# MULTI-MODAL NON-PROTOTYPICAL MUSIC MOOD ANALYSIS IN CONTINUOUS SPACE: RELIABILITY AND PERFORMANCES

**Björn Schuller**[1], **Felix Weninger**[1], **Johannes Dorfner**[2]
[1] Institute for Human-Machine Communication, [2] Institute for Energy Economy and Application Technology, Technische Universität München, Germany
schuller@tum.de

## ABSTRACT

Music Mood Classification is frequently turned into 'Music Mood Regression' by using a continuous dimensional model rather than discrete mood classes. In this paper we report on automatic analysis of performances in a mood space spanned by arousal and valence on the 2.6 k songs NTWICM corpus of popular UK chart music in full realism, i. e., by automatic web-based retrieval of lyrics and diverse acoustic features without pre-selection of prototypical cases. We discuss optimal modeling of the gold standard by introducing the evaluator weighted estimator principle, group-wise feature relevance, 'tuning' of the regressor, and compare early and late fusion strategies. In the result, correlation coefficients of .736 (valence) and .601 (arousal) are reached on previously unseen test data.

## 1. INTRODUCTION

Music mood analysis, i. e., automatic determination of the perceived mood in recorded music, has been an active field of research in the last decade. For instance, it can enable browsing through music collections for music with a specific mood, or to automatically select music best suited to a person's current mood as determined manually or automatically. In this study, we describe music mood by Russell's circumplex model of affect consisting of a two-dimensional space of *valence* (pleasure–displeasure) and degree of *arousal* which allows to identify emotional tags, such as the ones used for the MIREX music mood evaluations [9], as points in the 'mood space', avoiding the ambiguity of categorical taxonomies [21]. Note that in recent research, e. g. [11], new models have been proposed specifically for music emotion, which go beyond the traditional emotion models by including non-utilitaristic or eclectic emotions. However, the valence / arousal model is an emerging standard for describing human emotions in automatic analysis [4]. Thus, from an application point of view, it is, e. g., useful for matching human emotions and music mood, such as for automatic music suggestion [16]. For automatic music mood recognition, a great variety of features have been proposed, comprising low-level acoustic, such as spectral, cepstral, or chromagram features [18], higher-level audio features such as rhythm [14], as well as textual features derived from the lyrics [12]. Early (feature-level) and late (classifier-level) fusion techniques for the acoustic and textual modalities have been compared in [8].

A first major contribution of this study is to investigate regression in the continuous arousal / valence space by single modalities (spectrum, rhythm, lyrics, etc.), and by early as well as late fusion. To briefly relate our work to recent performance studies on music mood regression: In [18] regression in a purely acoustic feature space has been investigated; [10] evaluates automatic feature selection and classifiers, but not various feature groups individually; [2] compares prediction of dimensional and categorical annotation and highlights the relevance of single features without reporting their actual performance. In summary, the majority of research still deals with classification [8, 12, 14, 19], to refer to a few recent studies. Besides, to deal with reliability issues of human music mood annotation [9], we introduce the evaluator weighted estimator (EWE) [3] to the Music Information Retrieval domain and evaluate its influence on regression performance. The EWE has been proposed as a weighted decision taking into account reliabilities of individual annotators for emotion recognition from speech [3]. Furthermore, we extend late fusion approaches such as [8] by considering the regression performance of single modalities on the development set for determination of fusion weights, in analogy to the EWE used for reaching a robust ground truth estimate.

We evaluate our system on the "Now That's What I Call Music!" (NTWICM) database introduced in [19], containing 2 648 songs annotated by four listeners on 5-point scales for perceived arousal and valence on song level. In con-

trast to some earlier work on music mood recognition such as [2], no instance pre-selection has been performed in order to simulate real-life conditions where an automatic system has to deal with non-prototypical instances, in particular those characterized by low emotional intensity [10]. Our evaluation measure is the correlation coefficient between the regression output and the estimated continuous ground truth.

The remainder of this contribution briefly describes the evaluation database (Section 2), with a particular focus on annotation reliability, and the acoustic and linguistic features used for automatic regression (Section 3). Results of extensive regression runs are given in Section 4 before concluding in Section 5.

## 2. NTWICM DATABASE

### 2.1 Data Set

For building the NTWICM music database the compilation "Now That's What I Call Music!" (U. K. series, volumes 1–69) is selected. It contains 2 648 titles — roughly a week of total play time — and covers the time span from 1983 to 2010. Likewise it represents very well most music styles which are popular today; that ranges from Pop and Rock music over Rap, R&B to electronic dance music as Techno or House. The stereo sound files are MPEG-1 Audio Layer 3 (MP3) encoded using a sampling rate of 44.1 kHz and a variable bit rate of at least 128 kBit/s as found in many typical use-cases of an automatic mood classification system.

For 1 937 of 2 648 songs in the database (cf. Section 2.3, Table 2) lyrics can automatically be collected from two on-line databases: In a first run lyricsDB, (http://lyrics.mirkforce.net/) is applied, which delivers lyrics for 1 779 songs, then LyricWiki, (http://www.lyricwiki.org/) is searched for all remaining songs, which delivers lyrics for 158 additional songs. The only manual post-processing carried out was normalization of transcription inconsistencies, e. g., markers for chorus lines, among the databases.

### 2.2 Annotation and Reliability

Songs were annotated as a whole, i. e., without selection of characteristic song parts, to stick to real world use cases – such as music suggestion – as closely as possible. Respecting that mood perception is generally judged as highly subjective [9], we decided for four labellers. While mood may well change within a song, as change of more and less lively passages or change from sad to a positive resolution, annotation in such detail is particularly time-intensive. Yet, we are assuming the addressed music type – mainstream popular and by that usually commercially oriented – music to be less affected by such variation as, for example, found in longer arrangements of classical music. In fact, this can be very practical and sufficient in many application scenarios, as for automatically suggestion that fits a listener's mood. Details on the chosen raters (three male, one female, aged between 23 and 34 years; average: 29 years) and their professional and private relation to music are provided in Table 1. As can be seen, they were picked to form a well-balanced set spanning from rather 'naive' assessors without instrument knowledge and professional relation to 'expert' assessors including a club disc jockey (D. J.). The latter can thus be expected to have a good relationship to music mood, and its perception by the audiences. Further, young raters prove a good choice, as they were very well familiar with all the songs of the chosen database. They were asked to make a forced decision according to the two dimensions in the mood plane assigning values in -2, -1, 0, 1, 2 for arousal and valence, respectively. They were further instructed to annotate according to the perceived mood, that is, the 'represented' mood, not to the induced, that is, 'felt' one, which could have resulted in too high labelling ambiguity. The annotation procedure is described in detail in [19], and the annotation along with the employed annotation tool are made publicly available [1] .

| | age, g | $\rho$ | | CC | | CC-LORO | |
|---|---|---|---|---|---|---|---|
| | | Val | Aro | Val | Aro | Val | Aro |
| A | 34, m | .828 | .749 | .827 | .763 | .678 | .456 |
| B | 23, m | .267 | .623 | .304 | .640 | -.012 | .366 |
| C | 26, m | .797 | .633 | .800 | .656 | .651 | .442 |
| D | 32, f | .797 | .717 | .819 | .733 | .640 | .474 |

**Table 1**: NTWICM Database: Raters A–D by age and g(ender), and reliability of val(ence) and aro(usal) annotation by Spearman's $\rho$ and correlation coefficient (CC) with mean (A–D), as well as CC in leave-one-rater-out (LORO) analysis.

In this study, we aim at music mood assessment in the continous domain as determined by the four raters. Thus, a consensus has to be derived from the individual labellings for valence and arousal. A continuous quantity as needed for regression is obtained as follows. As a first step, we calculated the agreement (reliability) of rater $k \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ with respect to the arithmetic mean label $\overline{l_n^{(d)}}$ for each instance $n$, $d \in \{\text{valence, arousal}\}$,

$$\overline{l_n^{(d)}} = \frac{1}{4} \sum_k l_{n,k}^{(d)} \tag{1}$$

where $l_{n,k}^{(d)} \in \{-2, -1, 0, 1, 2\}$ is the label assigned by rater $k$ to instance $n$. As a measure of reliability for each $k$, we computed the correlation coefficient $\mathrm{CC}_k$ between $(l_{n,k}^{(d)})$ and $(\overline{l_n^{(d)}})$. Results are shown in Table 1, where we also pro-

[1] http://openaudio.eu/NTWICM-Mood-Annotation.arff

vide the values for Spearman's rho ($\rho$) for reference: Notable differences between CC and $\rho$ can mainly be seen for the valence annotation by rater B.

Evidently, the reliability in terms of $CC_k$ differs among the raters – especially for valence, where it ranges from .828 (rater A, club D. J.) down to .267 (rater B). Hence, as a robust estimate of the desired ground truth mood of each instance $n$, we additionally considered the EWE [3], denoted by $l_n^{(d)}$, in further analyses:

$$l_n^{(d)} = \frac{1}{\sum_k \mathrm{CC}_k} \sum_k \mathrm{CC}_k l_{n,k}^{(d)}. \quad (2)$$

We hypothesize that the EWE provides a robust ground truth estimate especially for the NTWICM database with only four annotators, where a single 'unreliable' annotator does not simply 'average out'. Note that we refrain from reporting the agreement of the raters with the EWE, as in the EWE information about their reliability is already integrated. Furthermore, the CC of raters with the mean of *all* raters is arguably a slight overestimate of the true reliability, since the rating to be evaluated is included in the ground truth estimate. Thus, we additionally performed a 'leave-one-rater-out' (LORO) reliability analysis. Thereby for each rater $k$ the CC is calculated between $(l_{n,k}^{(d)})$ and the EWE of *all raters except* $k$. It turns out that human agreement is considerably lower when measured in a LORO fashion – partly, this can be attributed to the fact that in the LORO analysis, each ground truth estimate is made up from only three raters. Again, rater A exhibits the highest reliability whereas rater B is ranked last, both for valence and arousal (cf. Table 1).

### 2.3 Partitioning

We partitioned the 2 648 songs into training, development, and test partitions through a transparent definition that allows easy reproducibility and is not optimized in any respect: Training and development are obtained by selecting all songs from odd years, whereby development is assigned by choosing every second odd year. By that, test is defined using every even year. The distributions of instances per partition are displayed in Table 2, together with the number of instances for which lyrics are missing – it can be seen that their proportion is roughly equal for all partitions.

Once development was used for optimization of classifier parameters, the training and development sets are united for training. Note that this partitioning resembles roughly 50 % / 50 % of overall training / test in order to favor statistically meaningful findings.

## 3. FEATURES

A summary of the feature groups discussed in this study is given in Table 3. They can be roughly categorized into features derived from the lyrics (Sections 3.1, 3.2), the song meta-information (Section 3.3), and finally the audio itself (Sections 3.5, 3.4, 3.6). A detailed explanation of the features is given in [19].

| Set | # songs | # lyrics |
|---|---|---|
| Train | 690 | 515 (75 %) |
| Devel | 686 | 509 (74 %) |
| Train+Devel | 1 376 | 1 024 (74 %) |
| Test | 1 272 | 913 (72 %) |
| **Sum** | 2 648 | 1 937 (73 %) |

**Table 2**: Partitioning of the NTWICM Database, and availability of lyrics.

### 3.1 Emotional Concepts

Semantic features are extracted from the lyrics by the *ConceptNet* [13] text processing toolkit, which makes use of a large semantic database automatically generated from sentences in the Open Mind Common Sense Project [2]. The software is capable of estimating the most likely emotional affect in a raw text input, which has already been shown quite effective for valence prediction in movie reviews [20].

The underlying algorithm starts from a subset of concepts that are manually classified into one of six emotional categories (happy, sad, angry, fearful, disgusted, surprised), and calculates the emotion of unclassified concepts extracted from the song's lyrics by finding and weighting paths which lead to those classified concepts. The algorithm yields six discrete features indicating a ranking of the moods from highest to lowest dominance in the lyrics, and six continuous-valued features contain the corresponding probability estimates.

### 3.2 Linguistic Features: From Lyrics to Vectors

Linguistic features are obtained from the lyrics by text processing methods proven efficient for sentiment detection [20]. The raw text is first split into words while removing all punctuation. In order to recognize different flexions of the same word (e. g. *loved*, *loving*, *loves* should be counted as *love*) the conjugated word has to be reduced to its word stem. This is done using the Porter stemming algorithm [15].

Word occurences are converted to a vector (Bag-of-Words, BoW) representation where each component represents a word stem that occurs at least 10 times. For each song, the relative frequency of the stem is computed, i. e., the number of occurences is normalized by the total number of words in the song's lyrics. The dimensionality of the resulting feature set is 393.

[2] http://openmind.media.mit.edu/

### 3.3 Metadata

Additional information about the music is sparse in this work because of the large size of the music collection used: Besides the year of release only the artist and title information is available for each song. While the date is directly used as a numeric attribute, the artist and title fields are processed in a similar way as the lyrics (cf. previous section): Only the binary information about the occurrence of a word stem is retained. While the artist word list looks very specific to the collection of artists in the database, the title word list seems to have more general relevance with words like "love", "feel" or "sweet". In total, the size of the metadata feature set is 152.

### 3.4 Chords

For chord extraction from the raw audio data a fully automatic algorithm as presented by Harte and Sandler [6] is used. Its basic idea is to map signal energy in frequency sub-bands to their corresponding pitch class which leads to a chromagram or pitch class profile. Each possible chord type corresponds to specific pattern of tones. By comparing the chromagram with predefined chord templates, an estimate of the chord type (e. g., major, minor, diminished) can be made. We recognize the nine chord types defined in [19] along with the chord base tone (e. g. C, F, G♯). Each chord type has a distinct sound which makes it possible to associate it with a set of moods [1]: For instance, major chords often correspond to happiness, minor ones to a more melancholic mood, while diminished chords are frequently linked to fear or suspense. For each chord name and chord type, the relative frequency per song is computed and augmented by the total number of recognized chords (22 features in total).

### 3.5 Rhythm

The 87 rhythm features rely on a method presented in [17]. It uses a bank of comb filters with different resonant frequencies covering a range from 60 to 180 bpm. The output of each filter corresponds to the signal energy belonging to a certain tempo, devliering robust tempo estimates for a wide range of music. Further processing of the filter output determines the base meter of a song, i. e., how many beats are in each measure and what note value one beat has. The implementation used can recognize whether a song has duple (e. g., 2/4, 4/4) or triple (e. g., 3/4, 6/8) meter. A detailed description of the rhythm features is found in [19].

### 3.6 Spectral

Spectral features are straightforward and derived from the Discrete Fourier Transform (DFT) of the songs, which is mixed down to a monophonic signal. Then, the centre of gravity, and the second to fourth moment (i. e., standard deviation, skewness, and kurtosis) of the spectrum are computed. Finally, band energies and energy densities for the following seven octave based frequency intervals are added: 0 Hz–200 Hz, 200 Hz–400 Hz, 400 Hz–800 Hz, 800 Hz–1.6 kHz, 1.6 kHz–3.2 kHz, 3.2 kHz–6.4 kHz and 6.4 kHz–12.8 kHz, which yields a total of 24 spectral features.

| Group | Description | # |
|---|---|---|
| *Cho* rds | rel. chord freq.; # distinct chords | 22 |
| *Con* cepts | ConceptNet's mood from lyrics | 12 |
| *Lyr* ics | Bag-of-Words (BoW) from lyrics | 393 |
| *Met* a | BoW from artist, title; song date | 153 |
| *Rhy* thm | Tatum vec. (57); meter vec. (19); tatum cand.; tempo + meter estim.; tatum max, mean, ratio, slope, peak dist. | 87 |
| *Spec* tral | DFT centre of gravity, moments 2–4; octave band energies | 24 |
| *All* | Union of the above | 691 |
| *NoLyr* ics | *All* \ ( *Lyr* ∪ *Con* ) | 286 |

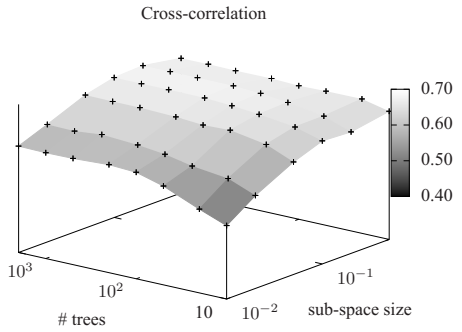
**Table 3**: Song-level feature groups and corresponding feature set sizes (#).

## 4. EXPERIMENTS AND RESULTS

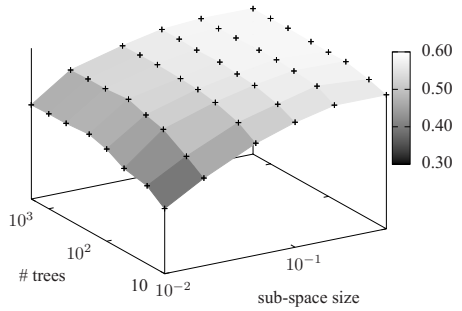
### 4.1 Setup

In our regression experiments we used ensembles of unpruned REPTrees with a maximum depth of 25 trained on random feature sub-spaces [7]. For straightforward reproducibility, we relied on the open-source implementation in the Weka toolkit [5].

We tuned the ensemble size (number of trees and sub-space size) on the development set for each combination of feature set and target (valence/arousal mean/EWE) to reflect varying sizes and complexities of the feature sets. The number of trees was chosen from $\{10, 20, 50, 100, 200, 500, 1\,000, 2\,000\}$ and the sub-space size from $\{.01, .02, .05, .1, .2, .5\}$. Results of the parameter tuning for selected feature groups can be seen in Figures 1 (a)–(b). As expected due to different sizes of the feature space, optimal parameters vary considerably. Interestingly, the best result for the *Met* feature set is obtained with 1 000 trees consisting of only 1–2 features, corresponding to a sub-space size of 1 %. Note that for the smallest feature set (*Con*), the number of possible trees is bounded by $\binom{12}{6} = 924$, so a larger number of trees will result in duplicates by the pigeon hole principle.




(a) *All*




(b) *Rhy*

**Figure 1**: Tuning of ensemble size on CC with valence EWE on development set for *All* (a) and *Rhy* (b) feature groups.

| CC | Valence | | Arousal | |
|---|---|---|---|---|
| | mean | EWE | mean | EWE |
| *Train vs. Devel* | .652 | .680 | .600 | .593 |
| *Train+Devel vs. Test* | .701 | .736 | .613 | .601 |

**Table 4**: Early fusion (*All* feature set): CC of regression on continuous valence and arousal (mean / EWE of annotators) by random sub-space learning with unpruned REPTrees. Ensemble size tuned on development set (20 % subspace, 500 trees, 2 000 for mean valence).

| | Valence | | | Arousal | | |
|---|---|---|---|---|---|---|
| | #t×sss | CC | | #t×sss | CC | |
| | | Dev | Test | | Dev | Test |
| *Cho* | 2k×.2 | .331 | .409 | 2k×.5 | .299 | .380 |
| *Con* | 500×.5 | .047 | .027 | 50×.2 | .079 | .081 |
| *Lyr* | 100×.1 | .249 | .266 | 200×.2 | .244 | .312 |
| *Met* | 1k×.01 | .209 | .241 | 500×.05 | .212 | .193 |
| *Rhy* | 100×.2 | .589 | .620 | 2k×.2 | .520 | .541 |
| *Spe* | 2k×.2 | .518 | .565 | 500×.2 | .452 | .418 |
| *NoL* | 2k×.2 | .678 | .735 | 1k×.2 | .594 | .602 |

**Table 5**: Single feature groups: CC of regression on continuous valence and arousal (EWE of annotators) by random sub-space learning with unpruned REPTrees. Number of trees (#t) and sub-space size (sss) optimized on development partition.

| CC | Valence | | Arousal | |
|---|---|---|---|---|
| | ALL | NOL | ALL | NOL |
| *Train vs. Devel* | .693 | .690 | .599 | .593 |
| *Train+Devel vs. Test* | .725 | .720 | .598 | .588 |

**Table 6**: Late fusion of modalities: CC of regression on continuous valence and arousal (EWE of annotators). REPTree ensembles for each modality parameterized as in Table 5. Fusion weights corresponding to CC on development set.

### 4.2 Results and Discussion

With the full feature set, CCs of .680 and .736 are obtained for valence on the development and test sets, respectively (cf. Table 4)—this corresponds to $R^2$ statistics of .462 resp. .542. In that case, regression on the EWE is considerably more robust than regression on the mean (absolute CC gains of .028 and .035 on development and test), which is probably due to the different reliabilities of the annotators. In contrast, for arousal, where annotator reliability is more consistent, the CC with the EWE is even slightly lower (by .007 and 0.012 absolute on development and test). In other terms, $R^2$ statistics of up to .36 (development) and .376 (test set) are obtained. For the sake of clarity, we will exclusively report on CC with the EWE in the following discussion.

Analysis of single feature groups (Table 5) reveals that spectral and rhythm features contribute most to the regression performance (CCs of .620 and .565 with the valence EWE on test). Chords (CC of .409) are in the mid-range while lyrics, meta information and concepts lag behind (CCs of .266, .241, .027). The same ranking of feature groups is obtained when considering the CC with the arousal EWE. We conclude that the feature groups that enable robust regression can be obtained directly from the audio (chords, spectral and rhythm information), and thus in full realism—though lyrics likely contribute to the annotation since the annotators were not explicitly told to ignore lyrics and all of them are experienced English speakers. In fact, the CC on the test set by the NoLyrics feature set (.735) is only slightly lower than that with the full feature set (.736).

The noticeable differences between the reliability of different modalities motivate a late fusion technique where the fused prediction is a weighted sum of the predictions of unimodal regressors. Thereby weights correspond to the individual regressors' CC on the development set, analogously to the EWE (Eqn. 2). Results obtained by this technique are shown in Table 6. On the development set, early fusion (cf. Table 4) is clearly outperformed for both recognition of valence (CC of .693 vs. .680) and arousal (CC of .599 vs. .593). However, this effect is almost reversed on the test set, where a CC of .725 as opposed to .735 (early fusion) is obtained for valence; results are similar for arousal. The latter result cannot be fully explained by overfitting fusion

weights on the development set, as there is no considerable mismatch between the reliabilities on the development compared with the test set.

## 5. CONCLUSIONS

We analyzed regression of music mood in contiuous dimensional space. Particular emphasis was laid on realism in the sense of automatically retrieving textual lyric information automatically from the web and by chosing a music database that is well defined in its own: 69 consecutive double CDs without pre-selection of high annotator agreement cases. As expected, the observed performances are clearly below the ones reported in studies on prototypical examples such as [2], yet in line with other studies on real-life data sets [10, 21]. To establish a reliable gold standard, i. e., ground truth, we proposed the usage of the evaluator weighted estimator. The best individual feature group were rhythm features based on comb-filter banks. In future work we will address unsupervised and semi-supervised learning for music mood analysis to exploit the huge quantities of popular music available on the internet.

## 6. REFERENCES


[1] W. Chase. *How Music REALLY Works!* Roedy Black Publishing, Vancouver, Canada, 2nd edition, 2006.

[2] T. Eerola, O. Lartillot, and P. Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proc. of ISMIR*, pages 621–626, Kobe, Japan, 2009.

[3] M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Proc. of ASRU*, pages 381–385, 2005.

[4] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey. In *Proc. International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE)*, Santa Barbara, CA, 2011. IEEE.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[6] C. A. Harte and M. Sandler. Automatic chord identification using a quantised chromagram. In *Proc. of the 118th Convention of the AES*, May 2005.

[7] T. K. Ho. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:832–844, 1998.

[8] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. Joint Conference on Digital Libraries (JCDL)*, pages 159–168, Gold Coast, Queensland, Australia, 2010.

[9] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In *Proc. ISMIR*, pages 462–467, Philadelphia, USA, 2008.

[10] A. Huq, J. P. Bello, and R. Rowe. Automated Music Emotion Recognition: A Systematic Evaluation. *Journal of New Music Research*, 39(3):227–244, 2010.

[11] P. N. Juslin and J. A. Sloboda, editors. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, New York, 2010.

[12] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Proc. International Conference on Machine Learning and Applications*, pages 688–693, Washington, DC, USA, 2008.

[13] H. Liu and P. Singh. ConceptNet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

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

[15] M. F. Porter. An algorithm for suffix stripping. *Program*, 3(14):130–137, October 1980.

[16] S. Rho, B.-J. Han, and E. Hwang. SVR-based music mood classification and context-based music recommendation. In *Proc. ACM Multimedia*, pages 713–716, Beijing, China, 2009.

[17] E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustic Society of America*, 103(1):588–601, January 1998.

[18] E. M. Schmidt, D. Turnbull, and Y. E. Kim. Feature selection for content-based, time-varying musical emotion regression. In *Proc. of MIR*, pages 267–274, Philadelphia, Pennsylvania, USA, 2010.

[19] B. Schuller, J. Dorfner, and G. Rigoll. Determination of non-prototypical valence and arousal in popular music: Features and performances. *EURASIP Journal on Audio, Speech, and Music Processing, Special Issue on Scalable Audio-Content Analysis*, 2010(ID 735854):19 pages, 2010.

[20] B. Schuller and T. Knaup. Learning and Knowledge-based Sentiment Analysis in Movie Review Key Excerpts. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, volume 6456 of *LNCS*, pages 448–472. Springer, Heidelberg, 2010.

[21] 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*, 16(2):448–457, 2008.