Journal of Mathematics and Music
Vol. 00, No. 00, January 2011, 1–11

RESEARCH ARTICLE

Automatic Detection of Emotions with Music Files

Angelina A. Tzacheva[ac*], Dirk Schlingmann[bc], and Keith J. Bell[ac]

[a]*Department of Informatics*, [b]*Division of Mathematics and Computer Science*,[c]*University of South Carolina Upstate, 800 University Way, Spartanburg, SC 29303, U.S.A.*

(*Received 00 Month 200x; final version received 00 Month 200x*)

The amount of music files available on the Internet is constantly growing, as well as the access to recordings. Music is now so readily accessible in digital form that personal collections can easily exceed the practical limits of the time we have to listen to them. Today, the problem of building music recommendation systems, including systems which can automatically detect emotions with music files, is of great importance. In this work, we present a new strategy for automatic detection of emotions with musical instrument recordings. We use Thayer's model to represent emotions. We extract timbre related acoustic features. We train and test two classifiers. Results yield good recognition accuracy.

Keywords: music information retrieval; emotion detection; timbre; automatic classification; data mining

*Corresponding author. Email: atzacheva@uscupstate.edu

ISSN: 1745-9737 print/ISSN 1745-9745 online

DOI: 10.1080/1745973YYxxxxxxxx
http://www.informaworld.com

1. Introduction

Music has accompanied mankind since ancient times in various situations. To many people music is an important part of life. It is often regarded as ordered and pleasant to listen to. It is a form of art whose medium is sound. Mathematics is believed to be the basis of sound.

Today, we hear music in advertisements, in films, at parties, at the philharmonic, etc. One of the most important functions of music is its effect on humans. Certain pieces of music have a relaxing effect, while others stimulate us to act, and some cause a change in or emphasize our mood. Music is not only a great number of sounds arranged by a composer, it is also the emotion associated with these sounds.

Nowadays, the quantity of sounds surrounding us is rapidly increasing and the access to recordings, as well as the amount of music files available on the Internet is constantly growing. At this time, the problem of building music recommendation systems, including systems which can automatically detect emotions associated with music files, is of great importance [1].

A popular approach called music emotion classification (MEC) divides the emotions into classes and applies machine learning on audio features to recognize the "emotion embedded in the music signal". Yang et.al. [2] speak of a semantic gap between the object feature level and the human cognitive level of emotion perception, which limits the MEC approaches.

The latter statement may be incorrectly phrased. We believe emotions are not something, which is embedded within a signal. We consider an emotion to be a feeling experienced by a human being. Therefore, is it possible for an emotion to be searched for and detected within a signal?

What we discover is that: certain information is present within the signal, which can be linked to the emotion that is invoked within a human while listening to the music recording at hand.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 describes the proposed approach, Section 4 shows the experiment results, finally Section 5 concludes and considers directions for the future.

2. Related work

We review previous work in classification and retrieval of music and sound by emotion, as well as other approaches related to music and emotion recognition.

Yang et.al. [2] propose an approach of linking song lyrics to emotions through text processing. Authors divide the music pieces into frames, and extract both low level audio features, as well as textual features. Next, support vector machine (SVM) based classifier is trained and tested. Authors report enhanced accuracy from 46% to 57% with a 4-class emotion classification by including lyrics features in addition to the audio features.

Authors continue with another study, where a system to music emotion recognition based on regression [3] is proposed. Volunteers rate a training collection of songs in terms of arousal and valence in an ordinal scale of 11 values from -1 to 1 with a 0.2 step. Authors then train regression models, with a variety of extracted features, where SVMs perform best. Finally, a user can retrieve a song by selecting a point in a two-dimensional arousal and valence mood plane.

Busso et.al. [4] study emotionally salient aspects of the fundamental frequency in human voice. During expressive speech, the voice is enriched to convey not only the intended semantic message but also the emotional state of the speaker. Authors find the pitch contour to be one of the important properties of speech that is affected by this emotional modulation. Pitch features have been commonly used to recognize emotions. Author's approach presents an analysis of the statistics derived from the pitch contour. A binary emotion detection system is built for distinguishing between emotional versus neutral speech.

Kim and Andre [5] investigate the potential of physiological signals in music listeners as reliable channels for emotion recognition. Physiological features from various analysis domains, including: time/frequency, entropy, geometric analysis, subband spectra, and multiscale entropy, are studied and correlated with emotional states. Classification is performed by extended linear discriminant analysis (pLDA) and the proposed emotion-specific multilevel dichotomous classification (EMDC).

Grekow and Ras [1] propose a strategy for emotion detection in classical music pieces, which are in MIDI format. Collection of harmonic and rhythmic attributes are extracted from the signal, after it is segmented into frames. Next, each frame is labeled with an emotion class. A database of 83 MIDI files is created for the experiment. Data mining software with k-NN classifier is used.

Trohidis et.al. [6] evaluate four algorithms in modeling detection of emotions as a multi-label classification task. Authors use the Tellegen-Watson-Clark model of mood [7] with 6 clusters of emotions. Results show random k-labelsets (RAKEL) dominates the other algorithms in terms of performance. Multi-label approaches aim to detect more than one emotion at a time. That comes at the price of being more complex, and hence more computationally expensive, in comparison with single-label approaches.

We propose a new method, based on single-label, and a 4-class emotion classification. We use a unique set of timbre related features. Little research has been done using timbre in music classification [9, 10], and emotion classification [6]. Timbre carries semantical information. Thus, it helps to bridge the aforementioned semantical gap between the object feature level and the human cognitive level of emotion perception [2], which has limited the MEC approaches until now. In addition, our proposed method compares Bayesian neural network and J48 decision tree classifiers. These have not been applied in any of the previous work on emotion classification we reviewed. We report good recognition accuracy.

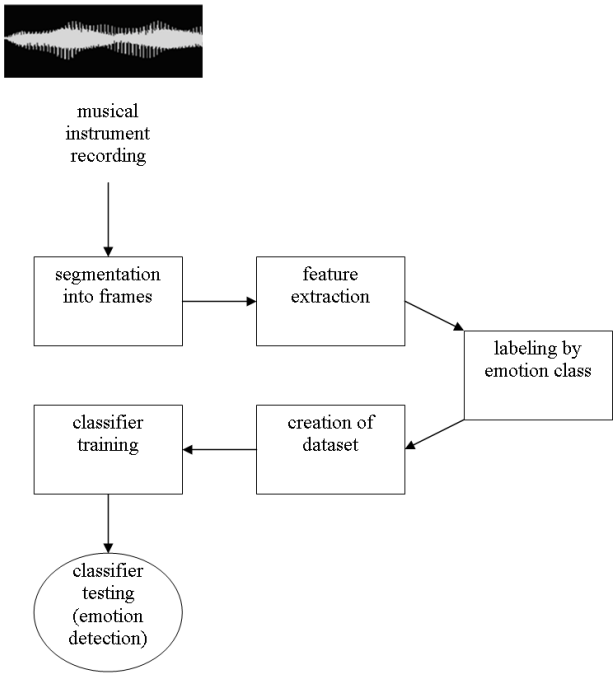

Figure 1. System diagram of the proposed method.

3. Proposed method

Concisely, our method consists of: we start with a set of musical instrument recordings; each recording is segmented into frames; acoustical features are extracted on each frame; a person, with formal education in music, listens to each music recording and assigns an emotion label to it; a dataset is created; with this data, we train Bayesian neural network, and J48 decision tree classifiers - to learn the description of each emotion class, based on the provided acoustical features extracted from the signal; finally, the classifier is presented with an unlabeled recording (no class) and asked to determine its emotion class (testing). This sequence of steps is illustrated in Figure 1.

3.1. Emotion Model

Humans, by nature, are emotionally affected by music. Hevner [11] was the first to study the relation between music and emotion. She discovered 8 clusters of adjective sets describing music emotion and created an emotion cycle of these categories, as shown on Figure 2. This model was used by Li et.al. [8] and Wieczorkowska et al. [12]. It is too complex for our purpose, however it illustrates the intricacy of describing emotions.

Another mood model is the Tellegen-Watson-Clark model shown on Figure 3. The 45

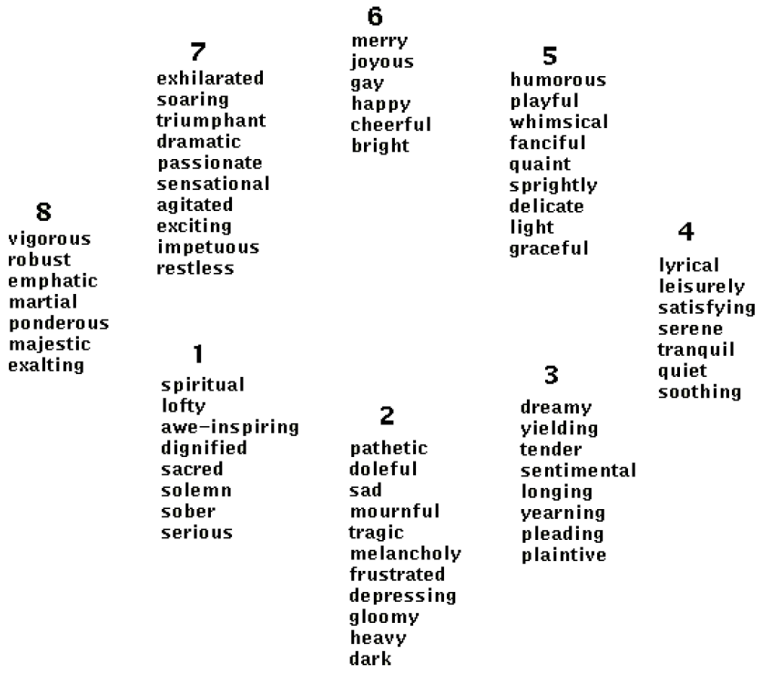

Figure 2. Hevner adjective circle.

degrees axes describe pleasantness (unpleasantness) versus engagement (disengagement) [7]. It was used by Trohidis et.al. [6] where 6 main emotional clusters were retained from the model: amazed-surprised; happy-pleasant; relaxing-calm; quiet-still; sad-lonely; angry-fearful.

In our proposed method we use the Thayer's model of mood, which consists of arousal-valence emotion plane [13]. In the horizontal axis the valence can change from negative to positive, and in the vertical axis arousal can change from low to high, as shown on Figure 4.

We adopt the 4 general emotions groups of Thayer's model, referencing energy and valence. Our mood model contains the emotion classes shown in Figure 5.

3.2. Input Data

We are using audio files of musical instrument recordings. We have chosen 6 instruments: viola, cello, flute, english horn, piano, and clarinet to our experiments. All recordings originate from MUMS CDs [14], which are used worldwide in similar tasks.

We split each recording into overlapping frames. We adopt the temporal cross-tabulation pre-processing from Tzacheva et.al. [9] to preserve temporal information. Next, we extract the acoustical features described in the next section 3.3.

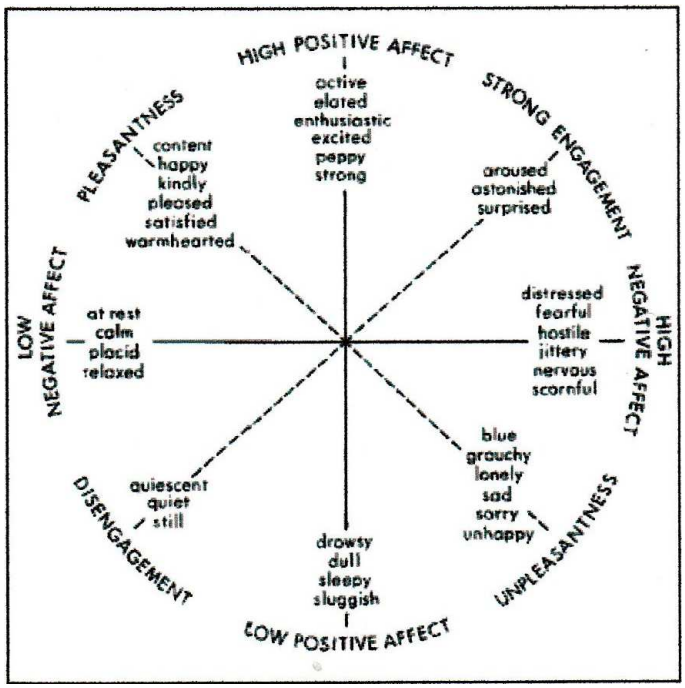

Figure 3. Tellegen-Watson-Clark model of mood.

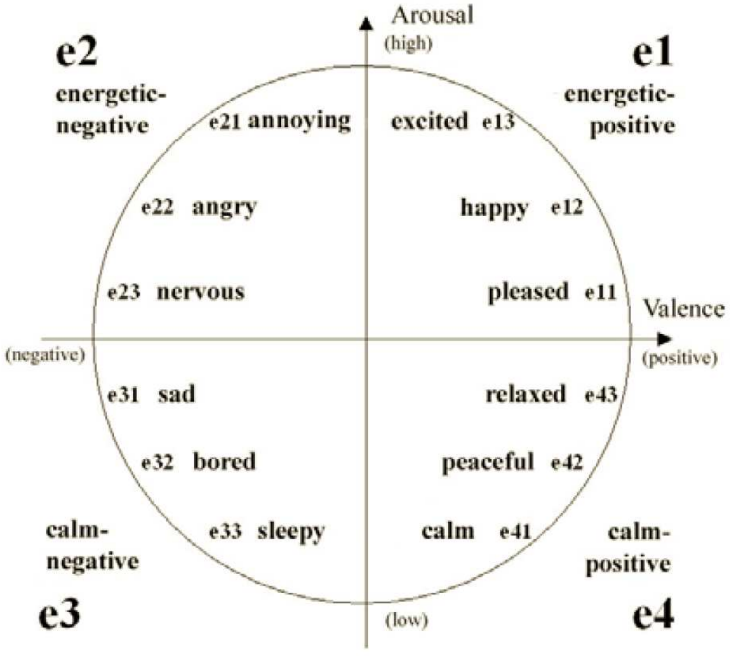

Figure 4. Thayer's arousal-valence emotion plane.

Abbreviation	Description
e1	energetic–positive
e2	energetic–negative
e3	calm–negative
e4	calm–positive

Figure 5. Description of the 4 emotion classes.

A person with formal education in music listens to each recording, and assigns one of the 4 emotion class label, specified in Figure 5., to each recording.

That produces a dataset with 3960 instances and 91 attributes.

3.3. Feature Descriptions

We use a unique set of timbre related features. Little research has been done using timbre in music classification [9, 10], and emotion classification [6]. Few systems exist for timbre information retrieval in the literature or market, which indicates it as a nontrivial and currently unsolved task [10].

The definition of timbre is: in acoustics and phonetics - the characteristic quality of a sound, independent of pitch and loudness, from which its source or manner of production can be inferred. Timbre depends on the relative strengths of its component frequencies; in music - the characteristic quality of sound produced by a particular instrument or voice; tone color. ANSI defines timbre as the attribute of auditory sensation, in terms of which a listener can judge that two sounds are different, though having the same loudness and pitch. It distinguishes different musical instruments playing the same note with the identical pitch and loudness. So it is the most important and relevant facet of music information. People discern timbre from speech and music in everyday life.

Musical instruments usually produce sound waves with frequencies, which are an integer (a whole number) multiples of each other. These frequencies are called harmonics, or harmonic partials. The lowest frequency is the fundamental frequency f0, which has close relation with pitch. The second and higher frequencies are called overtones. Along with fundamental frequency, these harmonic partials distinguish the timbre, which is also called tone color. The human aural distinction between musical instruments is based on the differences in timbre.

Timbre is rather subjective quality and not of much use for automatic sound timbre classification. To compensate, musical sounds must be very carefully parameterized to allow automatic timbre recognition.

Based on latest research in the area, MPEG published a standard group of features for digital audio content data. They are either in the frequency domain or in the time domain.

For those features in the frequency domain, a STFT (Short Time Fourier Transform) with Hamming window has been applied to the sample data. From each frame a set of instantaneous values is generated. We use the following timbre-related features from MPEG-7:

Spectrum Centroid - describes the center-of-gravity of a log-frequency power spectrum. It economically indicates the pre-dominant frequency range. We use *Log Power Spectrum Centroid*, and *Harmonic Spectrum Centroid*.

Spectrum Spread - is the Root of Mean Square value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame. Like Spectrum Centroid, it is an economic way to describe the shape of the power spectrum. We use *Log Power Spectrum Spread*, and *Harmonic Spectrum Spread*.

Harmonic Peaks - is a sequence of local peaks of harmonics of each frame. We use the *Top 5 harmonic peaks - Frequency*, and *Top 5 Harmonic Peaks - Amplitude*.

In addition, we use the *Fundamental Frequency* as a feature in this study.

3.4. Classification

We adopt data mining methods for classification. We use two classifiers - a Bayesian neural network and a J48 decision tree, and compare the results.

Neural networks process information with a large number of highly interconnected neurons working in parallel to solve a specific problem. Neural networks learn by example.

Decision trees represent a supervised approach to classification. It is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes.

4. Experimental results

We import the dataset into WEKA [15] data mining software for classification. We train two classifiers: Bayesian neural network and J48 decision tree. We use bootstrap testing. Results show Bayesian neural network has accuracy of 80.65%. J48 decision tree has accuracy of 75.27%. Summary results comparing the two classifiers are shown in Figure 6. The detailed results for Bayesian neural network and J48 decision tree are shown in Figure 7 and Figure 8 respectively.

classifier	no. correct instances	no. incorrect instances	accuracy
bayes neural net	3194	766	80.65%
J48 decision tree	2981	979	75.27%

Figure 6. Summary results - number of correctly and incorrectly classified instances, and accuracy percentage.

bayes	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.65	0.02	0.916	0.65	0.761	0.946	e4
	0.8	0.021	0.861	0.8	0.829	0.964	e3
	0.836	0.068	0.895	0.836	0.864	0.91	e2
	0.945	0.15	0.614	0.945	0.744	0.946	e1
avrg.	0.807	0.066	0.838	0.807	0.809	0.934	

Figure 7. Bayes neural network - detailed accuracy by class, and weighted average.

J48	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.768	0.118	0.683	0.768	0.723	0.921	e4
	0.317	0.002	0.957	0.317	0.477	0.903	e3
	0.816	0.198	0.739	0.816	0.776	0.9	e2
	0.914	0.049	0.825	0.914	0.867	0.981	e1
avrg.	0.753	0.12	0.773	0.753	0.738	0.922	

Figure 8. J48 decision tree - detailed accuracy by class, and weighted average.

5. Conclusions and directions for the future

We present a new content-based method for automatic recognition of emotions associated with music instrument recordings. We model emotions as 4 general class groups per Thayer's arousal-valence plane. We use a unique set of timbre related features. Little research has been done using timbre in music classification [9, 10], and emotion classification [6]. Timbre carries semantical information. Thus, it helps to bridge the semantical gap [2] between the object feature level and the human cognitive level of emotion perception, which has limited the MEC approaches until now. We train and test two classifiers. Results yield good recognition accuracy of 75%–80%.

This work contributes to solving the important problem of building music recommen-

dation systems; at a time, when the amount of music files available on the Internet is constantly growing, and personal music collections exceed practical limits.

This method can be applied towards mood-based systems [16] to aid users in finding objects of their liking within large online repositories.

In addition music recommendation systems, the retrieval of music by emotion is becoming an important task for various applications, such as: song selection in mobile devices [17], TV and radio programs, and music therapy.

This work can be extended by increasing the number of classes from 4 to 12 - to represent the detailed emotions within each general group per the Thayer's model in Figure 4. To achieve appropriate accuracy, more objects can be added to the dataset to enrich the learning process of the classifier with the increased number of classes.

Acknowledgements

This work is partially funded by the Center for Undergraduate Research and Scholarship (CURS) at the University of South Carolina Upstate, USA. The study is in collaboration with the Knowledge Discovery in Databases Laboratory at the University of North Carolina at Charlotte, USA.

References

[1] J. Grekow and Z.W. Ras, *Detecting emotion in classical music from MIDI files*, Foundations of Intelligent Systems, Proceedings of 18th International Symposium on Methodologies for Intelligent Systems (ISMIS'09), (Eds. J. Rauch et al), LNAI, Vol. 5722, Springer, Prague, Czech Republic, 2009, pp. 261–270.

[2] Y.H. Yang, Y.C. Lin, H.T. Cheng, I.B. Liao, Y.C. Ho, and H.H. Chen, *Toward multimodal music emotion classification*, in Proceedings of the 9th Pacific Rim Conference on Multimedia Advances in Multimedia Information Processing (PCM'08), Springer, 2008, pp. 70–79.

[3] 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 (TASL), vol. 16, no. 2, 2008, pp. 448-457.

[4] C. Busso, S. Lee, and S.S. Narayanan, *Analysis of emotionally salient aspects of fundamental frequency for emotion detection*, IEEE Transactions on Audio, Speech and Language Processing (TASL), vol. 17, no. 4, 2009, pp. 582–596.

[5] J. Kim and E. Andre, *Emotion recognition based on physiological changes in music listening.*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, 2008, pp. 2067–2083.

[6] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, *Multi-label classification of*

music into emotions, in Proceedings of the 9th International Conference of Music Information Retrieval (ISMIR 2008), Philadelphia, PA, USA, 2008, pp. 325–330.

[7] A. Tellegen, D. Watson, and L.A. Clark, *On the dimensional and hierarchical structure of affect*, Psychological Science, vol. 10 no.4, July 1999, pp. 297-303.

[8] T. Li and M. Ogihara. *Detecting emotion in music*, in Proceedings of the International Symposium on Music Information Retrieval, Washington D.C., USA, 2003, pp. 239-240.

[9] A.A. Tzacheva and K.J. Bell, *Music Information Retreival with Temporal Features and Timbre*, in Proceedings of 6th International Conference on Active Media Technology (AMT 2010), Toronto, Canada, LNCS 6335, 2010, pp. 212–219.

[10] W. Jiang, A. Cohen, and Z.W. Ras, *Polyphonic music information retrieval based on multi-label cascade classification system*, in Advances in Information and Intelligent Systems, Z.W. Ras, W. Ribarsky (Eds.), Studies in Computational Intelligence, Springer, vol. 251, 2009, pp. 117–137.

[11] K. Hevner, *Experimental studies of the elements of expression in music*, American Journal of Psychology, vol. 48, 1936, pp. 246-268.

[12] A. Wieczorkowska, P. Synak, Z.W.Ras, *Multi-label classification of emotions in music*, in Intelligent Information Processing and Web Mining, Proceedings of Advances in Soft Computing, IIS 2006 Symposium, Ustron, Poland, Springer, vol. 35, 2006, pp. 307–315.

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

[14] F. Opolko and J. Wapnick, *MUMS-McGillUniversityMasterSamples. CD's*, 1987.

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

[16] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.Y. Ma, *Musicsense: contextual music recommendation using emotional allocation modeling*, In Proceedings of the 15th international conference on Multimedia (MULTIMEDIA '07), 2007, pp. 553-556.

[17] M. Tolos, R. Tato, and T. Kemp, *Mood-based navigation through large collections of musical data*, in Proceedings of 2nd IEEE Consumer Communications and Networking Conference (CCNC 2005), 2005, pp. 71-75.