Extracting Emotions from Music Data

Alicja Wieczorkowska[1], Piotr Synak[1], Rory Lewis[2], and Zbigniew W. Raś[2,1]

[1] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
{alicja, synak}@pjwstk.edu.pl
[2] University of North Carolina, Charlotte, Computer Science Dept., 9201 University, City Blvd., Charlotte, NC 28223, USA
{rorlewis, ras}@uncc.edu

Abstract. Music is not only a set of sounds, it evokes emotions, subjectively perceived by listeners. The growing amount of audio data available on CDs and in the Internet wakes up a need for content-based searching through these files. The user may be interested in finding pieces in a specific mood. The goal of this paper is to elaborate tools for such a search. A method for the appropriate objective description (parameterization) of audio files is proposed, and experiments on a set of music pieces are described. The results are summarized in concluding chapter.

1 Introduction

Extracting information on emotions from music is difficult for many reasons. First of all, music itself is a subjective quality, related to culture. Music can be defined in various ways, for instance, as an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner [31], or as the art of combining sounds of voices or instruments to achieve beauty of form and expression of emotion [6]. Therefore, music is inseparably related to emotions. Musical structures itself communicate emotions, and also synthesized music aims at expressive performance [13], [19].

The experience of music listening can be considered within three levels of human emotion [12]:

- autonomic level,
- denotative (connotative) level, and
- interpretive (critical) level.

According to [17], music is heard:

- as sound. The constant monitoring of auditory stimuli does not switch off when people listen to music; like any other stimulus in the auditory environment, music is monitored and analyzed.
- as human utterance. Humans have an ability to communicate and detect emotion in the contours and timbres of vocal utterances; a musical listening experience does not annihilate this ability.

M.-S. Hacid et al. (Eds.): ISMIS 2005, LNAI 3488, pp. 456–465, 2005.

– in context. Music is always heard within the context of knowledge and environment, which can contribute to an emotional experience.
– as narrative. Listening to music involves the integration of sounds, utterances and context into dynamic, coherent experience. Such integration is underpinned by generic narrative processes (not specific to music listening).

Although music is such a delicate subject of scientific experiments, research has been already performed on automatic composition in given style [23], discovering principles of expressive music performance from real recordings [29], and labeling music files with metadata [24]. Also, research on recognizing emotions in audio files has been performed on speech data [7], [20], [27]. Emotions communicated in speech are quite clear. However, in experiments described by Dellaert et al. in [7] human listeners performed recognition of emotions in speech with about 80% correctness, in experiments with over a 1000 utterances from different speakers, classified into 4 categories: happy, sad, anger, and fear. The results obtained in machine classification were very similar, also reaching 80% correctness. Tato et al. [27] also obtained recognition rate approaching 80%, for 3 classes regarding levels of activation: high (angry, happy), medium (neutral), and low (sad, bored). The database of about 2800 utterances was used in these experiments.

Research on discovering emotions from music audio data have also been recently performed [18]. Li and Ogihara reported in [18] on detecting emotions in music, using 499 sound files representing 13 classes, labeled by a single subject. The accuracy ranged for particular classes from about 50% to 80%. Since emotions in music are more difficult to discover than in speech, and even the listener labeling the data reported difficulties with performing the classification task, the obtained results are very good.

2 Audio Data Parameterization

Parameterization of audio data for classification purposes can be based on various descriptors. For instance, Tato et al. in [27] applied speech-specific prosodic features, derived from pitch, loudness, and duration, which they associated with the activation or arousal dimension, and quality features, i.e. phonation type, articulation manner, voice timbre, which they associated with the evaluation and pleasure dimension. They achieved most important results for the speaker-independent recognition and three classes, with a accuracy about 80%.

In case of music audio data, other descriptors are used, see for instance [25], [28], [30]. These features include structure of the spectrum, time domain features, and also time-frequency description. Since the research on automatic detection of emotions in music is very recent, there is no significant comparison of descriptor sets and their performance for this purpose. Li and Ogihara applied parameters provided in [28], describing timbral texture features, rhythmic content features, and pitch content features. The dimension of the final feature vector was 30.

In our research, we based on assumption that emotions depend on chords and timbre, using Western music as audio samples. Long analyzing frame (32768

samples taken from the left channel of stereo recording, for 44100 Hz sampling frequency and 16-bit resolution), in order obtain more precise spectral bins, and to describe longer time fragment. Hanning window was applied to the analyzed frame. Spectral components up to 12 kHz were taken into account.

The following set of descriptors was calculated [30]:

- *Frequency*: dominating fundamental frequency of the sound
- *Level*: maximal level of sound in the analyzed frame
- $Tristimulus1, 2, 3$: Tristimulus parameters calculated for *Frequency*, given by [26]:

$$Tristimulus1 = \frac{A_1^2}{\sum_{n=1}^{N} A_n^2} \tag{1}$$

$$Tristimulus2 = \frac{\sum_{n=2,3,4} A_n^2}{\sum_{n=1}^{N} A_n^2} \tag{2}$$

$$Tristimulus3 = \frac{\sum_{n=5}^{N} A_n^2}{\sum_{n=1}^{N} A_n^2} \tag{3}$$

  where A_n denotes the amplitude of the n^{th} harmonic, N is the number of harmonics available in spectrum, $M = \lfloor N/2 \rfloor$ and $L = \lfloor N/2 + 1 \rfloor$
- *EvenHarm* and *OddHarm*: Contents of even and odd harmonics in the spectrum, defined as

$$EvenHarm = \frac{\sqrt{\sum_{k=1}^{M} A_{2k}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}} \tag{4}$$

$$OddHarm = \frac{\sqrt{\sum_{k=2}^{L} A_{2k-1}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}} \tag{5}$$

- *Brightness*: brightness of sound - gravity center of the spectrum, defined as

$$Brightness = \frac{\sum_{n=1}^{N} n\, A_n}{\sum_{n=1}^{N} A_n} \tag{6}$$

- *Irregularity*: irregularity of spectrum, defined as [9], [16]

$$Irregularity = \log\left(20 \sum_{k=2}^{N-1} \left| \log \frac{A_k}{\sqrt[3]{A_{k-1} A_k A_{k+1}}} \right| \right) \tag{7}$$

- $Frequency1, Ratio1, ..., 9$: for these parameters, 10 most prominent peaks in the spectrum are found. The lowest frequency within this set is chosen as $Frequency1$, and proportions of other frequencies to the lowest one are denoted as $Ratio1, ..., 9$

– $Amplitude1, Ratio1, ..., 9$: the amplitude of $Frequency1$ in decibel scale, and differences in decibels between peaks corresponding to $Ratio1, ..., 9$ and $Amplitude1$. These parameters describe relative strength of the notes in the music chord.

Since the emotions in music also depend on the evolution of sound, it is recommended to observe changes of descriptor values in time, especially with respect to music chords, roughly represented by parameters $Frequency1, Ratio1, ..., 9$; we plan such extension of our feature set in further experiments.

3 Data Labeling

One of difficulties in experiments on recognition of emotions in music is labeling of the data. The emotions can be described in various ways. One of the possibilities is presented in Figure 1, proposed by Hevner [11]. This labeling consists of 8 classes, although not all adjectives in a single group are synonyms, see for instance pathetic and dark in class 2.

Other labeling is also used. For instance, Li and Ogihara in [18] use 13 classes, each labeled by one, two, or three adjectives. These groups are based on redefined

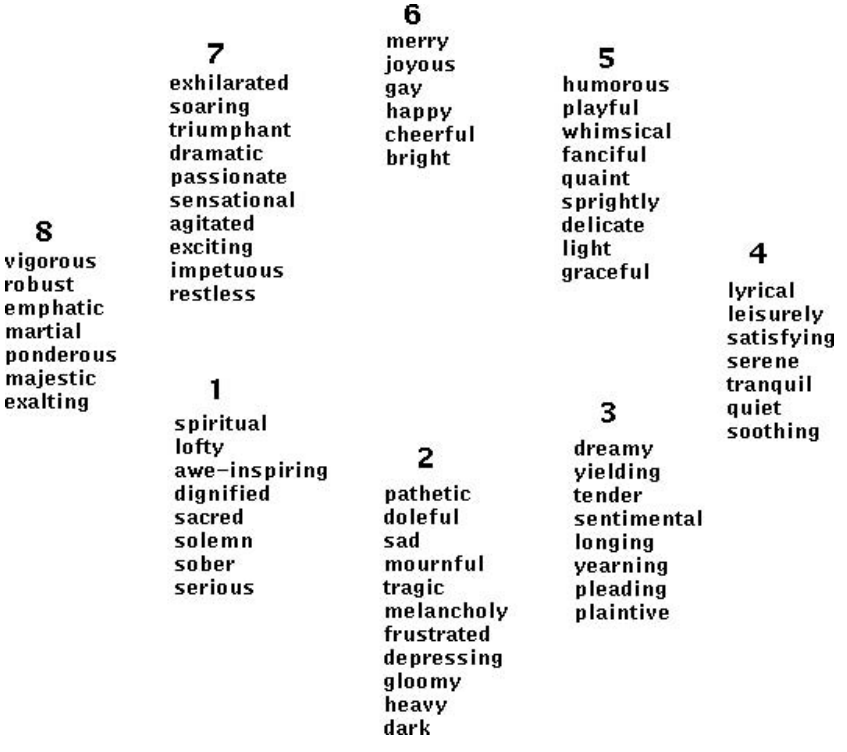

Fig. 1. Adjective Circle according to K. Hevner [11]

(by Farnsworth) Hevner adjectives, supplemented with 3 additional classes. Altogether, the following adjectives were used in this research: cheerful, gay, happy, fanciful, light, delicate, graceful, dreamy, leisurely, longing, pathetic, dark, depressing, sacred, spiritual, dramatic, emphatic, agitated, exciting, frustrated, mysterious, spooky, passionate, and bluesy [18].

Emotions can be also represented in 2 or 3 dimensional space, thus allowing labeling along the chosen axes. For instance, Tato et al. [27] performed research on detecting emotions in speech along activation dimension, using the following adjectives: angry, happy, neutral, sad, bored.

Apart from problem with choosing the appropriate representation, another issue to deal with is labeling the data by subjects, i.e. human listeners. Since emotions may vary from subject to subject, various listeners may provide different labeling. Labeling with emotions is inevitably subjective, and it is very difficult to compare objectively the classification results [5]. Also, even a single subject may have difficulties with choosing the most appropriate label, so multiple labels are usually needed for the same sample of the data. In our research, we decided to use single labeling of classes by a single subject. Such a setup allows checking exactly the quality of parameterization chosen as a tool for finding dependencies between subjective and objective audio description. When this is done, multiple labeling and multi-subject assessment can be performed, in order to check consistency of perceiving emotions from subject to subject.

4 Classifiers

The research on classification of audio data of various types has already been performed worldwide, although automatic detecting emotions is rather recent topic of research. Various classification algorithms were used, including:

- k-nearest neighbors (k-NN) [10], [14], [30],
- Gaussian classifiers [4], [8], [21],
- rough-set based algorithms [30],
- artificial neural networks (NN) [15],
- support vector machines [18],
- hidden Markov models [2],
- and other algorithms [10].

Since k-NN is one of more popular and at the same time successful algorithm is k-NN, we also decided to apply it. In this algorithm, the class of unknown sample is assigned on the basis of k nearest neighbors of known origin. Number k of considered neighbors is one of parameters of our classifier. We believe this algorithms is quite well suited to this task, since it reflects similarity or dissimilarity between samples representing the same or various classes respectively, therefore matching our subjective classification task.

5 Experiments and Results

The research described in this paper was performed using specially collected data. The audio files were gathered from personal collection and labeled by a single male subject, one of the authors (R. Lewis), a practicing musician. We decided to follow labeling used by Li and Ogihara in [18], since the class names they used cover a big set of possible emotions in music, yet represented by reasonable number of classes. Therefore, the experiments we performed were based on the classifying audio data into the following 13 classes:

- cheerful, gay, happy,
- fanciful, light,
- delicate, graceful,
- dreamy, leisurely,
- longing, pathetic,
- dark, depressing,
- sacred, spiritual,
- dramatic, emphatic,
- agitated, exciting,
- frustrated,
- mysterious, spooky,
- passionate,
- bluesy.

The collections consists of 303 music pieces. Each class is represented by 6 (dark and spooky) - 86 (dramatic) pieces, from CDs or bought iTunes [1]. Number of pieces for each class is shown in Figure 2.

Since some of the classes are underrepresented in comparison with the others, we decided to join the data into 6 superclasses as follows (see [18]):

1. happy and fanciful,
2. graceful and dreamy,
3. pathetic and passionate,
4. dramatic, agitated, and frustrated,

Class	No. of objects	Class	No. of objects
Agitated	16	Graceful	14
Bluesy	18	Happy	24
Dark	6	Passionate	18
Dramatic	88	Pathetic	32
Dreamy	20	Sacred	17
Fanciful	34	Spooky	7
Frustrated	17		

Fig. 2. Representation of classes in the collection of musical recordings for the research on automatic classifying emotions

Class	No. of objects	k-NN	Correctness
1. happy, fanciful	57	k=11	81.33%
2. graceful, dreamy	34	k=5	88.67%
3. pathetic, passionate	49	k=9	83.67%
4. dramatic, agitated, frustrated	117	k=7	62.67%
5. sacred, spooky	23	k=7	92.33%
6. dark, bluesy,	23	k=5	92.33%

Fig. 3. Results of automatic classification of emotions for the investigated database

5. sacred and spooky,
6. dark and bluesy.

The pieces were ripped into MP3 format using iTunes and Windows Media Player [22]. Before parameterization, they were converted to au/snd format. As we mentioned in Section 2, sampling frequency 44100 Hz was chosen and 32768 samples (2^{15}) frame length for analysis was selected. Parameterization was performed at the signal frame taken 30 seconds after the beginning of the piece; since Li and Ogihara in [18] also parameterized the same fragment, thus allows more reliable comparison of results.

Twenty-nine parameters were extracted for such a frame, according to description given in Section 2. The spectrum was limited to 11025 Hz (half of available spectrum) and to no more than 100 harmonics, since higher harmonics did not contribute significantly to the spectrum.

The obtained data set was next used in experiments with classification, using k-NN algorithm with k varying within range 1..20. The best k in each experiment was chosen. Standard CV-5 cross-validation was applied to test the classifier, i.e. 20% of data was removed from the set to train the classifier and then used for the tests; such procedure was repeated 5 times and the result was averaged. Since Li and Ogihara always divided data into 2 classes (the tested one against the rest of the data), we also followed the same technique, so these experiments and results can be quite easily compared. Such binary technique is a good basis for further classification, i.e. for construction of a general classifier, based on a set of binary classfiers [3]. The results of our experiments are presented in Figure 3.

To compare with, Li and Ogihara obtained accuracy ranging from 51% to 80% (for various classes), with use of 50% of data for training and the remaining 50% of data for testing of 599 sound files, with labeling into 13 classes, and then into 6 the same classes, also assigned by a single subject (39 year old male, i.e. comparable with our subject) and 30-element feature vector [18]. Our general accuracy in test with 6 classes tested in parallel (i.e. classic k-NN test) yielded 37% correctness. These result suggest necessity of further work. Especially, more balanced representation of all classes should be gathered, and evolution of sound features in time should be observed, in order to improve the quality of audio description.

We also performed similar experiments on the audio data obtained from M. Ogihara. These data (872 audio files) are presented in Figure 4. The best results

Class	No. of objects	Class	No. of objects
Agitated	74	Graceful	45
Bluesy	66	Happy	36
Dark	31	Passionate	40
Dramatic	101	Pathetic	155
Dreamy	46	Sacred	11
Fanciful	38	Spooky	77
Frustrated	152		

Fig. 4. Representation of classes in Ogihara's database

Class	No. of objects	Correctness
1. happy, fanciful	74	95.97%
2. graceful, dreamy	91	89.77%
3. pathetic, passionate	195	71.72%
4. dramatic, agitated, frustrated	327	64.02%
5. sacred, spooky	88	89.88%
6. dark, bluesy,	97	88.80%

Fig. 5. Results of automatic classification of emotions for the Ogihara's database

of binary classification for these data, grouped into 6 classes as previously, were obtained in k-NN experiments for $k = 13$. The results are presented in Figure 5. As we can see, the results are comparable or even better than for our database, albeit the size of the data was increased. These results convince us that our parameterization and classification perform quite well, although the final accuracy still is not satisfying, so our methods should be improved in further research.

6 Summary and Conclusions

Discovering emotions in music is a difficult issue for many reasons. First of all, emotions perceived with music are subjective and depend on numerous factors. Additionally, the same piece may evoke various emotions not only for various human listeners, but even the same person may feel various emotions when listening to this same piece of music. Therefore, our experiments may be extended if we ask other subjects to label the same data, here labeled by a single subject.

The purpose of our research was to perform parameterization of audio data for the purpose of automatic recognition of emotions in music, and the special collection of music pieces was gathered and turned into a data set. In order to check quality of parameterization, single labeling by a single subject was performed. As a classification algorithm k-NN was chosen, since this algorithm performed well in other experiments regarding classification of music, and since it also allows easy extension in case of increasing the size of the data set, which is our plan for the nearest future.

We are going to continue our investigations with use of other classifiers. Since support vector machines gain increasing popularity in audio classification research, we would also like to apply this method to our data. Also, neural networks could be applied in this research, since neural classifiers may perform well in such tasks, and decision trees. Additionally, we can analyze sequences of short audio samples, in order to follow changes of emotions in a music piece. Since we would like to obtain more universal classification, we also plan further listening experiments and labeling of the same data by other subjects. Although multi-label classification is more challenging in the training phase, the results should be more useful, i.e. better suited to emotions that various user may feel while listening to the same music.

Acknowledgements

This research was partially supported by the National Science Foundation under grant IIS-0414815, and by the Research Center at the Polish-Japanese Institute of Information Technology, Warsaw, Poland.

The authors express thanks to Professor Mitsunori Ogihara from the University of Rochester for supplying them with his valuable audio database.

References

1. Apple: iTunes (2004). http://www.apple.com/itunes/
2. Batlle, E. and Cano, P.: Automatic Segmentation for Music Classification using Competitive Hidden Markov Models. Proceedings of International Symposium on Music Information Retrieval, Plymouth, MA (2000). Available at http://www.iua.upf.es/mtg/publications/ismir2000-eloi.pdf
3. Berger, A.: Error-correcting output coding for text classification. IJCAI'99: Workshop on machine learning for information filtering. Stockholm, Sweden (1999). Available at http://www-2.cs.cmu.edu/ aberger/pdf/ecoc.pdf
4. Brown, J. C.: Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. J. Acoust. Soc. of America **105** (1999), 1933–1941
5. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics **22 (2)** (1996), 249–254. Available at http://homepages.inf.ed.ac.uk/jeanc/squib.pdf
6. Cross, I.: Music, cognition, culture and evolution. Annals of the New York Academy of Sciences, **930** (2001), 28–42. Available at http://www-ext.mus.cam.ac.uk/ ic108/PDF/IRMCNYAS.pdf
7. Dellaert, F., Polzin, T., Waibel, A.: Recognizing Emotion in Speech. Proc. ICSLP 96 **3** (1996), 1970–1973.
8. Eronen, A. and Klapuri, A.: Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2000, Plymouth, MA (2000). 753–756
9. Fujinaga, I., McMillan, K.: Realtime recognition of orchestral instruments. Proceedings of the International Computer Music Conference (2000) 141–143

10. Herrera, P., Amatriain, X., Batlle, E., and Serra X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In Proc. of International Symposium on Music Information Retrieval ISMIR 2000, Plymouth, MA (2000)
11. Hevner, K.: Experimental studies of the elements of expression in music. American Journal of Psychology **48** (1936) 246–268
12. Huron, D.: Sound, music and emotion: An introduction to the experimental research. Seminar presentation, Society for Music Perception and Cognition Conference. Massachusetts Institute of Technology, Cambridge, MA (1997).
13. Juslin, P., Sloboda, J. (eds.): Music and Emotion: Theory and Research. Series in Affective Science. Oxford University Press (2001)
14. Kaminskyj, I.: Multi-feature Musical Instrument Classifier. MikroPolyphonie **6** (2000).)nline journal at http://www.mikropol.net/
15. Kostek, B. and Czyzewski, A.: Representing Musical Instrument Sounds for Their Automatic Classification. J. Audio Eng. Soc. **49(9)** (2001). 768–785
16. Kostek, B, Wieczorkowska, A.: Parametric Representation Of Musical Sounds. Archives of Acoustics **22, 1** (1997) 3–26
17. Lavy, M. M.: Emotion and the Experience of Listening to Music. A Framework for Empirical Research. PhD. dissertation, Jesus College, Cambridge (2001).
18. Li, T., Ogihara, M.: Detecting emotion in music. 4th International Conference on Music Information Retrieval ISMIR, Washington, D.C., and Baltimore, MD (2003).
19. Mantaras, R. L. de, Arcos, J. L.: AI and Music. From Composition to Expressive Performance. AI Magazine, Fall 2002 (2002) 43–58
20. Marasek, K.: private communication (2004).
21. Martin, K. D., Kim, Y. E.: Musical instrument identification: A pattern-recognition approach. 136 meeting of the Acoustical Soc. of America, Norfolk, VA (1998)
22. Microsoft Corp.: Windows Media Player (2004). http://www.microsoft.com/
23. Pachet, F.: Beyond the Cybernetic Jam Fantasy: The Continuator. IEEE Computers Graphics and Applications, Jan./Feb. 2004, spec. issue on Emerging Technologies.
24. Pachet, F.: Knowledge Management and Musical Metadata. In: Schwartz, D. (Ed.), Encyclopedia of Knowledge Management. Idea Group (2005).
25. Peeters, G. Rodet, X.: Automatically selecting signal descriptors for Sound Classification. ICMC 2002 Goteborg, Sweden (2002)
26. Pollard, H. F., Jansson, E. V.: A Tristimulus Method for the Specification of Musical Timbre. Acustica **51** (1982) 162–171
27. Tato, R., Santos, R., Kompe, R., Pardo, J. M.: Emotional Space Improves Emotion Recognition. 7th International Conference on Spoken Language Processing ICSLP 2002, Denver, Colorado (2002).
28. Tzanetakis, G., Cook, P.: Marsyas: A framework for audio analysis. Organized Sound **4(3)** (2000) 169-175.
29. Widmer, G.: Discovering Simple Rules in Complex Data: A Meta-learning Algorithm and Some Surprising Musical Discoveries. Artificial Intelligence **146(2)** (2003)
30. Wieczorkowska, A., Wroblewski, J., Slezak, D., Synak, P.: Application of temporal descriptors to musical instrument sound recognition. Journal of Intelligent Information Systems **21(1)**, Kluwer (2003), 71–93
31. WordIQ Dictionary (2004). The Internet http://www.wordiq.com/dictionary/