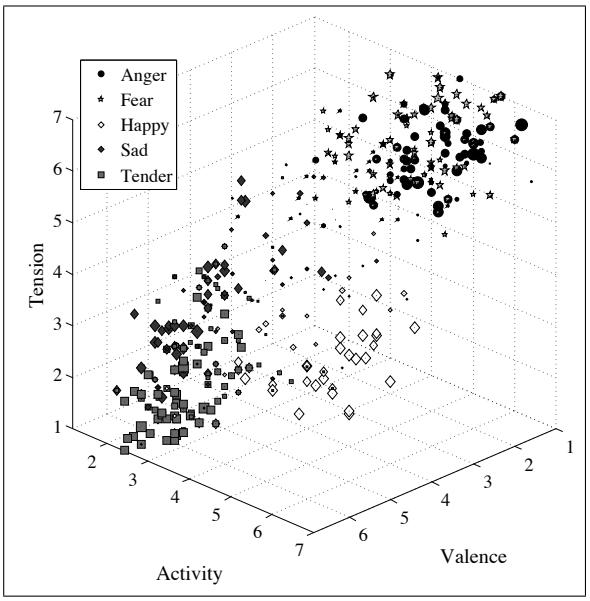

Figure 1. Average ratings of the three dimensions and basic emotions for the 360 soundtrack excerpts.

3.1 Evaluation

The expert panel (music students with extensive musical background) rated the examples, using both basic emotion concepts and dimensional ratings, on Likert scales (cf. Figure 1). Then a sampling of the 360 excerpts using both conceptual frameworks was carried out.

- For the basic emotion examples, the excerpts were categorized and ranked according to the basic emotion concept that received highest rating. From these ranked lists, the top five examples and five moderately high examples were chosen for each basic emotion (happiness, sadness, tenderness, anger and fear), yielding 50 basic emotion examples ([5 top + 5 moderate] $\times$ 5 categories).

- For the dimensional model, each dimension was sampled at 4 percentiles along its axis whilst the other two dimensions were kept constant, resulting in 60 audio examples that cover the affect space.

This set of 110 examples will be called *Soundtrack110* set hereafter. The mean duration of the excerpts was 15.3 seconds (SD 1.9 s).

In the next phase, 116 university students aged 18-42 years rated the Soundtrack110 set using both 3D set and

	3D	2D
	R^2 (β)	R^2 (β)
Happiness	.89 ($V_{.93}$,$A_{.79}$,$T_{-.35}$)	.89 ($V_{.85}$,$A_{.49}$)
Sadness	.63 ($V_{-.20}$,$A_{-.84}$,$T_{-.22}$)	.63 ($V_{-.05}$,$A_{-.69}$)
Tenderness	.77 ($V_{.33}$,$A_{-.45}$,$T_{-.58}$)	.74 ($V_{.50}$,$A_{-.51}$)
Fear	.87 ($V_{-.83}$,$A_{.07}$,$T_{.63}$)	.87 ($V_{-.90}$,$A_{.24}$)
Anger	.64 ($V_{-.52}$,$A_{.32}$,$T_{.35}$)	.68 ($V_{-.55}$,$A_{.35}$
Mean	.76	.76

Table 1. Ridge regression summary of dimensional models explaining basic emotion model categories. For instance, 89% of the variance (R^2) of Happiness can be explained with Valence (V) and Activity (A), with respective linear coefficient (β) .85 and .49.

basic emotions (on Likert scales). For the ensuing analyses, the means of the ratings across the participants were used as high consensus existed (Cronbach $\alpha > .99$ for each concept).

3.2 Basic emotions vs. dimensional ratings

As could be seen from the Figure 1, at least two emotion dimensions correlated heavily. In numerical terms, tension and valence correlate highly ($r = -.83$) and activity and tension in moderate way ($r = .57$), while valence and activity do not exhibit such a relation ($r = -.08$). The high correlation has implications in the task of constructing regression models for predicting categorical ratings based on the dimensional rating data, because multicollinear variables are problematic for standard versions of the regression. Hence we employed ridge regression since this technique is less influenced by collinearity due to the inclusion of constant variance parameter. This enables to attenuate the influence of collinearity in the calculation of the least squares optimization in regression. Ridge regression was used to predict the dimensional ratings from the categorical ratings and vice versa. The results – displayed in Tables 1 and 2 – demonstrate that the basic emotion model can more accurately explain the results obtained with the three-dimensional model than contrariwise. Nevertheless, the difference is not large (17%, the difference between the mean R^2 from the Tables 1 and 2) and this high degree of overlap between the conceptual frameworks suggests that the conceptual frameworks are highly compatible.

To further examine the validity of the three-dimensional model, its underlying coefficients of determination were also compared with the 2-dimensional circumplex model [7]. The results suggest that these two-dimensional models can explain the results obtained with the basic emotion model virtually as accurately as the three-dimensional model, with the exception of anger and tenderness (minor differences in R^2 values, see Table 1). It is worth pointing out that sadness was explained equally modestly (R^2 = .63), in comparison to other emotion categories, by all the dimensional models. This may reflect the participants' difficulty to rate the valence of sad music, for sadness in music is rarely perceived to represent an unpleasant emotion. Despite this irregularity, these analyses suggest fairly high mutual correspondence between the two conceptual frameworks and stimulus sets.

	Basic emotion model
	R^2 (β)
Valence	.97 ($H_{.35}$, $S_{-.11}$, $T_{.20}$, $F_{-.50}$, $A_{-.14}$)
Activity	.88 ($H_{.47}$, $S_{-.32}$, $T_{-.42}$, $F_{-.05}$, $A_{.36}$)
Tension	.93 ($H_{-.29}$, $S_{-.23}$, $T_{-.55}$, $F_{.18}$, $A_{.12}$)
Mean	.93

Table 2. Ridge regression summary of dimensional models explained by basic emotion model categories: Happiness (H), Sadness (S), Tension (T), Fear (F) and Anger (A).

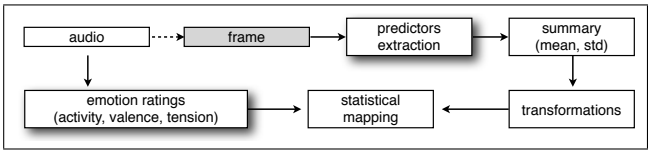

Figure 2. General design of the methodology.

4. AUDIO AND MUSIC FEATURE EXTRACTION AND TRANSFORMATION

The methodology proposed in this study is summarised in Figure 2: The *Soundtrack110* collection has been analysed using *MIRtoolbox* [23], and a set of features has been selected, explained below. We assume that a theoretical selection of features combined with a suitable data reduction techniques will result to the most parsimonious model. In addition, the features may require transformation to linearity before statistical mapping, described in the final section.

4.1 Theoretical selection of features

First, a theoretical selection is made based on the traditional categories of musical elements (rhythm, timbre, pitch, form, etc.) and by representing these categories by a few, non-redundant (non-correlating) features, in total 29. A synthetic description of the complete feature extraction process is given in Figure 3.

4.1.1 Timbre

Based on a spectrogram with a frame length of .046 s and half overlapping, three timbral descriptions are computed: centroid, spread and entropy, the latter predicting the presence of strong peaks. The mean correlation between features, computed using the *sountrack110 set*, is $r = .10$.

4.1.2 Harmony

The peaks configuration in the spectrogram enables to estimate a measure of roughness [24]. The entropy of each spectrum, collapsed into one single octave, indicates the presence of important chroma components. Or more precisely, the spectrum is turned into a chromagram, wrapped into one octave, and tonal information is computed – such

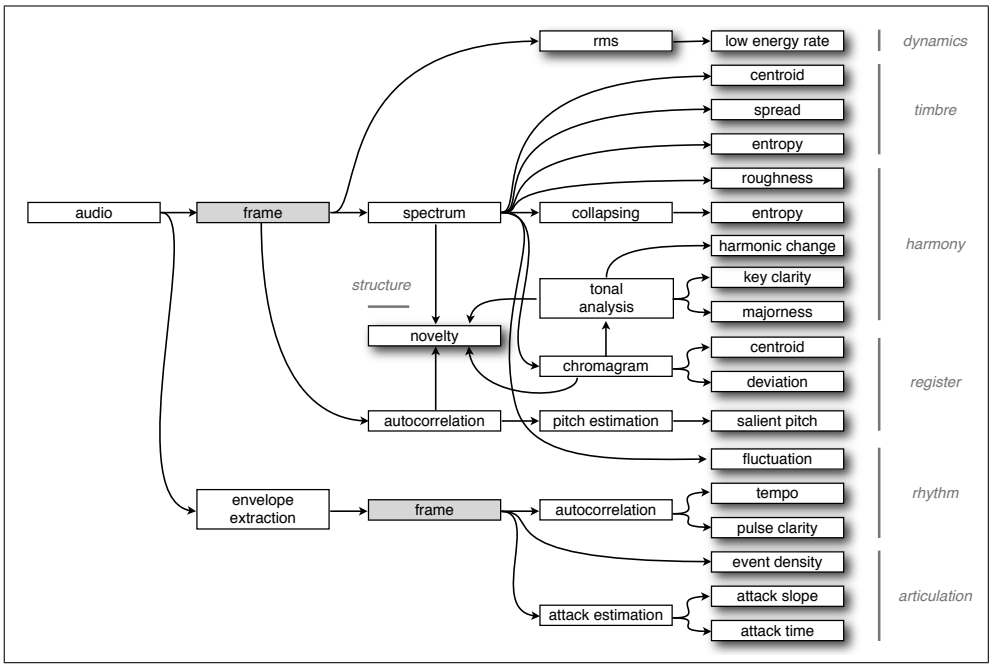

Figure 3. Flowchart of predictor extraction.

as key clarity or harmonic change [25] – based on tonal profile [26, 27]. We also designed a new measure of majorness, related to the difference of amplitude, observed on the tonal profile, between the best major score and the best minor score. For this dimension we obtain a within-feature correlation of $r = .04$

4.1.3 Register

Broad description of the localisation of pitch energy is performed through an estimation of the centroid and deviation of the unwrapped chromagram, and also in parallel a statistic description of pitch component based on advanced pitch extraction method [28]. $r = .27$

4.1.4 Rhythm

Rhythmic periodicity is estimated both from a spectral analysis of each band of the spectrogram, leading to a fluctuation pattern [29], and based on the assessment of autocorrelation in the amplitude envelope extracted from the audio. The clarity of the pulsation can also be assessed through an observation of the global characteristic of the autocorrelation function [30]. $r = .03$

4.1.5 Articulation

Onsets indicated by peaks picked from the amplitude envelope leads to the estimation of the relative amount of event density. For each successive onset, the slope and temporal duration of the corresponding attack phase is also estimated. $r = -.23$

4.1.6 Structure

The multidimensional structure of the pieces of music is estimated through the computation of novelty curves [31] based on various functions already computed such as the spectrogram, the autocorrelation function and the chromagram. $r = .85$

As a whole, the features represent the categories in a non-redundant way, as within-feature correlation is lower than .30, except for structural features.

4.2 Statistical selection of features

The second selection is based on statistical selection of relevant features, in which we compare Multiple Linear Regression (MLR) with a stepwise selection principle, Principal Component Analysis (PCA) followed by a selection of an optimal number of components, and Partial Least Squares Regression (PLS). Linear mapping via regression is known to be problematic as the predictors-to-cases ratio should be 1:10 or larger (we have 29 features, we would need at least 290 observations or more). Moreover, high number of predictors will probably be highly collinear, which is problematic for the establishment of a linear modeling of the data. Principal component analysis will eliminate the problem of collinearity, as the components are orthogonal and enables to use a low number of predictors (PCA components) in the regression. However, this data reduction method is not sensitive to the covariance between the features and the predicted data and thus may discard important features. The third technique, PLS regression [32], carries out simultaneous data reduction and maximization of covariance between features and predicted data, thus preserving any interesting correlational pattern between them. The output from the PLS is similar to PCA, individual, orthogonal components. To select the optimal number of features, Bayesian Information Criterion (BIC) was used.

	Prediction rate (R^2)		
Model	Valence	Activity	Tension
MLR	.64	.75	.67
PCA	.42	.74	.51
PLS	.70	.77	.71
MLR_λ	.66	.74	.69
PCA_λ	.51	.73	.63
PLS_λ	.72	.85	.79

Table 3. Prediction rates of the different models for circumplex model of emotions. $_\lambda$ denotes Box-Cox transformed variables.

	Prediction rate (R^2)				
Model	Angry	Scary	Happy	Sad	Tender
MLR	.46	.55	.46	.38	.38
PCA	.66	.67	.60	.59	.54
PLS	.66	.62	.61	.61	.50
MLR_λ	.56	.55	.63	.54	.45
PCA_λ	.56	.47	.53	.52	.45
PLS_λ	.70	.74	.68	.69	.58

Table 4. Prediction rates for the 5 basic emotions.

4.3 Data Transformation

To apply linear least-squares models, the distribution of the data should be approximately normal. Each feature was tested for normality (Lilliefors $p < .001$) and each non-normally distributed feature was transformed by means of Box-Cox power transform [33] by testing λ values between -2 and 2 in .1 increments and taking the one that yielded the maximal normality. Finally, all features were normalized.

5. RESULTS AND DISCUSSION

Table 3 displays the prediction rate of linear regression models using first 5 components in stepwise linear regression (MLR), and first 5 PCA components, and 2 first components from PLS, with or without data transformations ($_\lambda$). 5-fold cross-validation (80% for training, 20% for prediction) was used in all cases to avoid overfitting. In general, about 70 % of the variance in participants ratings could be predicted with features extracted from the audio. Data transformation has an important contribution to the models. MLR provides fairly successful model but it is problematic due to the serious over optimization stepwise regression does when using 29 predictors to explain 110 observations. PCA with 5 components has less power to predict the ratings but is nevertheless fairly adequate model. It suffers especially from the skewness and lack of normalization of the data. Finally, PLS (normalized) provides the highest prediction rate with only two components. The model adequacy is largely similar for basic emotions, displayed in Table 4.

The resulting predictive models vary depending on the chosen mapping method. Table 5 shows for instance the important features contributing to the perception of the categories of anger and tenderness, as predicted by the MLR method. The predictive models given by the PCA and PLS methods are less easy to represent clearly, are their underlying dimensions are formed by a high number of audio and musical features.

Anger		Tenderness	
Feature	β	Feature	β
Fluctuation peaks	-.14	RMS variance	-.44
Key clarity	-.07	Key clarity	.08
Roughness	.05	Majorness	-.08
Sp. centroid variance	-.04	Sp. centroid	-.05
Tonal novelty	.004	Tonal novelty	-.01

Table 5. Components and standardized beta weights of the MLR_λ model for two chosen basic emotions.

When mapping the dimensional ratings onto each of the five basic emotions, the regression models could explain 63 to 89 percent of the variance. No significant improvement was observed with the 3D model over the 2D model, with the exception of anger, for which adding the third dimension increased the variance explained by five per cent. When mapping basic emotions onto the emotion dimensions, even higher proportions of variance could be explained by the models, these ranged from 88 to 97 per cent. These results suggest that there is a high mutual correspondence between the two emotion spaces.

Using a five-fold cross-validation, about 70% of the variance in the participants ratings could be explained by the PLS models. The highest proportion of variance explained (85%) was obtained when predicting activity with the PLS model using transformed features. We examined the effect of the Box-Cox transform on the predictive power of the regression models. In most cases this transform improved the models significantly. This observation suggests that the distributions of the extracted features are a crucial factor in the performance of such predictive models.

The emotion prediction model has been written in *Matlab* and has been integrated into the new version (1.3) of *MIRtoolbox* [23].

6. REFERENCES

[1] X. Hu and J. S. Downie. Exploring mood metadata: relationships with genre, artist and usage metadata. In *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[2] P. N. Juslin and J. A. Sloboda. *Music and Emotion: Theory and Research*. OUP, Oxford, UK, 2001.

[3] M. Zentner, D. Grandjean, and K. Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4):494–521, 2008.

[4] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

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

[6] P. R. Farnsworth. *The social psychology of music*. The Dryden Press, 1958.

[7] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[8] M. Leman, V. Vermeulen, L. De Voogdt, D. Moelants, and M. Lesaffre. Prediction of musical affect using a combination of acoustic structural cues. *Journal of New Music Research*, 34(1):39–67, 2005.

[9] J. Skowronek, M. McKinney, and S. van de Par. Ground-truth for automatic music mood classification. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 295–296, 2006.

[10] J. Skowronek, M. McKinney, and S. van de Par. A demostrator for automatic music mood estimation. In *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[11] R. E. Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, USA, 1989.

[12] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, 19(8), 2005.

[13] U. Schimmack and R. Reisenzein. Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, 2(4):412–417, 2002.

[14] X. Hu, M. Bay, and J. S. Downie. Creating a simplified music mood classification ground-truth set. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 309–310, 2007.

[15] Y. E. Kim, E. Schmidt, and L. Emelle. Moodswing: A collaborative game for music mood label collection. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 231–236, 2008.

[16] T. Li and O. M. Ogihara. Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval*, 2003.

[17] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 462–467, 2008.

[18] D. Yang and W. Lee. Disambiguating music emotion using software agents. In *Proceedings of the International Symposium on Music Information Retrieval*, 2004.

[19] Y. Feng, Y. Zhuang, and Y. Pan. Popular music retrieval by detecting mood. In *Proceedings of the 26th annual international ACM SIGIR conference*, Toronto, 2003.

[20] L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans.on Audio, Speech, and Language Processing*, 14(1):5–18, 2006.

[21] K.F. MacDorman. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, 36(4):281–299, 2007.

[22] Y.H. Yang, Y.C. Lin, Y.F. Su, and H.H. Chen. Music emotion classification: A regression approach. In *Proc. IEEE Int. Conf. Multimedia and Expo*, pages 208–211, 2007.

[23] O. Lartillot and P. Toiviainen. MIR in matlab (II): A toolbox for musical feature extraction from audio. In *Proceedings of 5th International Conference on Music Information Retrieval*, 2007.

[24] W. A. Sethares. , *Timbre, Spectrum, Scale*. Springer-Verlag, 1998.

[25] M. B. Sandler C. A. Harte. Detecting harmonic change in musical audio. In *Proceedings of Audio and Music Computing for Multimedia Workshop*, 2006.

[26] C. L. Krumhansl. *Cognitive foundations of musical pitch*. OUP, Oxford, UK, 1990.

[27] E. Gomez. *Tonal description of music audio signal*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2006.

[28] T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8-6:708–716, 2000.

[29] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 570–579.

[30] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari. Multi-feature modeling of pulse clarity: Design, validation, and optimization. In *Proceedings of the International Symposium on Music Information Retrieval*, 2008.

[31] M. Cooper J. Foote. Media segmentation using self-similarity decomposition. In *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, volume 5021, pages 167–175, 2003.

[32] S Wold, M. Sjostrom, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130, 2001.

[33] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, Series B 26:211–246, 1964.