INVESTIGATION OF THE RELATIONSHIPS BETWEEN AUDIO FEATURES AND INDUCED EMOTIONS IN CONTEMPORARY WESTERN MUSIC

Konstantinos Trochidis
LEAD-CNRS
Université de Bourgogne
Konstantinos.Trochidis@u-bourgogne.fr

Charles Delbé
LEAD-CNRS
Université de Bourgogne
Charles.Delbe@u-bourgogne.fr

Emmanuel Bigand
LEAD-CNRS
Université de Bourgogne
Emmanuel.Bigand@u-bourgogne.fr

ABSTRACT

This paper focuses on emotion recognition and understanding in Contemporary Western music. The study seeks to investigate the relationship between perceived emotion and musical features in the fore-mentioned musical genre. A set of 27 Contemporary music excerpts is used as stimuli to gather responses from both musicians and non-musicians which are then mapped on an emotional plane in terms of arousal and valence dimensions. Audio signal analysis techniques are applied to the corpus and a base feature set is obtained. The feature set contains characteristics ranging from low-level spectral and temporal acoustic features to high-level contextual features. The feature extraction process is discussed with particular emphasis on the interaction between acoustical and structural parameters. Statistical relations between audio features and emotional ratings from psychological experiments are systematically investigated. Finally, a linear model is created using the best features and the mean ratings and its prediction efficiency is evaluated and discussed.

1. INTRODUCTION

The expressive aspects of music are the most difficult to analyze structurally, inducing a large variety of emotional responses in humans. The richness of these responses is what motivates an engagement with music [1]. Many studies indicate the important distinction between one's perception of the emotion(s) expressed by music and the emotion(s) induced by music. Studies of the distinctions between perception and induction of emotion have demonstrated that both can be subjected to not only the social context of the listening experience, but also to personal motivation [2].

Modeling the perception of expressive musical content is highly useful in MIR applications such as emotion based classification and recommendation systems, radio and TV broadcasting programs, and music therapy defining appropriate musical repertories for research in patients suffering from Alzheimer or Bibromilagic. Due to the highly conceptual elusiveness of emotions and the limitation of computational methods mainly based on low level features, modeling and prediction of emotion in music remains a particularly difficult task. Research on music and emotion has always focused in music genres such as Western Popular or Classical music.

To the best of our knowledge limited research has been conducted in the field of Contemporary Western music. The term "Contemporary art music" is used for Western art-tradition music written since 1945. Characteristic structures in Western music like systematic variation of tempo, mode and timbre, which are identified in Classical or popular modern music, do not necessarily exist in Contemporary art music. This raise questions such as:
1) Can an emotional response be triggered when the mentioned features are not present?
2) Which other features contribute to emotional reaction? and 3) Are the same features contribute to emotion generation in both musicians and non musicians?

The present paper deals with the above issues. Section 2 provides background material on previous research on music studies, while Section 3 presents the ground truth and the experiments carried out with musicians and non-musicians volunteers. Section 4 describes the audio feature extraction and representation while, Section 5 discusses the statistical selection of features and modeling of music emotional regression. Discussion and conclusions are drawn in Section 6.

2. RELATED WORK

Many psychological models have been used in studies concerning music and emotion. The main approaches existing in the literature are the discrete and the dimensional models [3]. A comparison of the discrete and dimensional models of emotions in music can be found in [4]. According to the categorical approach, emotions are conceptualized as discrete unique entities and contain a certain basic number of emotion categories from which all the emotional states are derived. There is an agreement towards researchers representing this approach as to five basic emotions: happiness, sadness, anger, fear and disgust [5], [6].

In the dimensional approach, emotions are expressed on a Cartesian coordinate system according to two axes those of valence and arousal. The model depicted in figure 1 shows Russell's [7] circumplex model of affect, where the axes measure activation and pleasure. Happiness and Anger are located at the top of the vertical (arousal) axis, Serenity and Sadness are located at the bottom. On the horizontal axis (pleasure), Happiness and Serenity are more positive emotions than Anger and Sadness, and so these pairs are located on the right and left side respectively of this axis of Russel's space. In the late 1990s, Thayer [8] proposed a two-dimensional mood model that uses individual adjectives which collectively form a mood pattern. This dimensional approach adopts the theory that mood is entailed from two factors: Stress (happy/anxious) and Energy (calm/ energetic), and divides music mood into four clusters: Contentment, Depression, Exuberance and Anxious/Frantic.

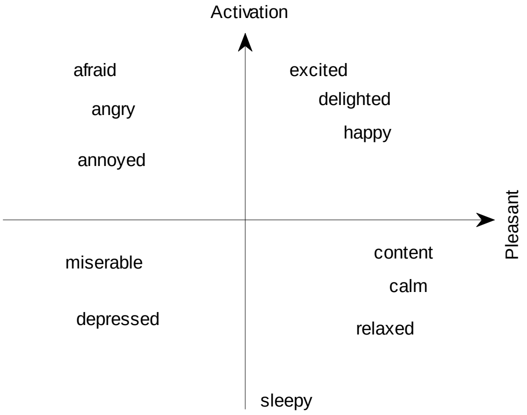

Figure 1. Russell's Circumplex model of emotion.

Music emotion representation in an affective space gained lots of interest among researchers. Hevner's study [9] arranges music emotion in a circle of eight clusters. Each cluster represents a certain emotion and contains seven to eleven adjectives grouped together. Shubert [10] argues that emotional reaction depends on the interaction between different factors such as musical mode, tempo, loudness and pitch. In [11] the time course of emotional responses to music is investigated in both untrained and trained musicians using musical excerpts of increasing duration ranging from 250 ms to 20 seconds. Findings show that less than 1s of music is enough to instill elaborated emotional responses in listeners

The four main emotion classes of Thayer's model were used as the emotion model in [12]. Three different feature sets were adopted for music representation, namely intensity, timbre and rhythm. Gaussian mixture models were used to model each of the four classes. An interesting contribution of this work was a hierarchical classification process, which first classifies a song into high/low energy (vertical axis of Thayer's model), and then into one of the two high/low stress Classes.

Emotion recognition is modelled as a regression task in [13]. Volunteers rated a training collection of songs in terms of arousal and valence in an ordinal scale of values from -1 to 1 with a 0.2 step. The authors then trained regression models using a variety of algorithms and a variety of extracted features.

A model predicting perceived emotions based on a set of features extracted from soundtrack music is given in [14]. Three separate data reduction techniques, namely stepwise regression, principal component analysis, and partial least squares are compared.

The effectiveness of current music understanding processes and music intelligent systems is mainly hampered by the so called semantic gap between human perception and cognition and on the other hand by the low level music features which are mostly statistics of spectral and temporal characteristics in the signal.

Therefore, many researchers [15], [16], [17] try to bridge the semantic gap between the low level features and high level semantics, which humans perceive and understand, by merging different modalities such as low level acoustical features and social data including lyrics, tags and web logs.

3. EXPERIMENTAL SETUP

We decided to analyze and explore Western Contemporary art music because several of its features are shared differently by other musical genres. The main concept that best describes contemporary music is confusion in listening. Harmony does not necessarily play an important role. Thus, there is a difficulty of extracting and interpreting harmonic information because of the lack of tonal reference system. Composition contains clustered sounds and disharmonic intervals very different compared to the ones found in Western Classical music or modern popular music. Instrumentation is very complex and musicians use their instruments for producing sounds very different from those encountered in the Classical repertoire. Mixed Electro acoustic and traditional instruments raises the problem of which information of inharmonic sounds related to timbre can be captured and represented.

3.1 Method

The data used in this paper are based on a previous study [11]. Twenty participants without musical training (referred to as non musicians) and 20 with an average of 10 years of musical training and instrumental practice participated in this experiment. A set of 27 musical excerpts of Contemporary music is selected by music theorists and psychologists according to several constraints. All excerpts were expected to convey a strong emotional experience. They were chosen to illustrate a large variety of emotions, and to be representative of key musical periods of Contemporary Western music. The excerpts showed a great variation in musical structure including harmony, rhythm, tempo, timbre and instrumentation. The participants were asked first to listen to all excerpts and then focus their attention on their private emotional experience. Next, they were asked to look for excerpts that induced similar emotional experience based on arousal and valence dimensions (whatever that may be) and to drag the corresponding icons in order to group these excerpts. They corresponded either to the beginning of a musical movement, or to the beginning of a musical theme or idea. An average duration of 30s sounded appropriate. We adopted a dimensional approach for emotional labelling because it avoids a strict classification and accounts for similarity and dissimilarity of emotions. Linguistic labels remain problematic and may simplify the emotional reaction and further disregard the difference between induced and perceived emotions.

The groupings of participants were then converted into a 27x27 matrix of co-occurrence. Each cell of the matrix indicated the average number of times that two excerpts were grouped together. The subtraction of the average matrix of occurrence from 1 resulted in a matrix of dissimilarity. The matrices obtained were highly correlated for musicians and non musicians, (r= .65, p < .001). The resulting matrices were analysed with MDS and cluster analysis methods. The locations of the 27 excerpts along the two principal dimensions are presented in Figures 2 and 3. The vertical axis represents musical excerpts that varied obviously by their arousal level (with low arousal pieces at the bottom, and high arousal pieces at the top). The horizontal axis represents presumably musical excerpts that differ by their emotional valence (with positive valence on the right and negative valence on the left).

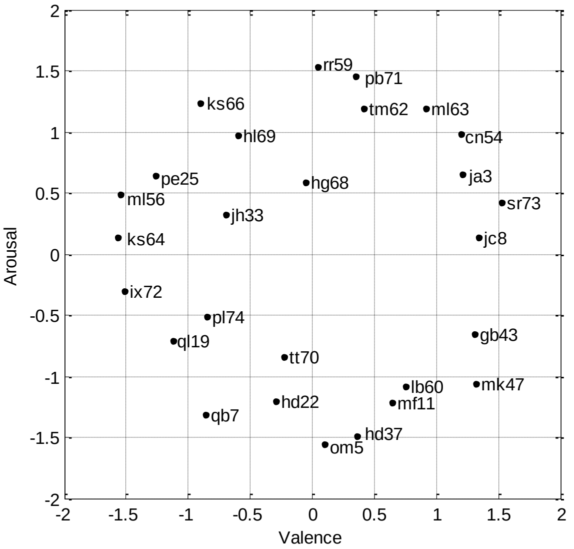

Figure 2. Geometrical representation for the 27 Contemporary music excerpts in musicians, resulting from the MDS analysis.

4. AUDIO FEATURE EXTRACTION

4.1 Low-level acoustical features

A theoretical selection of musical features was made based on music characteristics such as timbre, harmony, rhythm and dynamics. A total of 324 features where extracted from the music excerpts representing information related to the above concepts. The MIR Toolbox for MATLAB was used to compute the various low and high level descriptors [18].

4.1.1 Rhythmic features

A rhythmic analysis of the music signals was performed. Descriptors such as the fluctuation (the rhythmic periodicity along auditory frequency channels), the estimation of notes onset times and the number of onsets per second were computed. Finally, the tempo was estimated.

4.1.2 Timbre features

Mel Frequency Cepstral Coefficients (MFCCs) are used for speech recognition and music modeling. To derive the MFCCs, the signal was divided into frames and the amlit ude spectrum for each frame was calculated. Next, its logarithm was taken and converted to Mel scale. Finally,

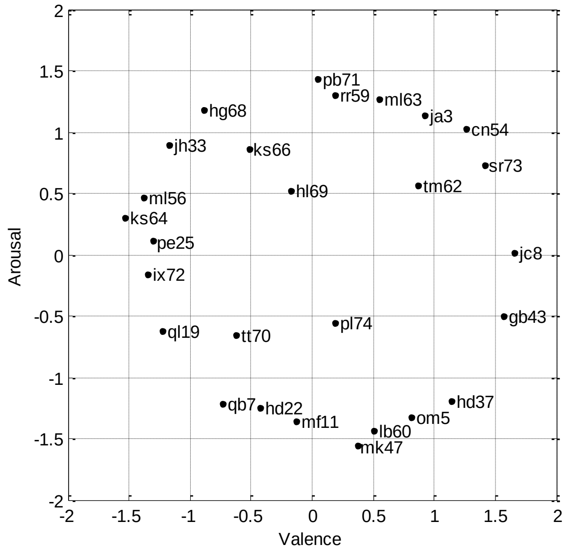

Figure 3. Geometrical representation for the 27 Contemporary music excerpts in non-musicians, resulting from the MDS analysis.

the discrete cosine transform was implemented. We selected the first 13 MFCCs. Another set of 4 features related to timbre textures were extracted from the Short-Term Fourier Transform: spectral centroid, spectral rolloff, spectral flux and flatness which indicate whether the spectrum distribution is smooth or spiky. The size of the frames used to compute the timbre descriptors was 0.05 sec half-overlapped.

4.1.3 Tonal features

The signals were also analyzed according to the tonality context. Descriptors such as the Chromagram (energy distribution of the signals wrapped in the 12 pitches), the key strength (i.e. the probability associated with each possible key candidate, through a cross correlation with the Chromagram and all possible key candidates), the tonal Centroid (a six dimensional vector derived from the Chromagram corresponding to the projection of the chords along circles of fifths or minor thirds) and the harmonic change detection function (flux of the tonal Centroid) were extracted.

4.1.4 Dynamic features

We computed information related to the dynamics of the music signals such as the RMS and the percentage of low energy frames to see if the energy is evenly distributed throughout the signals or certain frames are more contrasted than others For all features a series of statistical descriptors were computed such as the mean, the standard deviation and the linear slope of the trend along frames, i.e. the derivative. For all Mel Frequency Cepstral Coefficients the first and second derivative were computed. Also the maximal periodicity detected in the frame-by frame evolution of the values was estimated through the computation of the autocorrelation sequence and the amplitude of this periodicity.

4.2 High-level Contextual features

While low-level descriptors such as loudness (perception of sound intensity) or pitch (perception of fundamental partials) account for perceptual aspects of music, higher-level ones, such as pulse, harmony or complexity account for contextual aspects, i.e. they refer to the cognitive perception and aspects of music. Many models using low level features successfully predicted the dimension of Arousal. The retrieval, however, of Valence has proved to be difficult to measure by using only low level information [19]. In order to tackle this problem, we used a set of five high level features in conjunction with the low level descriptors which are described above.

4.2.1 Pulse Clarity

This descriptor measures the sensation of pulse in music. Pulse can be described as a fluctuation of musical periodicity that is perceptible as “beatings” in a sub-tonal frequency band below 20Hz. The musical periodicity can be melodic, harmonic or rhythmic as long as it is perceived the listener as a fluctuation in time.

4.2.2 Articulation

Articulation usually refers to the way in which a melody is performed. If a pause is clearly noticeable in between each note in the melodic, the articulation of the melody is *staccato*, which means “detached”. On the other hand, if there is no pause in between the notes of the melody then the melody is *legato*, meaning “linked”. This feature attempts to estimate the articulation from musical audio signals by attributing to it an overall grade that ranges continuously from zero (staccato) to one (legato).

4.2.3 Mode

This feature refers to a computational model that detects between major and minor excerpts. It calculates an overall output that continuously ranges from zero (minor mode) to one (major mode).

4.2.4 Event density

This descriptor measures the overall amount of simultaneous events in a musical excerpt. This can be melodic, harmonic and rhythmic, as long as they can be perceived as independent entities by the human cognition.

4.2.5 Brightness

This descriptor measures the sensation of how bright of a music excerpt is felt to be Attack, articulation, or the unbalance or lacking of partials in other regions of the frequency spectrum can influence its perception.

5. FEATURE SELECTION

From the selected features, only those whose correlation with the ratings is sufficiently statistically significant (with a p-value lower than .05) are selected. The selected features are ordered from the most correlated to the least correlated ones. Features that are not sufficiently independent with respect to the better scoring ones (with a normalized cross correlation exceeding 0.6) were removed as well. In order to see how these acoustic features may account for the present data, a normalized stepwise regression of the coordinates of the pieces on the two axes was performed using the best features. Table 1 and 2 provide the outcome of the multiple linear regression analysis of the acoustic features over the coordinates of the pieces for musicians and non musicians. The resulting model provides a good account of the arousal for musicians (adjusted R² = 0.72, see Table 1), with the periodicity amplitude of flatness (β = 0.60) and the entropy of the magnitude of the highest peak in the chromagram (β = -0.43) contributing the most, followed by the flux of the tonal centroid and the mean derivative of the 3d mfcc band. On the other hand, the regression model provided a moderate account of valence with R² = 0.57, with the mean of pulse clarity (β = 0.74) (i.e. the perceived sensation of pulse) contributing the most, followed by the mean of articulation (β =0.68) and brightness (β =-0.40).

Valence	β	Arousal	β
Pulse_clarity	0.74	FlatnessPeriodAmp	0.60
Articulation	0.68	chromagramPeakstd	-0.43
brightness	-0.40	Tonal_hdcf	0.33
Event_density	-0.19	Dmfcc_mean_3	-0.18
Mode_Mean	0.17	FlatnessPeriodFreq	-0.06

Table 1. Outcome of the multiple linear regression analysis of the acoustic features over the coordinates for musicians.

The outcome of the multiple regression analysis of the acoustic features over the coordinates of the pieces for non musicians is presented in Table 2. One can see that the results are very similar to that of the musicians. The resulting model provides a good account R² = 0.67 of the arousal for the non musicians, with the periodicity amplitude of flatness (β = -0.58) and the entropy of the magnitude of the highest peak in the chromagram contributing the most (β = -0.46) followed by the flux of the tonal centroid. The regression model provided a moderate account of valence with R² = 0.62, with the mean of pulse clarity and the mean derivative of the 10th mfcc band contributing the most, followed by the mean of brightness.

Valence	β	Arousal	β
Articulation	0.83	FlatnessPeriodAmp	0.58
Pulse_clarity	0.68	chromagramPeakstd	-0.46
brightness	-0.50	Tonal_hdcf	0.25
Event_density	-0.19	FlatnessPeriodFreq	-0.19
Mode_mean	0.05	RoughnessPeriodFreq	-0.009

Table 2. Outcome of the multiple linear regression analysis of the acoustic features over the coordinates for non musicians.

6. CONCLUSIONS

In the present paper the relationships between music features and emotion perception in the case of Contemporary Western music are investigated. A systematic analysis of the musical stimuli shows that low level spectral and temporal features such as flatness and chroma features are efficient in modeling the emotion perception of arousal dimension, while high-level contextual information such as articulation, pulse clarity, mode and brightness succeed to measure the more cognitive nature of valence. The results contradict the widespread opinion that understanding of contemporary western music is restricted to highly trained listeners. It is shown that the emotion processing mechanism is quite similar for musicians and non musicians with the same low level spectral, temporal features correlated with arousal and high level contextual features correlated with valence dimension. Contemporary Western music can serve successfully as stimulus for studying the emotional processing mechanism in music. An emotional response can be still triggered when characteristic structures and features of Western popular or Classical music are not present.

Future work will explore the effectiveness of new features extracted from physiological signals such as EEGs to bridge the semantic gap between high level knowledge related to the cognitive aspects of emotion and low level acoustical features. Furthermore, a larger Contemporary music dataset will be constructed and new audio features will be designed and tested to allow for better statistical results.

7. REFERENCES

[1] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau & A. Dacquet, “Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts”, in Cognition & Emotion, 2005, 19(8), 1113–1139.

[2] P. N Juslin and P. Luakka, “Expression, perception, and induction of musical emotions: A review and questionnaire study of every day listening”, in Journal of New Music Research, 2004, 33, 217–238.

[3] P. Juslin and J. Sloboda, Music and emotion: Theory and research. Oxford, England: Oxford University Press, 2001.

[4] T. Eerola, &J.K Vuoskoski. “A comparison of the discrete and dimensional models of emotion in music”, in Psychology of Music, 2011, 39(1), 18-49.

[5] R. Plutchik. The psychology and biology of emotion. Harper Collins, New York, 1994.

[6] T.D. Kemper. How many emotions are there? Wedding the social and the autonomic components. American Journal of Sociology, 93:263–289, 1987.

[7] J. A. Russell, “A circumplex model of affect,” in Journal of Psychology and Social Psychology, 1980, 39, 6, 1161-1178.

[8] R. E, Thayer. The biopsychology of mood and arousal. Oxford University Press, 1989.

[9] K. Hevner, “Expression in music: a discussion of experimental studies and theories” in Psychological Review, 1935, 42, 186-204.

[10] E. Schubert. “Measuring emotion continuously. Validity and reliability of the two-dimensional emotion-space”, in Australian Journal of Psychology, 1999, 51(3), 154-165.

[11] E. Bigand, S. Filipic, & P. Lalitte. “The time course of emotional responses to music”, in Annals of the New York Academy of Sciences, 2005, 1060, 429-437.

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

[13] Y.-H. Yang, Y.-C. Lin, Y.-F .Su, and H.-H. Chen. “A regression approach to music emotion recognition” in IEEE Transactions on Audio, Speech and Language Processing, 2008, 16(2):448–457.

[14] T. Eurola, O. Lartillot, P. Toiviainen. “Prediction of Multidimensional Emotional ratings in Music from Audio Using Multivariate Regression Models”, in Proc. Int. Conf. in Music Information Retrieval, Kobe, 2009.

[15] C. Laurier, M. Sordo, J. Serra, and P. Herrera, “Music mood representation from social tags,” in Proc. of the Int. Society for Music Information Conf., Kobe, 2009.

[16] X. Hu, J. S. Downie, and A. F. Ehmann, “Lyric text mining in music mood classification,” in Proc .of the Int. Conf in Music Information Retrieval, Kobe, Japan, 2009.

[17] Y.E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, A. Speck and D. Turnbull. “Music emotion recognition: A state of the art approach”, in Proc. of the Int. Conf. in Music Information Retrieval, Utrecht, 2010.

[18] Lartillot, O., and P. Toiviainen. “MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio”, Proceedings of the International Conference on Music Information Retrieval, Wien, Austria, 2007

[19] J. Fornari, & T. Eerola. “Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music”, in Lecture Notes in Computer Science, chapter The Pursuit of Happiness in Music: Retrieving Valence with Contextual Music Descriptors, 2009, 5493, 119-133. Springer.