

Some Issues on Detecting Emotions in Music

Piotr Synak and Alicja Wieczorkowska

Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
synak,alicja@pjwstk.edu.pl

Abstract. Investigating subjective values of audio data is both interesting and pleasant topic for research, gaining attention and popularity among researchers recently. We focus on automatic detection of emotions in songs/audio files, using features based on spectral contents. The data set, containing a few hundreds of music pieces, was used in experiments. The emotions are grouped into 13 or 6 classes. We compare our results with tests on human subjects. One of the main conclusions is that multi-label classification is required.

Keywords: music information retrieval, sound analysis.

1 Introduction

Automatic recognition of emotions in music is a difficult task because of many reasons. First of all, there is no any universal way or any standard of describing sound files. Several kind of descriptors can be generated from sounds without any warranty that they reflect any emotions. Moreover, especially in the case of emotions, any classification (also subjective one) can be ambiguous – every subject may classify emotions in a little bit different way. However, for listeners from similar cultural background, one may expect to obtain similar classification. Therefore, we considered this topic worth investigations.

We present some initial experiments performed on a database of 870 sound files classified to 13 or 6 classes of emotions. To describe the files we used a number of spectral descriptors. In the paper, we discuss the obtained results and draw the conclusions how to detect the emotions better.

2 Data Parametrization

Automatic parametrization of audio data for classification purposes is hard because of ambiguity of labeling and subjectivity of description. However, since a piece of music evokes similar emotions in listeners representing the same cultural background, it seems to be possible to obtain parametrization that can be used for the purpose of extracting emotions. Our goal was to check how numerical parameters work for classification purposes, how good or low is classification accuracy, and how it is comparable with human performance.

Objective descriptors of audio signal characterize basic properties of the investigated sounds, such as loudness, duration, pitch, and more advanced properties, describing frequency contents and its changes over time. Some descriptors come from speech processing and include prosodic and quality features, such as phonation type, articulation mannered etc. [12]. Such features can be applied to detection of emotions in speech signal, but not all of them can be applied to music signals, which require other descriptors. Features applied to music signal include structure of the spectrum - timbral features, time domain features, time-frequency description, and higher-level features, such as rhythmic content features [7], [9], [13], [14].

When parameterizing music sounds for emotion classification, we assumed that emotions depend, to some extent, on harmony and rhythm. Since we deal with audio, not MIDI files, our parametrization is based on spectral contents (chords and timbre). Western music, recorded stereo with 44100 Hz sampling frequency and 16-bit resolution was used as audio samples. We applied long analyzing frame, 32768 samples taken from the left channel, in order obtain more precise spectral bins, and to describe longer time fragment. Hanning window was applied, and spectral components calculated up to 12 kHz and no more than 100 partials, since higher harmonics did not contribute significantly to the spectrum.

The following set of 29 audio descriptors was calculated for our analysis window [14]:

- *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 [10]:

$$Tristimulus1 = \frac{A_1^2}{\sum_{n=1}^N A_n^2} \quad (1)$$

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

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

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

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

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

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

$$Brightness = \frac{\sum_{n=1}^N n A_n}{\sum_{n=1}^N A_n} \quad (6)$$

- *Irregularity*: irregularity of spectrum, defined as [5], [6]

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

- *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.

3 Experiment Setup

Investigations on extracting emotions from music data were performed on a database of 870 audio samples. The samples represented 30 seconds long excerpts from songs and classic music pieces. This database was created by Dr. Rory A. Lewis from the University of North Carolina at Charlotte. Therefore, all audio file were labeled with information about emotions by a single subject. The pieces were recorded in MP3 format and next converted to au/snd format for parametrization purposes. Sampling frequency 44100 Hz was chosen. Parametrization was performed for 32768 samples (2^{15}) frame length. The data set is divided into the following 13 classes, covering wide range of emotions [7]:

1. frustrated,
2. bluesy, melancholy,
3. longing, pathetic,
4. cheerful, gay, happy,
5. dark, depressing,
6. delicate, graceful,
7. dramatic, emphatic,
8. dreamy, leisurely,
9. agitated, exciting, enthusiastic,
10. fanciful, light,
11. mysterious, spooky,
12. passionate,
13. sacred, spiritual.

| 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. 1. Representation of classes in the 870-element database

Number of samples in each class is shown in Figure 1.

Some classes are underrepresented, whereas others are overrepresented in comparison with the average number of objects in a single class. Moreover, labeling of classes is difficult in some cases, since the same piece may evoke various emotions. Therefore, we decided to join the data into 6 superclasses as follows (see [7]):

1. happy and fanciful,
2. graceful and dreamy,
3. pathetic and passionate,
4. dramatic, agitated, and frustrated,
5. sacred and spooky,
6. dark and bluesy.

The classification experiments were performed using k -NN algorithm, with k varying within range 1..20, and the best k in each experiment was chosen. We decided to use k -NN because in the first experiments it outperformed classifiers of other types. For training purposes, 20% of the data set was removed and then used as test data after finishing training; this procedure was repeated 5 times (i.e., standard CV-5 procedure was applied). Next, the results were averaged. In order to compare results with Li and Ogihara [7], we performed experiments for each class separately, recognizing them in a binary way - one class against the rest of the data. The binary classification can be a good basis for construction of a general classifier, based on a set of binary classifiers [1].

4 Results

The experiments described in the previous section were first performed on a smaller data set, containing 303 objects, as presented in Figure 2. These experiments yielded results presented in Figure 3.

The results can be well compared with the results obtained by Li and Ogihara [7]. They obtained accuracy ranging from 51% to 80% for various classes and 30-element feature vector, with use of 50% of data for training and the remaining

| 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 303 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 303-element database using k -NN

50% of the data set, consisting of 599 audio files, also labeled by a single subject into the same 3 classes, and then into 6 classes, as described in section 3. We also performed experiments for the same 6 classes, using k -NN classifier, i.e., examining all 6 classes in parallel. These experiments yielded 37% correctness (and 23.05% for 13 classes), suggesting that further work was needed. Since we suspected that uneven number of objects in classes and not too big data set could hinder classification, the 870-element data set was used in further experiments.

The results of experiments with binary classification performed on the full data set, containing 870 audio files, are presented in Figure 4. The best results of experiments were obtained in k -NN for $k = 13$. As we can see, the results have been even improved comparing to the small data set. However, general classification for all classes examined in parallel was still low, comparable with results for 303-element data set, since we obtained 20.12% accuracy for 13 classes and 37.47% for 6 classes.

Because of the low level of accuracy in general classification, we decided to compare the results with human performance. Two other subjects with musical background were asked to classify a test set of 39 samples, i.e., 3 samples for each class. The results convinced us that the difficulty is not just in the parametrization or method of classification, since the correctness of assessment

| 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. 4. Results of automatic classification of emotions for the 870-element database

yielded 24.24% and 33.33%, differing essentially on particular samples. This experiment suggests that multi-class labeling by a few subjects may be needed, since various listeners may perceive various emotions while listening to the same file, even if they represent the same cultural and musical background.

5 Multi-class Labeling

In our experiments we used a database of music files collected by a professional musician acting as an expert. Every file was labeled and classified to exactly one class representing particular kind of emotion. One of the first remarks of the expert was that in several cases it is impossible to classify a song to exactly one class. First of all, the nature of the song and the melody can be labeled by more than one adjective. Secondly, labeling is very subjective and different people may use various associations. This is because the perception of sounds by humans is not uniform. The perception can be dominated by different factors, e.g., by particular instrument or by vocal, and thus, different labels can be attached to the same piece of sound.

The results of our initial experiments (i.e., with 13 decision classes) one may interpret as not satisfactory. One reason of low results is that the set of used descriptors is relatively small. However, we claim that the most important factor is that the files were initially classified to single classes only. To confirm this we conducted the following experiment. We asked another two musicians (female and male) to classify the sound files to the same categories as the initial labeling. As we stated in the previous section, the results were very surprising in that the quality of recognition by a human was worse than one obtained by k -NN.

From the discussions with the experts it follows that the main difficulty while performing the classification was that they had to choose one class only, whilst in most cases they found at least two class labels appropriate. Therefore, we suggest to use multi-class labeling, i.e., to allow labeling each piece of sound with any number of labels.

The data that can be classified to more than one class are known in the literature as multi-label data [2, 8, 4]. This kind of data is often being analyzed

in text mining and scene classification, where text documents or pictures may have been attached several labels describing their contents.

There are several problems related to multi-label data analysis, including: selecting training model with multi-label data, using testing criteria, and evaluating multi-label classification results.

5.1 Training models

One of the basic questions of training phase of classifier's induction is how to use training examples with multiple labels? There are a few models commonly used.

The simplest model (*MODEL-s*) assumes labeling of data by using single label – the one which is most likely.

MODEL-i assumes ignoring all the cases with more than one label. That means that there can no data to be used in the training phase if there are no data with single label.

In *MODEL-n* there are created new classes for each combination of labels occurring in the training sample. The main problem of this model is that the number of classes easily becomes very large, especially when we consider not only two, but three and more labels attached to one sample. Therefore, the data become very sparse, and, as result of that, several classes can have very few training samples.

The most efficient model seems to be *MODEL-x*, **cross**-training, where samples with many labels are used as positive examples, and not as negative examples, for each class corresponding to the labels.

5.2 Testing criteria

We assume that we build models for each base class only, and not for combination of classes (*MODEL-n*) because of sparseness of data as discussed above. As an exemplary classifier we use Support Vector Machines (SVM) [3] as they are recognized to give very good results in text and scene classification, i.e., in multi-label data.

Now, let us see how can we obtain multiple labels from the outputs of each of the models. In standard 2-class SVM the positive (negative) output of a SVM for a testing object means that it is a positive (negative) example. In the case of multi-class problems there are several SVMs built – one for each class. The highest positive output of SVMs determines the class of a testing object. However, it can happen that no SVM gives positive output. This approach can be extended to multi-label classification.

Let us consider the following three testing (labeling) criteria.

P-criterion labels the testing object with all classes corresponding to positive output of SVM. If no output is positive than the object is unlabeled.

T-criterion works similarly to *P-criterion*, however, if no output is positive than the top value is used for labeling.

C-criterion evaluates top values that are close each other no matter whether they are positive or negative.

5.3 Evaluating classification results

Evaluation of results differs from the classical case of single-label classification, where testing object is classified either correctly or incorrectly. In the case of multi-label data classification we can have more cases. If all the labels assigned to a testing object are proper then it is classified correctly – if all are wrong then incorrectly. However, what makes it different from single-label classification, only some of the labels can be attached properly – this is the case of partial correctness.

Thus, except standard measures of quality of classification like precision or accuracy, we need additional ones that take into account also partial correctness. Some examples of such measures, for example one-error, coverage, and precision, have been proposed in the literature (see, e.g., [11]). In [2] there are proposed two methods, *α -evaluation* and *base class evaluation* of multi-label classifier evaluation that make it possible to analyze results of classification in a wide range of settings.

6 Conclusions and Future Work

Difficult task of automatic recognition of emotions in music pieces was investigated in our research. The purpose of this investigations was not only testing how numerical parameters perform in objective description of subjective features, but also assessment of the recognition accuracy, and comparison of results with human subjects. The obtained accuracy is not high, but is of the same quality as human assessment. Since humans differ in their opinions regarding emotions evoked by the same piece, inaccuracies in automatic classification are not surprising.

In the next experiments we plan to apply multi-class labeling of sounds and develop the methodology for multi-label data classification.

We also plan to extend the set of descriptors used for sounds parametrization. In particular, we want to investigate how values of particular parameters change with time and how it is related to any kind of emotions.

Acknowledgements

This research was partially supported by the National Science Foundation under grant IIS-0414815, by the grant 3 T11C 002 26 from Ministry of Scientific Research and Information Technology of the Republic of Poland, and by the Research Center at the Polish-Japanese Institute of Information Technology, Warsaw, Poland.

The authors express thanks to Dr. Rory A. Lewis from the University of North Carolina at Charlotte for elaborating the audio database for research purposes.

References

1. 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>
2. Boutell, M., Shen, X., Luo, J., Brown, C.: Multi-label Semantic Scene Classification. Technical Report, Dept. of Computer Science, U. Rochester, (2003).
3. Burges, C.J.: A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* **2(2)** (1998) 121–167.
4. Clare, A., King, R.D.: Knowledge Discovery in Multi-label Phenotype Data. *Lecture Notes in Computer Science* **2168** (2001) 42–53.
5. Fujinaga, I., McMillan, K.: Realtime recognition of orchestral instruments. *Proceedings of the International Computer Music Conference* (2000) 141–143
6. Kostek, B., Wierzchowska, A.: Parametric Representation Of Musical Sounds. *Archives of Acoustics* **22, 1** (1997) 3–26
7. Li, T., Ogihara, M.: Detecting emotion in music. 4th International Conference on Music Information Retrieval ISMIR, Washington, D.C., and Baltimore, MD (2003). Available at <http://ismir2003.ismir.net/papers/Li.PDF>
8. McCallum, A.: Multi-label Text Classification with a Mixture Model Trained by EM. AAAI'99 Workshop on Text Learning, (1999).
9. Peeters, G., Rodet, X.: Automatically selecting signal descriptors for Sound Classification. ICMC 2002 Goteborg, Sweden (2002)
10. Pollard, H. F., Jansson, E. V.: A Tristimulus Method for the Specification of Musical Timbre. *Acustica* **51** (1982) 162–171
11. Schapire, R., Singer, Y.: Boostexter: a boosting-based system for text categorization. *Machine Learning* **39(2/3)** (2000) 135–168
12. 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).
13. Tzanetakis, G., Cook, P.: Marsyas: A framework for audio analysis. *Organized Sound* **4(3)** (2000) 169-175. Available at <http://www-2.cs.cmu.edu/~gtzan/work/pubs/organised00gtzan.pdf>
14. Wierzchowska, A., Wroblewski, J., Synak, P., Slezak, D.: Application of temporal descriptors to musical instrument sound recognition. *Journal of Intelligent Information Systems* **21(1)**, Kluwer (2003), 71–93