# AUTOMATIC AND MANUAL ANNOTATION OF TIME-VARYING PERCEPTUAL PROPERTIES IN MOVIE SOUNDTRACKS

**Vedant Dhandhania**
Fraunhofer IDMT, Germany
Manipal Institute of Technology, India
vedant1990@gmail.com

**Jakob Abeßer, Anna Marie Kruspe, Holger Großman**
Fraunhofer IDMT, Germany
[abr,kpe,grn]@idmt.fraunhofer.de

## ABSTRACT

In this paper, we present an automated tool as a part of the *SyncGlobal* project for time continuous prediction of loudness and brightness in soundtracks. The novel *AnnotationTool* is presented where manual time continuous annotations can be performed. We rate well-known audio features to represent two perceptual attributes—loudness and brightness. A regression model is trained with the manual annotations and the acoustic features in order to model both attributes. Five different regression methods are implemented and their success in tracking the two perceptions is studied. A coefficient of determination ($R^2$) of 0.91 is achieved for loudness and 0.35 for brightness using Support Vector Regression (SVR) yielding a better performance than Friberg et al. [1].

## 1. INTRODUCTION

The development of automated tools for recognizing various emotions within music has gained the interest of many research groups. In the last years, an increasing amount of publications tackled the modeling of semantic properties such as the emotion dimensions valence and arousal as well as their temporal progression over time. Inspired by their success, we aim to develop a system for tracking important perceptual attributes such as loudness, brightness, tension, valence, and arousal over time. This paper presents first results of the temporal modeling of loudness and brightness.

The main focus of this work is on movie soundtracks. They play an important role in supporting the story line and the emotional progression of a film [2]. Both the audio signal and the visual signal form a cohesive force and enhance the viewer's experience by conveying the movie's atmosphere and scenery.

Loudness is a very fundamental perception in music. The measurement of short-term and long-term perceived loudness is presented in [3]. Loudness can be related to tension, anger, fear, and joy depending on the dynamics of the piece [4]. The perceived brightness mainly depends on the spectral energy distribution of the audio signal and the relationship between high and low spectral energy in particular. Brightness is often associated with contentment and exuberance [5].

The presented research is part of the *SyncGlobal* project—a 2-year research project with a special focus on multimodal synchronization between video and audio content. In order to allow users to retrieve suitable audio content for a given video recording, novel search query types need to be implemented that incorporate the temporal development of perceptual attributes. One such search query could be: “Find a music segment of approx. 20 s length with slowly increasing loudness, low arousal, and medium brightness.” In this paper, we aim at modeling the temporal behavior of the perceptual attributes loudness and brightness and finding segments with similar tendencies. These segments can then be used to handle the above-mentioned search queries.

The contributions of this paper are two-fold: First, the novel version of the *AnnotationTool* software is applied to initially capture manual annotations and later verifying (or modifying) the automatically predicted annotations. Second, we compare various configurations of audio features and regression systems and investigate optimal parameter settings such as the optimal window size.

This paper is organized as follows. After a brief review of related work in Sect. 2, we present our novel approach for time-continuous modeling of perceptual attributes in Sect. 3. The evaluation of the proposed system is detailed in Sect. 4, starting with an overview over the *AnnotationTool* software and the data acquisition process. Then, two different experiments towards modeling of loudness and brightness and finding the optimal window-size are explained in Sect. 4.3 and Sect. 4.4 and the results are discussed. Sect. 5 gives a short conclusion of our work.

## 2. RELATED WORK

Labeling and annotating music is an integral part in developing any automated system for the prediction of various perceptions or moods in music. Up to now, a major hindrance in music information retrieval (MIR) research has been the collection of high-quality semantic annotations. Commonly, these annotations are based on complete songs or time segments. Time-continuous annotations on a frame-level still is a challenging task due to time and cost issues. In order to obtain human annotations, hand-labeling music [6, 7] is a common approach. However, interactive and effective multiplayer games such as *MoodSwings* [8] and *Listen Game* [9] have also been proposed. Users are asked to suggest new words to describe music or mark labels in these games.

Some online companies like Last.fm [10] allow all users to tag songs with relevant musical terms. However, the tags might not be trustworthy and is also difficult to cluster. In this publication, we use the *AnnotationTool*—previously introduced in [11]—for data acquisition and present extensions for time-continuous annotations and event based annotations in Sec. 4.1.

The annotations obtained are needed to train statistical models based on acoustic features representing the perceptual attributes. Regression is a well known statistical method used in music emotion recognition [12,13]. These methods often outperform standard classification methods [14]. In this paper, we compare the performance of 5 different regression methods for the prediction of loudness and brightness values based on previously extracted audio features.

Friberg et al. [1] identify suitable features for 9 different perceptions such as speed, dynamics (loud / soft), pitch, and brightness. The authors use step wise regression to predict the perceptual features and report a performance of $R^2 = 0.29$ for brightness and a $R^2 = 0.67$ for dynamics [1] in ringtones. The annotators in their study showed a high agreement concerning the annotation of both attributes: Cronbach alpha values of 0.93 for dynamics (soft / loud) and 0.88 for brightness were reported.

## 3. NEW APPROACH

We explore four methods of regression namely Multiple Linear Regression (MLR), Partial Least Square Regression (PLSR), Support Vector Regression (SVR) [2] and Robust Regression (RR) to predict loudness and brightness values in different time frames. Our aim is not only to identify suitable acoustic features to model loudness or brightness but furthermore to analyze the optimal window size for calculating these features. In this section, we describe our system and the applied methodologies and proceed to the evaluation in the next section.

### 3.1 System Overview

Fig. 1 illustrates our approach for generating automatic prediction of perceptual music properties. First, audio features are extracted from soundtrack excerpts in the training set. These features and the corresponding ground truth annotations are used to train a regression model. Unlabeled music data is processed similarly: Audio features are computed and attribute values are predicted on a frame-wise basis. The predicted values are post-processed by a smoothing filter, which later allows to detect segments of similar trajectories such as increasing or decreasing values. These segments can be visualized, verified, and manually corrected by human experts in the *AnnotationTool* software as described in Sect. 4.1.

| Feature Name | Meaning |
|---|---|
| ASE | Audio Spectral Envelope |
| CENT | Spectral Centroid |
| Chroma | Energy in Pitch Class |
| EPCP | Enhanced Pitch Class Profile |
| LogLoud | Bandwise Log Loudness |
| NormLoud | Bandwise Normalized Loudness |
| OSC | Octave-based Spectral Crest |
| SCF | Spectral Crest Factor |
| SFM | Spectral Flatness Measure |
| ZCR | Zero-crossing rate |

**Table 1**. List of audio features used to represent loudness

#### *3.1.1 Pre-processing*

To the best knowledge of the authors, no dataset containing time-continuous annotations of perceived loudness and brightness in movie soundtracks has been published so far. Therefore, we assembled a novel database of 20 soundtrack clips, each about 45 to 90 seconds in length. The soundtracks were selected from popular films. The genre of the movies varied from horror, sci-fi, action, to romance. Each of the soundtrack excerpts was converted to a uniform format (44.1 kHz sampling rate, 16 bit quantization, mono channel PCM WAV) and normalized to the same volume level.

#### *3.1.2 Feature Extraction*

We are interested in features that correlate with the perceived loudness and brightness in music. Since we aim to track both attributes over time, we focus on a frame-wise feature extraction. Table 1 lists the audio features applied in this publication. Most of the features were used for modeling loudness (compare Sect. 4.3)

Spectral centroid is well known to measure brightness. The MIR toolbox [15] has been used in [1] to calculate acoustic features representing brightness. Mainly two audio features in the toolbox —*mirbrightness* and *mirmode* —yielded the highest correlation [3]. The default parameters of the two features were used in all the experiments. However, we calculate the features using different window sizes and compare the performance of the regression system as described in Sec. 4.4.

### 3.2 Regression

Regression was used to model several aspects in music including emotion and several perceptual attributes. We use regression for the prediction of loudness and brightness due to its reliable prediction performance and easy optimization [16]. We compare the performance of MLR, RR, PLSR and SVR.

Both the feature extraction and the attribute annotation using the *AnnotationTool* are based on the same time resolution. The $i$-th time-frame is represented by a feature vector $\boldsymbol{x}_i \in \mathbb{R}^{N_{\text{feat}}}$ and a scalar attribute annotation $y_i \in \mathbb{R}$ on

[1] This attribute roughly corresponds to loudness as investigated in this paper.

[2] Both the $\nu$-SVR and $\epsilon$-SVR methods are considered.

[3] The *mirbrightness* feature is closely related to spectral centroid

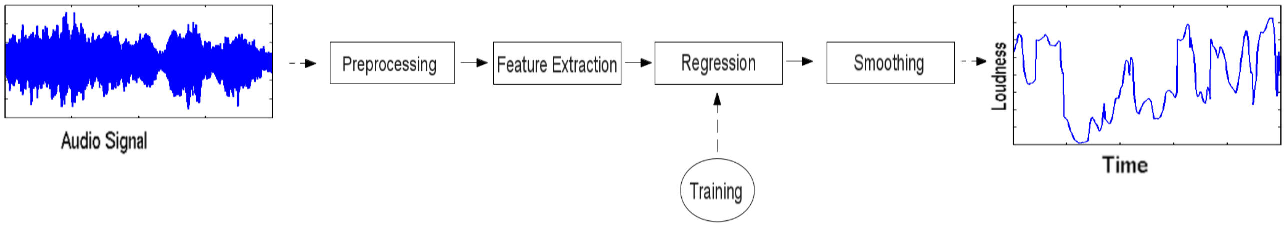

**Figure 1**. Flowchart of the System

a scale of $-1 \leq y_i \leq 1$.

The MLR algorithm models the linear relationship between a dependent variable and one or more independent variables. The RR assigns a weight to each data point within the training data, which causes outliers to have a smaller influence on the regression model. In the PLSR algorithm, a smaller number of less correlated predictor variables is derived from a linear combination of the initial feature dimensions. The optimal model order is determined by minimizing the Residual Sum of Errors. For the SVR, we compare $\nu$-SVR and $\epsilon$-SVR as provided by the LibSVM toolbox [17] with a linear kernel function. The optimal parameters $\{C, \gamma, \nu\}$ for the $\nu$-SVR and $\{C, \gamma, \epsilon\}$ for the $\epsilon$-SVR are selected via grid-search as proposed in [17] by minimizing the mean squared error (MSE) value. For more details on the regression methods, see for instance [17] and [18].

The number of cross validation (CV) folds was kept at 10 for all experiments. The frames from the same songs were never used for training and testing at the same time to avoid the well known "album effect" [19].

### 3.3 Smoothing

The values predicted by the regression algorithm are usually noisy. Smoothing the predicted data provides a better visual appeal and understanding of the predicted value trajectory over time. Once the prediction is smoothed, it is fed back into the *AnnotationTool* where it can be verified or changed accordingly. As illustrated in Fig. 2, time segments showing approximated linear functions are used for visualization. Various methods such as Savitzky-Golay [20] filter, moving-average (MA) filter, and smoothing via folding with a Gaussian kernel were compared. The noisy regression predictions were not completely removed using Savitzky-Golay filters. However, the simplest method—moving-average filtering— provides fully satisfying results for the given task.

## 4. EVALUATION

In this section, we first describe the *AnnotationTool* software in Sect. 4.1 and then proceed to our data collection methods. Finally, we evaluate different audio features for modeling brightness and loudness and then compare the performance of various regression methods in Sec. 4.4.

### 4.1 Annotation Tool

The *AnnotationTool* is a software that was developed at Fraunhofer IDMT during the ongoing research for the *Sync-Global* project. Its purpose is the annotation of music files with a large multitude of musical properties. The software is written in C++, using the Qt framework for GUI development, and it is available for Windows, Linux, and OS X. The tool displays different representation of the audio file such as the waveform and the spectrogram as well as all human annotations. It allows the user to edit these annotations and create new ones.

The annotation is based on the manual or automatic time segmentation of the audio track. This allows the user to annotate the song based on its temporal structure. The software allows for multi-domain labeling [21], e.g., semantic annotations of the same attributes related to different musical domains such as timbre, harmony, or rhythm. Users may define their annotation schema very flexibly using an XML configuration file. The tool in its newest version allows for three types of annotations:

- **Discrete annotations**: For discrete annotations, the user has to pre-define a number of possible textual or discrete values that can be assigned to different time segments. For example, "Pop", "Rock", and "Urban" could be applied for annotating the musical genre.

- **Gradual annotations**: Gradual annotations are based on continuous values, they are useful to annotate certain strengths of a property. For instance, gradual annotations can be used for the emotional dimensions "valence" or "arousal". We use a generalized value range between -1 and 1, which for instance could correspond to "very low valence" and "very high valence". Gradual annotations allow to capture a certain change of an attribute over time.

- **Event annotations**: Events are short occurrences of properties, e.g. a drum roll or a sudden noise. The *AnnotationTool* allows the user to mark these appearances, either with or without a duration.

The resulting annotations are saved in XML format and can then be used for further tasks such as regression or classification model training or similarity searches. A screenshot of the tool is shown in figure 2. A gradual annotation is visible in the center window.

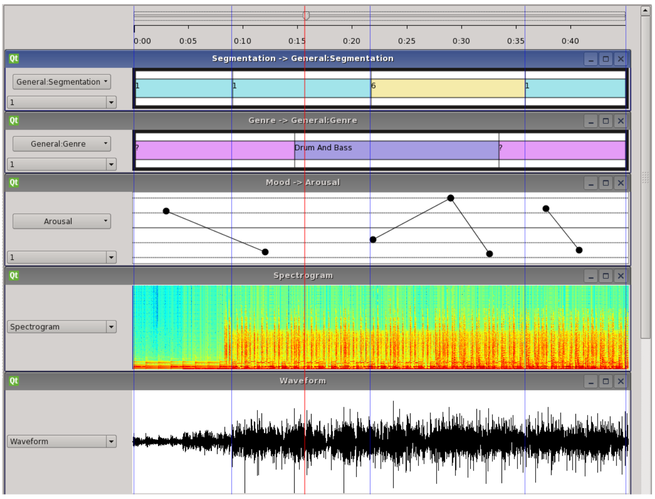

**Figure 2**. A screenshot of the *AnnotationTool*

### 4.2 Data Collection

In our experiments, we used the *AnnotationTool* to annotate time-continuous progressions of the properties loudness and brightness. All soundtrack excerpts in our dataset were annotated by 5 musical experts in terms of perceived loudness and brightness on a scale between -1 and 1. A high value on this scale corresponds to high loudness or high brightness, respectively. The following considerations were made to obtain annotations: The 5 musical experts annotated the soundtracks for loudness and brightness in different sessions. They were allowed to review their annotations and to change them at a later stage. A maximum of 20 minutes was allotted for every annotation session. After the annotation of each excerpt, the users were asked to review their annotation before proceeding to the next. All experts were made to listen to the soundtracks at the same volume with the same hearing headphones. After the annotation process, the agreement between the experts was investigated for each annotated soundtrack. The frames where the individual annotations of at least 2 annotators differed by a value of at least 0.25 (12.5% of range of -1 to 1) were discarded. After this outlier detection, the remaining annotations were averaged in each time frame. The loudness annotations of 5 experts of a soundtrack excerpt (about 90 seconds) from the movie "Buffy The Vampire Slayer" is shown in Fig. 3. Fig. 4 depicts the brightness annotations of 5 experts from a soundtrack excerpt (about 50 seconds) of the movie "The Lord of the Rings: The Fellowship of the Ring." Interestingly, we found a high Cronbach's alpha of 0.93 for the loudness annotations and 0.92 for the brightness annotations. Comparable values of 0.93 and 0.88 were found for the same two attributes in [1]. This shows that loudness and brightness annotations across annotators have very similar tendencies to one another.

### 4.3 Experiment 1

For predicting loudness, we first investigated the correlation between the annotations of the perceived loudness and the computed features in Table 2 by calculating the correlation coefficient $r$. The most correlated features were that of LogLoud and NormLoud across all bands as well of some bands of the OSC.

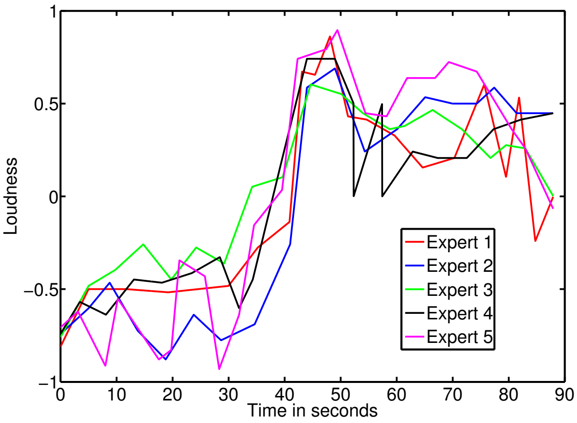

**Figure 3**. Segmented Loudness Annotations of 5 experts from a soundtrack clip of "Buffy The Vampire Slayer"

We use the 'LogLoud' and 'NormLoud' feature vectors for representing our audio signal as almost all bands in this feature have a correlation of $r \geq 0.5$. The regression methods described in Sect. 3.2 are used for predicting loudness. Tab. 3 presents the regression results of the 5 different methods in terms of the coefficient of determination $R^2$ and the root-mean squared error (RMSE). Very high $R^2$ values were observed for almost all regression methods. SVR outperforms the other regression methods. While MLR, RR and PLSR have almost the same performance.

### 4.4 Experiment 2

As mentioned earlier, the MIR toolbox was used to calculate acoustic features to represent brightness. The toolbox calculates features that estimate the brightness of the audio signal. We window our audio signal and estimate the brightness for each window frame. We consider win-

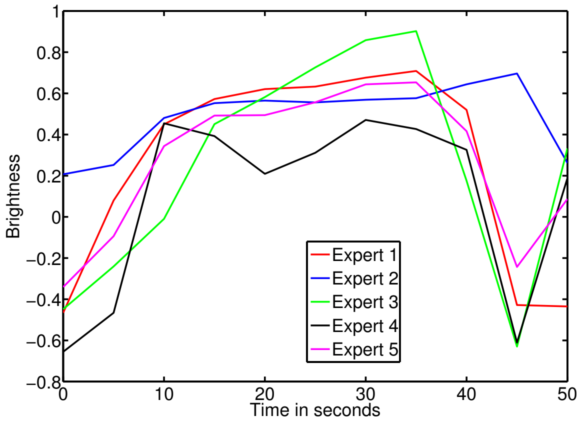

**Figure 4**. Segmented Brightness Annotations of 5 experts from a soundtrack clip of “The Lord of the Rings: The Fellowship of the Ring”

| Audio Feature | Feature Dimension | $r$ value (min) | $r$ value (max) |
|---|---|---|---|
| ASE | 14 | 0.07 | 0.47 |
| CENT | 16 | -0.06 | 0.31 |
| CHROMA | 12 | -0.03 | 0.38 |
| EPCP | 36 | -0.09 | 0.38 |
| LogLoud | 12 | 0.45 | 0.77 |
| MFCC | 12 | -0.27 | 0.18 |
| NormLoud | 12 | -0.58 | 0.50 |
| OSC | 16 | -0.24 | 0.76 |
| SCF | 16 | -0.16 | 0.22 |
| SFM | 16 | -0.40 | 0.25 |
| ZCR | 1 | -0.29 | -0.29 |

**Table 2**. Cross-correlations between computed features and loudness; No of observations, N=8000.

dow sizes as small as 1 second up till 8 seconds (in steps of 1 second) and analyze the performance. Smaller window sizes (less than 1 second) were also assessed leading to poor performance. Fig. 5 and Fig. 6 illustrates the $R^2$ and RMSE for different window-sizes respectively. The best results for the individual regression methods is shown in Tab. 4. The SVR outperforms MLR, RR, and PLSR again. The highest $R^2$ values were obtained for window sizes of 1 to 3 seconds. It should be noted that as the window size increases the amount of data also proportionately decreases. This might have a small effect on the regression results as they are dependent on the amount of training and testing data.

| Regression Method | $R^2$ | RMSE | Window size(ms) |
|---|---|---|---|
| MLR | 0.85 | 0.16 | 30 |
| RR | 0.85 | 0.17 | 30 |
| PLSR | 0.85 | 0.17 | 30 |
| $\nu$_SVR | 0.91 | 0.13 | 30 |
| $\epsilon$_SVR | 0.91 | 0.13 | 30 |

**Table 3**. Results of the 5 types of regression for loudness

| Regression Method | $R^2$ | RMSE | Window size (ms) |
|---|---|---|---|
| MLR | 0.29 | 0.27 | 2000 |
| RR | 0.29 | 0.27 | 2000 |
| PLSR | 0.29 | 0.27 | 1000 |
| $\nu$_SVR | 0.35 | 0.26 | 1000 |
| $\epsilon$_SVR | 0.35 | 0.26 | 1000 |

**Table 4**. Best results of the 5 types of regression for brightness

## 5. CONCLUSION

In this paper, we proposed a novel method of annotating movie soundtracks. Although the experiments were performed using movie soundtracks, the same methodology can be applied for estimating loudness or brightness in a more general set of musical genres. Work has already begun on automatic time-continuous tracking of other high level perceptual attributes such as tension, happiness, excitement, and fear in movie soundtracks. Predictions of such perceptions or moods would have wide applications. In order to make the developed systems more robust, we aim to deploy more musical experts and expand our database in the future.

## 6. ACKNOWLEDGMENTS

This research work is a part of the SyncGlobal project. It is a 2-year collaborative research project between piranha womex AG from Berlin and Bach Technology GmbH, 4FriendsOnly AG, and Fraunhofer IDMT in Ilmenau, Germany. The project is co-financed by the German Ministry of Education and Research in the frame of an SME innovation program (FKZ 01/S11007).The authors would like to thank Dr. Scott Beveridge for all the fruitful discussions.A special thanks to Estefanía Cano, Pedro Mercado, Jens Hansen, Martin Bocksberger, and Sebastian Maulbeck for participating in the experiments.

## 7. REFERENCES

[1] A. Friberg and A. Hedblad, “A comparison of perceptual ratings and computed audio features,” in *Proceedings of the 8th Sound and Music Computing (SMC) Conference, Italy*, 2011, pp. 122–127.

[2] A. Cohen, *Music as a source of Emotion in Film*. Oxford University Press, Music and Emotion, 2001, pp. 248–268.

[3] B. Glasberg and B. Moore, “A model of loudness applicable to time-varying sounds,” *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.

[4] A. Gabrielsson and E. Lindström, *The influence of mu-*

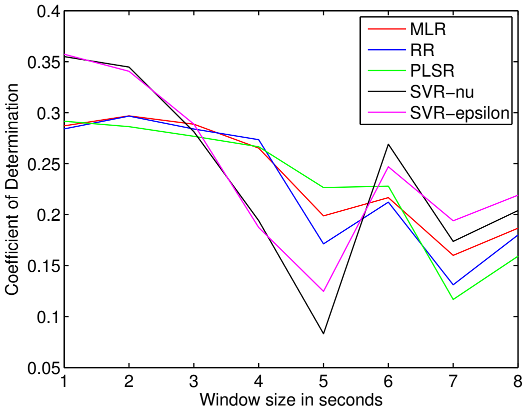

**Figure 5**. $R^2$ values of the 5 regression methods with different window sizes for brightness prediction

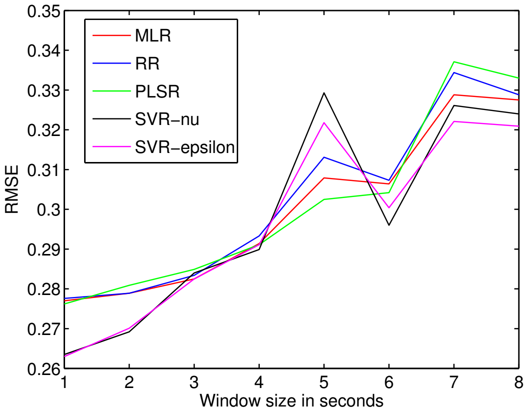

**Figure 6**. RMSE values of the 5 regression methods with different window sizes for brightness prediction

*sical structure on emotional expression*. Oxford University Press, 2001, pp. 223–248.

[5] 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*, vol. 14, no. 1, pp. 5 – 18, jan. 2006.

[6] M. Goto, "AIST annotation for the RWC music database," in *Proceedings of the 7th International Conference on Music Information Retrieval, Victoria, Canada*, 2006, pp. 359–360.

[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] Y. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), Philadelphia, USA*, 2008.

[9] D. Turnbull, R. Liu, L. Barrington, and G. R. G. Lanckriet, "A game-based approach for collecting semantic annotations of music." in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna, Austria*, 2007, pp. 535–538.

[10] Last.fm. Last accessed on 03-30-2012. [Online]. Available: http://www.last.fm

[11] P. Woitek, P. Bräuer, and H. Großmann, "A novel tool for capturing conceptualized audio annotations," in *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*. ACM, 2010, pp. 15:1–15:8.

[12] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio." in *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR),Utrecht, Netherlands*, 2010, pp. 465–470.

[13] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, "A regression approach to music emotion recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 448 –457, feb. 2008.

[14] B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression," in *Proc. of the 10th Intl. Society for Music Information Conf., Kobe, Japan*, 2009.

[15] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, September,2007, pp. 127–130.

[16] A. Sen and M. Srivastava, *Regression analysis*, ser. Springer texts in statistics. New York: Springer, 1990.

[17] C.-w. Hsu, C.-c. Chang, and C.-j. Lin, "A Practical Guide to Support Vector Classification," vol. 1, no. 1, pp. 1–16, 2010.

[18] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," vol. 12, pp. 1207–1245, 2000.

[19] Y. E. Kim, D. S. Williamson, and S. Pilli, "Towards quantifying the "album effect" in artist identification." in *Proceedings of the 7th International Conference on Music Information Retrieval, Victoria, Canada*, 2006, pp. 393–394.

[20] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, pp. 1627–1639, 1964.

[21] H. M. Lukashevich, J. Abeßer, C. Dittmar, and H. Großmann, "From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification." in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*. International Society for Music Information Retrieval, 2009, pp. 459–464.