MODELING MUSICAL MOOD FROM AUDIO FEATURES AND LISTENING CONTEXT ON AN IN-SITU DATA SET

Diane Watson
University of Saskatchewan
diane.watson@usask.ca

Regan L. Mandryk
University of Saskatchewan
regan.mandryk@usask.ca

ABSTRACT

Real-life listening experiences contain a wide range of music types and genres. We create the first model of musical mood using a data set gathered in-situ during a user's daily life. We show that while audio features, song lyrics and socially created tags can be used to successfully model musical mood with classification accuracies greater than chance, adding contextual information such as the listener's affective state or listening context can improve classification accuracy. We successfully classify musical arousal with a classification accuracy of 67% and musical valence with an accuracy of 75% when using both musical features and listening context.

1. INTRODUCTION

Musical mood – the emotion expressed by a piece of music – is conveyed to a listener through a variety of musical cues in the form of auditory features. These auditory features (e.g., mode, rhythm, articulation, intensity and timbre of a musical track) have previously been used to model musical mood with fairly good classification results [1-2]. However, current high performing models of musical mood have two main problems: first, music is constrained to a single genre (usually Western classical or Western popular); and second, the data is collected and labeled in laboratory contexts. Previous work has shown that the data sets used in previous research modeling musical mood do not correspond to real-life listening experiences in a number of ways [3]. First, people listen to music of various genres in their daily life; second, music is listened to as part of social activities or in a public venue; third, music is attended to as a secondary activity while driving, working, or exercising; finally, people listen to music for various reasons, such as to relax, to entertain, or to influence emotion [3].

Previous high-performing musical mood models, based on data from a single genre and gathered in a laboratory setting may fail when applied to data sets gathered in daily life. Systems implementing these previous models – such as music recommender systems – may also fail when using data collected from real-life listening experiences, which may lead to negative user experiences. However, musical mood classifiers built on a broad data set, containing several genres and labeled during real-life activities rather than in a laboratory, may be unusable in real systems if they yield weak classification results.

To solve the problem of building good musical mood classifiers that are effective for real-life listening experiences, we include context-sensitive features in addition to the previously used auditory features. Our data set of real-life listening experiences was gathered through an experience-sampling study using smartphones. Participants were handed phones for a period of two weeks. Phones would randomly poll the user about once per hour and ask them to fill out a survey collecting musical mood, the listener's affective state and the context of the listening experience. Genre, title and artist were optionally captured. Previous analysis of our data set shows that real-life listening experiences are far from the homogenous data sets used in current models, and cover a wide range of genres, artists, and songs [3]. In the present paper, we used our naturally-gathered data set to model musical mood using Bayesian networks and feature sets including musical features (audio features, song lyrics, socially-created tags), the affective state of the listener, and listening context. Listening context included reason for listening, activity, location, social company, level of choice over the song and mental associations.

In this paper we make two main contributions. First, we successfully model musical mood from a data set gathered in-situ during a user's daily life; we are the first to do so. Second, we show that while musical features (audio features, song lyrics and socially created tags) can successfully model musical mood with classification accuracies better than chance, adding contextual information, such as the listener's affective state or the listening context of the musical experience, can further improve classification accuracies. We successfully classify musical arousal with a classification accuracy of 67% and musical valence with an accuracy of 75% when using both musical features and listening context.

2. RELATED WORK

2.1 Affective State

It is well documented that music can induce specific affective experiences in the listener. Affective state, or the emotion or mood a person is experiencing, can be described using either a categorical or dimensional approach. The categorical approach breaks emotions into

discrete labeled categories (e.g., happiness, fear, joy) [4]. In contrast, the dimensional approach, which we use in this paper, represents affective state using two orthogonal dimensions: arousal and valence [5]. Arousal can be described as the energy or activation of an emotion. Low arousal corresponds to feeling sleepy or sluggish while high arousal corresponds to feeling frantic or excited. Valence describes how positive or negative an emotion is. Low valence corresponds to feeling negative, sad or melancholic and high valence to feeling positive, happy or joyful. Most categorical emotions can be described by Arousal-Valence (A-V) space (e.g., angry in Figure 1).

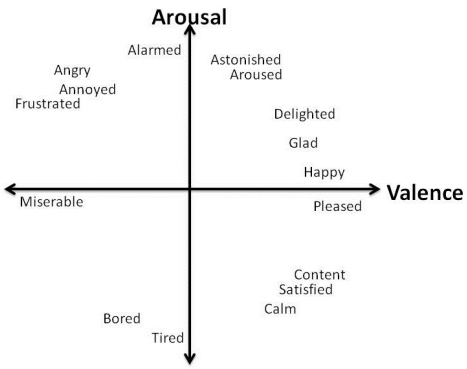

Figure 1 shows A-V space labeled with several of the categorical emotions.

2.2 Musical Mood

Musical mood, the emotion expressed by a piece of music, is to some degree perceived consistently across different listeners and even different cultures. Studies by Juslin and Sloboda have shown that listeners of different musical training classify musical mood into the same categories [6]. Fritz et al. found that the Mafa natives of Africa – without any exposure to Western music – categorized music into the same three basic emotional categories as Westerners [7]. Musical mood is frequently measured in arousal and valence [8] and we have used this approach in this paper. It should be noted that the affective state induced in the listener is not necessarily the same as the musical mood of the music [9], [10]. For example, an individual who is herself feeling frustrated (i.e., mood of the listener) can still perceive a piece of music as calm (i.e., musical mood).

2.3 Musical Mood Classification

Musical mood can be manually categorized by the listener, but researchers have also algorithmically classified musical mood using audio features extracted from the musical track. Work by Juslin [11] has identified seven musical features that are important in the interpretation of musical mood. He asked performers to play the same musical scores in such a way as to express four different musical moods (anger, sadness, happiness and fear) and then had listeners rate the strength of each mood. He found that performers and listeners used the same features to identify each mood, but weighted their importance differently. These features are:

- *Mode:* Mode refers to the key of the music. (e.g. A-)
- *Rhythm:* Rhythm is the pattern of strong and weak beat. It can be described through speed (tempo), strength, and regularity of the beat.
- *Articulation:* Articulation refers to the transition and continuity of the music. It ranges from legato (connected notes) to staccato (short abrupt notes).
- *Intensity / Loudness:* Intensity is a measure of changes in volume.
- *Timbre / Spectrum:* Timbre describes the quality of the sound. It is often defined in terms of features of the spectrum gathered from the audio signal.

Musical mood has previously been modeled using only audio features. Lu et al. classified classical music into the four quadrants of A-V space using a collection of audio features with an accuracy of 86.3% [1]. Their algorithm also detected places within the song where the mood changed. Experts labeled musical mood. Feng et al. classified Western popular music into four moods using only two features: tempo and articulation. They achieved a precision of 67% and a recall of 66% [2]. They do not specify how they gathered musical mood.

Some effort has been made to incorporate other musical context with audio features to improve classification. Yang et al., working with a set of Western Rock music, made small gains in their classification rates by adding lyrics to the audio features (from 80.7% to 82.8%) [12]. Musical mood was gathered in a laboratory study. Bischoff et al. integrated socially created tags with audio features, and while their classification rates were low due to problems with their ground truth data, they achieved better results using tags and audio features than audio features alone [13]. Their poor results may be due to the fact they were using a diverse, online, data set with multiple genres. Musical mood was specified in this data set by users of the AllMusic site.

2.4 Music Recommenders

Many commercial music recommender systems exist (e.g., Last.fm, Pandora, Apple's Genius, StereoMoods). In 2010, Han et al. created COMUS, a context-based music recommender system that accounts for mood, situation and musical features [14]. Their system was limited to recommending music for only one listening purpose – to transition between emotional states – and assumed a prior explicit knowledge about how a specific individual changes their music habits depending on situation.

3. EXPERIENCE SAMPLING SOFTWARE

To gather an in-situ data set of musical mood and listening context, we surveyed participants using an experience-sample methodology [15]. We created an application that ran on Android 2.1 smartphones, which generated custom surveys from XML files. Participants were asked to carry the phone with them at all times. While it would be possible to create a plug-in for an existing computer media player such as iTunes, we wanted to capture listening experiences in all contexts. For example, some activities, such as exercising, do not usually occur simultaneously with computer use. Participants were not required to use the phone as a media player as this would

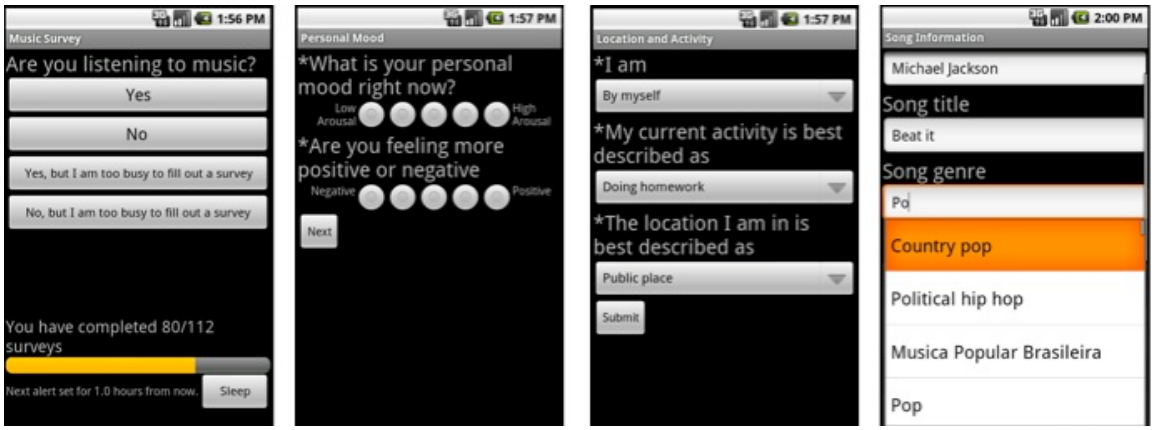

Figure 2 shows screenshots of the experience-sampling software. Participants answered a short survey about their affective state, listening context and the music they were listening too.

further limit listening contexts (e.g., music playing in the background at a restaurant). The tradeoff is that we could not automatically capture song title, artist, or audio features such as tempo.

The program would query the user randomly (approximately hourly) by vibrating the phone. A participant could fill out a survey or dismiss the program by indicating they were too busy. Surveys were completed in less than five minutes and were filled out regardless of whether participants were listening to music. This was done to encourage survey completion. Participants were paid per number of surveys completed, between 5 and 40 CAD. To obtain the maximum payout, 112 surveys were required, which is roughly 8 surveys per day. A progress bar in the software provided feedback about how many surveys had been completed.

Four types of information were collected: musical mood, affective state, musical context and listening context. See Figure 2 for screenshots of the experience-sampling application.

Musical Mood: Participants were asked to describe the musical mood of the song they were listening to using two five-point differential scales. They were asked to rate the arousal of the music by selecting one of five radio buttons between low arousal and high arousal. Similarly, they rated the valence of the music on a scale between sad and happy. Definitions were given to participants before the study and available from a help menu.

Affective State: Participants were asked to describe their personal arousal and valence using five-point differential scales similar to musical mood.

Artist, Title and Genre: Artist and title could optionally be entered in free-text fields that autocompleted to previously entered answers. A genre field was provided that autocompleted to a list of common genres taken from Wikipedia, but also allowed participants to enter their own genre.

Listening Context: Participants were asked questions describing their current listening context. Participants selected their current activity from a list (waking up, bathing, exercising, working, doing homework, relaxing, eating, socializing, romantic activities, reading, going to sleep, driving, travelling as a passenger, shopping, dancing, getting drunk, other). These activities were taken from [8], which lists the most common activities to occur in conjunction with music. Participants also selected their location (home, work, public place, other) and social company (by myself, with people I know, with people I do not know). Participants selected their reason for listening (to express or release emotion, to influence my emotion, to relax, for enjoyment, as background sound, other) as well as whether or not they choose the song (yes, yes as part of a playlist, no). A text field was provided for participants to enter any terms or phrases they associated with the song.

4. DATA SET GATHERED IN-SITU

Twenty participants, (14 male) with an average age of 25, used the experience-sampling software for two weeks.

For a full description of the data set, see [3]. Here we summarize for the purposes of guiding the development of our musical mood classifiers. In total 1803 surveys were filled out; 610 of those surveys were completed when the participant was listening to music. Only the results of the music surveys are included in this paper.

Participants had an average arousal of 2.28 (SD=0.92) on our 5-pt scale (0 low, 2 neutral, 4 high) and average valence of 2.64 (SD=0.90). The music they were listening to had an average arousal of 2.64 (SD=1.05) and average valence of 2.66 (SD=1.14).

The most common activities while listening to music were working (37%) and relaxing (21%). Users also listened to music while eating (6%), driving (5%), travelling (as a passenger)(5%), other (5%), and socializing (4%). Participants were by themselves 57% of the time, with people they knew 37% and with people they did not know 6%. They were at work 39% of the time, at home 38%, in a public place 21% and in other locations 2%.

The most common reason for listening was to use the music as background sound (46%) or enjoyment (25%). Participants chose the song 74% of the time; 50% of the time it was as part of a playlist.

Participants entered 102 unique song genres a total of 486 times. Genres were coded into their parent genre and the most common genres were pop (28%), rock (23%),

electronic (14%), jazz (7%), hip-hop & rap (6%), other (5%), modern folk (4%) and country (3%). The remaining genres were classical, traditional/indigenous music, soundtrack, blues, easy listening and R&B.

Participants entered musical associations for 335 songs. These were then coded into themes, from a list partially taken from [8]. Participants mostly described emotions (45%), lyrics or instruments (20%), imagery (15%), or specific people, locations or memories (7%).

Songs were not limited to Western genres or even the English language. At least 14% of the songs with artist and title specified were non-English; however, all participants listened to at least some English music.

5. CLASSIFICATION FEATURES

To create classifiers of musical mood, we included musical features used in previous work, but also added context-based features from our data set.

5.1 Musical Features

Songs were downloaded from iTunes and other sources where possible using the artist and title specified.

5.1.1 Audio Features

Audio features describing the mode, rhythm, articulation, and timbre of the music were extracted using MIRtoolbox [16] and Matlab.

Mode: These features included the most probable key of the music as well as an estimation of whether the key was major or minor.

Rhythm: These features included an estimation of tempo (number of beats per minute) and pulse clarity (relative strength of the beat, related to how easily a listener can perceive the tempo[17]).

Articulation: These features included the attack slope, (an indicator of how aggressively a note is played) as well as the Average Silence Ratio (ASR)[2].

Timbre: These features were taken from the audio spectrum and include brightness (amount of energy above a cutoff point in the spectrum), rolloff (the frequency such that 85% of total energy is contained below that frequency), spectral flux (average distance between the spectrum of successive frames), spectral centroid (the frequency around which the spectrum is centered), MFCC (description of the sound separated into different bands), low energy (percentage of frames with less than average energy), and average sensory roughness. Sensory roughness corresponds to when several sounds of nearly the same frequency are heard, causing a "beating" phenomenon. High roughness corresponds to harsher music with more "beating" oscillations.

5.1.2 Lyrics

Lyrics were downloaded from various sources using the artist and title. Some included mark-ups indicating non-word sounds or names of singers responsible for a section of lyrics. Only English lyrics were collected. Songs that were mainly English but included a few foreign words were included. Some songs contained notations indicating that a section was repeated (e.g., "x2"). These were manually removed and replaced with the repeated text. Lyrics were analyzed using the Linguistic Inquiry Word Count Tool (LIWC) [18], a textual analysis tool that provides a word count in 80 categories and the output of LIWC was used as the feature set.

5.1.3 Tags

Socially created tags from the website Last.fm were downloaded and analyzed using LIWC. This output was used as features.

5.2 Affective Features

This included the personal arousal and valence of the listener on a 5-point scale.

5.3 Listening Context

Listening context included: reason for listening, activity, location, social company, and level of choice over the song. The associations were categorized (see section 4) and this category was included as a feature. Associations were also analyzed using LIWC.

6. MODEL RESULTS

6.1 Feature Sets

We used a number of feature combinations in creating our models, which can be summarized as three feature sets.

Musical Features: Our first feature set used audio features, lyrical features, and tag features, as these features were used in previous models based on laboratory-gathered data sets of a single genre. There were 198 different features in this set.

Musical Features + Affective Features: Our second feature set used all the musical features but added personal arousal and valence for a total of 200 different features.

Musical Features + Listening Context: Our third feature set combined musical features with the listening context collected in our study for a total of 296 features.

6.2 Models of Musical Mood

Due to "in the wild" nature of the study, musical arousal and musical valence had an uneven distribution of responses. Participants were much more likely to indicate that they were listening to songs with high arousal and high valence. To prevent over fitting of the model due to class skew, musical arousal and musical valence were binned into two levels, low, and high. Neutral instances were ignored. Since only songs with song titles could be downloaded and audio features extracted, instances without a song title were also ignored. Also, undersampling, a technique that selects a random number of instances to obtain an equal distribution, was used. This lowered the total number of instances from 610 to 122 when modeling musical arousal and 156 when modeling musical valence. To avoid any effects caused by the specific set of random instances chosen, this process was completed five times, and the average accuracies of all runs are reported.

All models were created in Weka [19] using Bayes Net classifiers, Markov Estimation and tenfold cross validation. We modeled musical arousal and musical valence

separately, using each feature set. See Figure 3 for classification accuracies.

Musical Features: Using only musical features (audio features, lyrics and tags), musical arousal has a classification accuracy of 59.5% (SD=3.1, kappa=0.1984). Musical valence has an accuracy of 53.0% (SD=6.8, kappa=0.0604). While both models are higher than chance (50%), a one sample t-test shows that only musical arousal (t_4=6.74, p<0.01) was significantly higher. However, the results are lower than previously reported classification accuracies of homogenous lab-based data sets.

Musical Features + Affective State: When we combined affective features (personal arousal and valence) with musical features, musical arousal has a classification accuracy of 60.3% (SD=4.6, kappa=0.2121). Musical valence has an accuracy of 60.2% (SD=4.6, kappa=0.2074). Both musical arousal (t_4=4.99 p<0.01) and musical valence (t_4=4.70, p<0.01) performed significantly better than chance (50%), and we achieved improved classification accuracies of musical arousal and musical valence by using a combination of affective and musical features.

Musical Features + Listening Context: When we combined musical features with listening context features, musical arousal has a classification accuracy of 67.4% (SD=1.7, kappa=0.3437). Musical valence has an accuracy of 75.7% (SD=1.5, kappa=0.5133). Both musical arousal (t_4=23.54, p <0.0001) and musical valence (t_4=38.75, p<0.0001) performed significantly better than chance (50%), and we achieved gains in classification accuracy in both models over using only musical features or musical features and affective features combined.

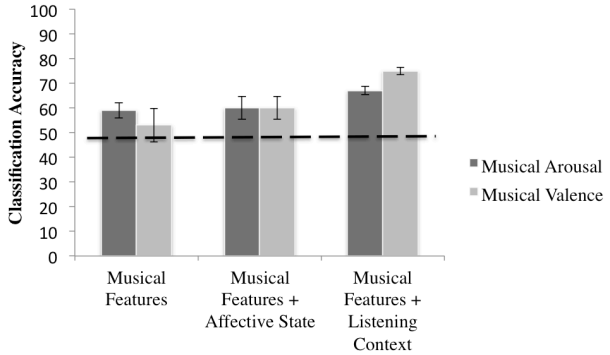

Figure 3 shows the classification accuracies for each feature set. The dotted line shows chance (50%).

7. DISCUSSION

Our experience sampling study collected in-situ data that reflects real-life listening experiences. Unlike previous models, our data included multiple genres and different listening contexts.

We have shown that musical mood can be successfully modeled from in-situ data, although with a lower classification accuracy than previous attempts. Adding affective state to the model resulted in an improvement in classification accuracy while modeling musical valence; adding listening context to the model resulted in improvements in both musical arousal and musical valence. Our results show that listening context is an important aspect of modeling musical mood, when using real-life data.

7.1 Importance of Context

It may be possible that context is important when modeling musical mood because participants rate musical mood differently depending on their context. For example, a user may rate the same song differently depending on whether they are working alone or cooking with friends. We cannot confirm this with our data set, as one would need the same songs played in a variety of listening contexts – in our study, songs and artists were only encountered once on average.

It is also possible that people listen to music with certain musical moods based on their context. For example, a user may generally choose to listen to music with high arousal when exercising and low arousal when eating dinner. In that case our model predicts the type of musical mood listeners want to listen to, based on context, which is useful for automatically generating playlists.

Similarly, participants may rate musical mood differently depending on their affective state. This is a tricky relationship to investigate as the music itself has a hand in inducing an affective state in a listener. Any correlation found between musical mood and affective state does not show directionality of the relationship.

To examine the relationships between listening context, musical mood, and affective state, we could provide users with representative samples in a music library. By listening to (and rating) the same song in a variety of contexts and affective states, the relationship between these three factors might be made clear.

7.2 Limitations

There are several limitations with our study. The first is that participants are unlikely to answer a survey during some activities (e.g., driving). Second, all categories in our data may not be mutually exclusive (e.g., reading while running on the treadmill). Third, the number of participants and length of the study may have been too small to collect a fully representative sample of listening context. Finally, previous studies have assumed that people listen to music with four emotional categories (happy, sad, fear, anger) [11]; however, in our study we found that people tended to listen to happy music. The other three emotions may not be equally represented when capturing in-situ data [3].

While a classification accuracy of 75% is much improved over a random classifier, or one based on auditory features, a music recommender suggesting songs with the wrong mood a quarter of the time may result in a negative user experience. This can be circumvented in a few ways. First a music recommender can select tracks from a personal music library; users are more likely to enjoy their own music even if the recommendation is off. Second, a playlist rather than a single song could be recommended so that a majority of the music recommended is suitable. Third, combining this model with existing recommendation systems that use clustering of similar genres and artists could further improve existing prediction rates. Finally, we could improve the classification rates and avoid possible overfitting caused by the small number of instances in our models by collecting a more comprehensive data set.

8. FUTURE WORK

Based on the results of this work, we will create a context-aware music recommender system. This system will take in the context of the listening experience and use this context to compile a playlist. Based on our models, the system will recommend a musical mood listeners are likely to enjoy, and will create playlists of songs with this specific musical mood, (based on a data set labeled from musical features). The system could also make suggestions of songs for purchase the user might enjoy. We will evaluate the predictions through a user study, conducted in-situ, to preserve the importance of context.

To create the underlying model for this music recommender, a larger in-situ data set will be collected. The study will run for a longer time period (i.e., months) with a larger pool of participants. Participants will receive bonuses for filling out genre, title and artist and will be asked to provide a copy of their music library at the end of the study for audio feature processing. This larger, more comprehensive data set will help improve classification accuracies.

9. CONCLUSIONS

We successfully model musical mood from a data set gathered in-situ during a user's daily life. We show that musical features (audio features, song lyrics and socially created tags) can successful model musical mood with classification accuracies better than chance. We successfully classify musical arousal with a classification accuracy of 59% and musical valence with an accuracy of 53% when using only musical features on an in-situ data set.

Adding contextual information, such as the listener's affective state or the listening context of the musical experience can further improve classification accuracies. We successfully classify musical arousal and musical valence with a classification accuracy of 60% when using both musical features and affective state. We classify musical arousal with a classification accuracy of 67% and musical valence with an accuracy of 75% when using both musical features and listening context.

10. ACKNOWLEDGEMENTS

This work was funded in part by the GRAND NCE.

11. REFERENCES

[1] L. Lu, D. Liu, and H.J. Zhang, "Automatic mood detection and tracking of music audio signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 5 – 18, Jan. 2006.

[2] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2003, pp. 375–376.

[3] D. Watson and R. L. Mandryk, "An In-Situ Study of Real-Life Listening Context," in *SMC 2012*.

[4] P. Ekman, *Basic Emotion*. John Wiley & Sons, Ltd., 2005.

[5] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological Review*, vol. 110, no. 1, pp. 145–172+, 2003.

[6] P. Juslin and J. Sloboda, *Music and Emotion: Theory and Research*. Oxford University Press, 2001.

[7] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici, and S. Koelsch, "Universal Recognition of Three Basic Emotions in Music," *Current Biology*, vol. 19, no. 7, pp. 573 – 576, 2009.

[8] P. Juslin and P. Laukka, "Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening - Journal of New Music Research," *Journal Of New Music Research*, vol. 33, no. 3, pp. 217–238, 2004.

[9] K. . Scherer and M. R. Zentner, "Emotional Effects Of Music: Production Rules," in *Music and Emotion:Theory and Research*, New York, NY, USA: Oxford University Press, 2001, pp. 361–392.

[10] M. R. Zentner, S. Meylan, and K. Scherer, "Exploring 'musical emotions' across five genres of music.," in ICMPC, Keeyle, UK, 2000.

[11] P. N. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 6, pp. 1797–1812, 2000.

[12] D. Yang and W. Lee, "Disambiguating Music Emotion Using Software Agents," in *ISMIR* Barcelona

[13] K. Bischoff, C. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music Mood and Theme Classification-a Hybrid Approach," in *ISMIR*, Kobe, Japan, 2009.

[14] B.-J. Han, S. Rho, S. Jun, and E. Hwang, "Music emotion classification and context-based music recommendation," *Multimedia Tools Appl.*, vol. 47, pp. 433–460, May 2010.

[15] R. Larson and M. Csikszentmihalyi, "The Experience Sampling Method.," *New Directions for Methodology of Social & Behavioral Science*, vol. 15, pp. 41–56, 1983.

[16] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction from Audio," in *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, France, 2007.

[17] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "Multi-Feature Modeling of Pulse Clarity: Design Validation and Optimization," presented at the ISMIR 2008, 2008.

[18] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic Inquiry and Word Count LIWC 2001," *Word Journal Of The International Linguistic Association*, pp. 1–21, 2001.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.