# Extraction of Emotional Content from Music Data

Marcin Bartoszewski Halina Kwasnicka Urszula Markowska-Kaczmar Pawel B. Myszkowski
Wroclaw University of Technology
Institute of Applied Informatics
Wyb. Wyspiańskiego 27, 50-370 Wroclaw, Poland
urszula.markowska-kaczmar@pwr.wroc.pl

## Abstract

*This paper presents the system for automatic emotion detection from music data stored in MIDI format files. First, the piece of music is divided into independent segments that potentially represent different emotional states. For this task the method of segmentation is used. The most important part is a features extraction from the music data. On this basis similar emotional parts are grouped by clustering algorithm. Music domain knowledge is used to extract features which are then grouped hierarchically by agglomerative clustering algorithm. Obtained results are visualised by the SOM neural network. The results prove that in the music structure exist features that affect on the human emotion. A novelty of the proposed approach lies in extracted features that discriminate emotional charge of music and application of agglomerative clustering.*

## 1. Introduction

In the last years, the number of researches referring to emotion in music has grown up significantly. This phenomenon is stimulated by development of technology delivering the equipment that needs an efficient choice of songs from a huge discography. Current mechanisms of data search bases are insufficient for such huge data sets.

Though query by humming solves the problem when a title of a song is unknown, but generates new complications. The main disadvantage of this approach is an efficiency, which is proportional to the ability of clear singing. Sometimes a part of a piece of music is very difficult to imitate.

Query by example seems to be a remedy for this problem but it requires existence of the appropriate pattern being a reference. A combination of selected methods does not solve all problems.

A new trend is an emotional content extraction from music. Unfortunately, this approach is complex and generates some new difficulties. The main argument against its application in the system of knowledge extraction is its subjectivity and a strong connection with character and personality of human but our intuition says that between emotions we can distinguish fuzzy classes that are perceived by people in the similar way. This hypothesis can be partially confirmed by musicotheraphy. Experiments have shown that there is a relationship between emotional reaction and a concrete musical structure.

A mood taxonomy is the next problematic aspect because of its expression and perception. The next problem is a diversity of music recording formats. Each standard has it own restrictions, i.e in MIDI files there is no information about acoustic and characteristic tones of instruments. Audio recording causes distortions in many cases. Nevertheless already mentioned problems, extraction of emotion from music seems to be very promising approach to efficient search of music pieces in the huge discography.

The aim of the research presented in this paper was to prove the existence of specific relationships in music that are responsible for the moulding emotional charge. The problem domain has been narrowed to strict music structure. Other music features like the sounding of musical instrument or reverberation are passed over. To realize this task the system of the automatic detection of emotional charge has been built.

The paper consists of four sections. In the next section the related works are described. Then, the details of our automatic emotional content detection system are presented. In the subsequent section the experimental results of the method are presented. The summary is presented in conclusion part.

## 2. Related works

The typical approach to emotional content extraction from music consists of the following steps:

- *Domain recognition* – this is an initial step, where the aim and the range of the problem is defined. In this step

978-0-7695-3184-7/08 $25.00 © 2008 IEEE
DOI 10.1109/CISIM.2008.46

the decision referring to the data format, as well as the choice of mood taxonomy is made.

- *Feature extraction* – This is the first and essential step in emotion extraction. Its aim is to reduce information about song to the description that fully characterize it. Because of many details it is very complex task that needs high accuracy and knowledge from music domain.

- *Classification and solution evaluation* – Data extracted in the previous step create input for classification algorithms. A set of extracted, the size of training set and its quality play an essential role for obtained results. In a large degree it decides about successful results.

As we have mentioned, emotions and mood are very fuzzy concepts. Human being usually is able to recognize emotional state but has difficulties with its proper defining [7]. One of the oldest method of mood description is a list of adjectives expressing a given emotional state, for instance one category can be defined by *dark* and *depressing* another one by *delicate* and *graceful*. The main disadvantage of such approach is ambiguity of concepts. Accurate covering of different emotional states needs huge number of categories that makes a fuzzy boundary between them.

A quite new approach applies dimensional models. The example of this method is circumplex dimensional model. It represents emotion as linear combination of *quality* defining a level of the pleasure feeling and *activation* describing a value of activation. A combination of these two dimensions define a space of emotions. For instance, a bored state can be perceived as a small value of activation and a negative valence. In an opposite direction is an excited state described by a high activation and a positive quality (valence). Thayer's model also uses two independent features - *activation* and *quality* [4]. Typically, visualisation of emotional state is realized by representing it by a point in two dimensional space, where coordinates are *activation* and *quality*. Four clusters are defined in [5], which have been named as follows: *contentment*, *exuberance*, and . Taking into account mood two similar pieces of music will be represented by two points lying closely to each other.

The feature extraction depends on the format of music data. For audio data it starts with signal division into samples [6]. Then subband filters are used in order to acquire detailed information about a part of signal. A further analysis of each part acquires music features: a tone, a rhythm etc. Then, they are merged. Average values and standard deviation are calculated. For the MIDI format a features extraction starts on a higher level of abstraction. In this case data are represented as a sequence of events defining some musical features, for instance as a pitch of sound or its duration. Other features like key, the most frequent intervals or average pitch of sound need further data processing based on music and statistics theories.

The most part of papers which describe different approaches to the emotion discovery in a music focuses on the classification problems. In [4] a mood detection is treated as a multilevel classification problem. This means that a piece of music can belong to more than one class of emotion represented by a set of adjectives. The authors of the paper [5] have proposed a hybrid method joining the Thayer's taxonomy and a hierarchical approach. Practical application of such systems one can see in Internet as a form of search engines or music libraries [1], [2]. As the search criterion they use texts that describe given emotional states. Because of a small set of songs it is difficult to assume that they are full professional systems. A graphical expression of emotions is the subject of research of Hiraga and Matsuda [3], where each note is assigned to a colored rectangle, expressing its emotional charge. Next, after the serialization of notes, a greater area is created that informs about emotion of the piece of music.

The most frequently used supervised classification method is SVM (Support Vector Machine) which is able to classify patterns into two classes only. Unsupervised classification is performed by an application of SOM (Self Organizing Map) mostly. Such system is resistant to ambiguity of emotional concepts and to an inappropriate choice of taxonomy.

## 3. The proposed approach

The aim of our system has been defined as a search for emotionally similar songs on the basis of the musical structure. Because symbolic recording retains it in the highest degree, the system is based on the MIDI format. The next assumption refers to the classification method. The application of unsupervised methods allows to avoid all problems connected with the acquiring an expert knowledge in the taxonomy of emotions and the training set which contains hand-labeled examples by an expert. Therefore we have applied clustering method that searches for groups of similar objects (music pieces). This solution is a human like way of searching a piece of music including a given emotional charge. In this case a listener does not need any knowledge about the name of the music class however information about songs that are close to the pattern is needed.

In some clustering methods there is a need to define the number of clusters. In the complex emotional domain estimation of the number of clusters is practically impossible,because of application of an agglomerative clustering. Its main advantage is a smooth regulation of the classes number.

Many systems of an emotional content extraction treat a piece of music as a self-contained whole but usually in a

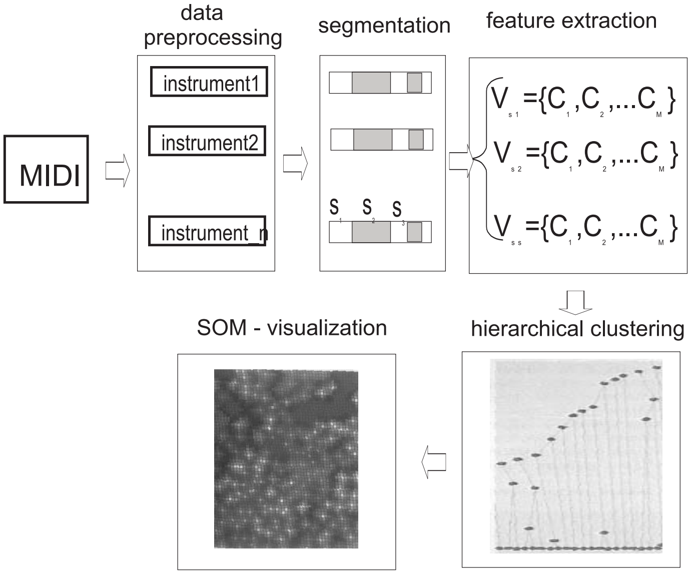


**Figure 1. The structure of the emotional content extraction system**

song there are many parts affecting in different way for human mood. In theory of music there is no rule determining musical structures with a different emotional charge. A mood tracking in a piece of music relies in its division into coherent parts in such a manner that an emotional state can be assigned in unambiguous way to each one.

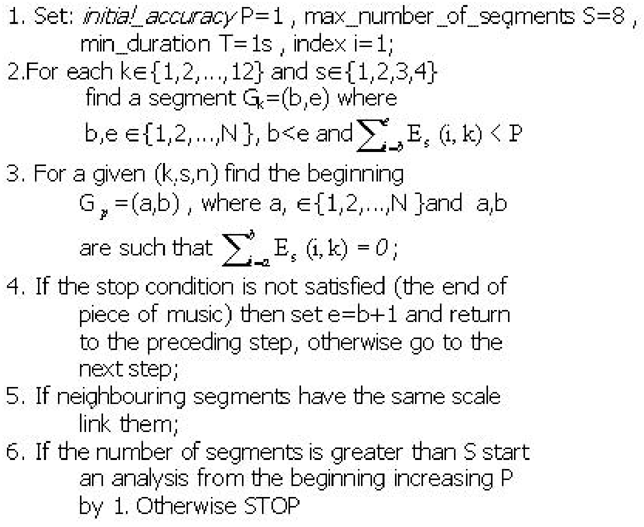
The structure of the system is presented in the Fig.1. The sequences of events from MIDI files are ordered and transformed to the standard form which is modeled in a similar way to a note recording. It is represented by three basic features of sound: *pitch*, *duration* and *velocity*. The set of these ”notes” is assigned to the instruments. In the next step the piece of music is divided into independent segments that potentially can represent different emotional states. Then, feature vectors are extracted, which represent the composition. They are used in the next step to cluster segments and to create a hierarchy of clusters. In the last step clusters are visualised on the SOM network. In the next sections, the details of each step are described.

### 3.1. Data preprocessing

In the MIDI format a sound is represented by start and stop time measured in milliseconds. This means that relative time of note duration is missed in this recording. Melody lines of instruments are collected together in this format. Nevertheless this unordered structure (in the sense of music) makes feature extraction difficult. The aim of preprocessing is to transform a composition to consistent format that contains information about *pitch*, *duration* and *velocity* of sound. In the first phase music events are separated from meta data containing information about title and composer. Then each of them is assigned to the proper instrument. This solution enables an analysis of lines melody separately. The next step relies in creation of a single occurrence of coherent music units that are described by basic features (*pitch*, *duration* and *velocity*). Pitch and duration represent frequency and relative time of duration while velocity define the way of uttering a sound.

### 3.2. Segmentation

The algorithm which assigns segments is based on a music scale which is defined as a sequence of sounds that have specific distances between them. It has a significant influence on a piece of music tone character. For instance, a major scale is perceived as bright and cheerful while minor one – as sorrow and melancholic. In music, beside scale, a key is used. It starts from a concrete sound. Because each octave is composed of 12 different sounds 12 keys are defined (for instance C-dur, a-mol). The applied segmentation algorithm is based on this propriety and estimates parts fitting to a given scale. It treats every change as potentially new segment. The version of the algorithm proposed here is a modified approach used in the paper [8]. The modification refers to the number of scales, error function and accuracy. Our system recognizes four popular scales *s*: diatonic $K_1 = \{0, 2, 4, 5, 7, 9, 11\}$, harmonic minor $K_2 = \{0, 2, 3, 5, 7, 8, 11\}$, melodic minor $K_3 = \{0, 2, 3, 5, 7, 9, 11\}$ and bluesy $K_4 = \{0, 3, 5, 6, 7, 10\}$, where the numbers in brackets define sounds of the scale in question. A diatonic scale exists in 7 different forms: Ionian, Dorian, Phrygian, Lydian, Mixolydian, Aeolian and Locrian. To determine it an estimation algorithm is applied.



```
1. Set: initial_accuracy P=1 , max_number_of_segments S=8 ,
        min_duration T=1s , index i=1;
2.For each $k \in \{1,2,...,12\}$ and $s \in \{1,2,3,4\}$
        find a segment $G_k=(b,e)$ where
        $b,e \in \{1,2,...,N\}$, b<e and $\sum_{i=b}^{e} E_s(i,k) < P$
3. For a given (k,s,n) find the beginning
        $G_p=(a,b)$ , where a, $\in \{1,2,...,N\}$and  a,b
        are such that $\sum_{i=a}^{b} E_s(i,k) = 0$;
4. If the stop condition is not satisfied (the end of
        piece of music) then set e=b+1 and return
        to the preceding step, otherwise go to the
        next step;
5. If neighbouring segments have the same scale
        link them;
6. If the number of segments is greater than S start
        an analysis from the beginning increasing P
        by 1. Otherwise STOP
```

**Figure 2. The pseudocode of segmentation algorithm**

It defines the membership level that a given segment belongs to the given scale on the basis of three sounds: the first, the fifth and the third ones of a scale (tonic, dominant, mediant). The choice is determined by two conditions:

- the existence of all levels of a scale in the segment (tonic, dominant or mediant),
- the number of occurrences of the tonic and the dominant should be maximal

The segment which does not satisfy the first condition for any forms is judged as undefined one.

An error function $E(s, n, k)$ which defines a membership value expressing that a sound $n$ belongs to a given scale $s$ starting from sound $k$ has the following form:

$$E(s,n,k) = \begin{cases} 0 & \text{if } mod(p(n)-k) \in K_s, \\ 1 & \text{otherwise,} \end{cases} \quad (1)$$

where $s \in \{1, 2, 3, 4\}$ assigns a music scale, $n \in \{1, 2, ..., N\}$ is a note from an input sequence, $k \in \{1, 2, ..., 12\}$ is the first sound in a scale (tonic). A value of function $E$ is equal to 0 if the sound belongs to the scale with tonic that lies in $K$ and equals to 1 otherwise.

The segmentation algorithm in pseudocode is shown in Fig 2. According to the given scale the algorithm is searching for the longest segments, that altogether cover a piece of music. If their number is greater than assumed value of accuracy it is decreased. This means an existence of accidental notes inconsistent with a scale. A smaller accuracy allows to treat a segment as one consistent unit.

### 3.3. Feature extraction

Feature extraction phase relies in extraction of essential information that groups similar emotional parts and discriminates those having nothing in common.

In our system feature vector includes:
*music scale* – 11 scales are taken into account and they are ranked. It causes that scales with a similar emotional charge are close to each other.
*accuracy* – a value of the scale estimation error, a less value of accuracy allows to ignore some sounds in order to assign a scale.
*sound intensity* – it is defined here as the total number of sound occurrences in 1 sec. It informs about a tempo of a given piece of music.
*basic sound* – an average pitch of sounds in a segment.
*interval* – an average distance between sounds in a segment. A high value describes a piece of an energetic music while low value refers to the calm music.
*direction* – it says about tendency of sound arrangement in the piece of music (growth or fall).
*velocity* – the average velocity of sounds in a segment.
*duration of notes* – the average duration of note in a segment.
Each feature is normalized to the range [0,1].

### 3.4. Hierarchical clustering

Hierarchical clustering is an unsupervised method of learning used to gather similar objects together in groups (clusters). In the system bottom-up approach is applied. This means that in one step two clusters are linked together if they have similar emotional content. This procedure ends when there is only one cluster containing all pieces of music from data base. Moving from the root of this tree (dendrogram) to the leafs we can see all procedure phases of the arose hierarchy. In this method two elements are essential – the *distance* between vectors (clustered objects) and the way of *linkage* of clusters. The distance defines similarity between objects (pieces of music). The Euclidian distance is applied in our system. The linkage technique says which two clusters ($A$ and $B$) may be linked in a given step. The following methods of cluster linkage are tested in the experiments:
*single linkage*

$$D(A,B) = min(d(i,j)), where \quad i \in A, \quad j \in B, \quad (2)$$

| accuracy value | 1 | 3 | 5 | 7 | 11 | 15 | 25 | 33 | 45 | 47 | 101 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| class label | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| number of songs | 83 | 4 | 7 | 3 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| group label | 1 | 2 |  | 3 |  |  | 4 |  |  |  |  |
| number of songs | 83 | 11 |  | 6 |  |  | 5 |  |  |  |  |

**Table 1. The results of a song segmentation**

*complete linkage*

$$D(A,B) = max(d(i,j)), \quad where \quad i \in A, \quad j \in B, \tag{3}$$

*average linkage*

$$D(A,B) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} d(i,j)}{M * N}, where \quad i \in A, \quad j \in B. \tag{4}$$

In equation eq.4 M and N are the number of elements in clusters A and B,respectively. The average linkage takes the mean distance $d(i, j)$ between all $(M * N)$ possible pairs of entities of the two clusters in question ($A$ and $B$). It is therefore more computationally expensive than the aforementioned methods.

### 3.5. Visualisation

The aim of visualisation methods is presentation the results in a form enabling their analysis and evaluation. Two forms are considered in the system. The levels of hierarchy are visualised with the aid of a dendrogram. This presentation of clustering results enables to find the level where the number of cluster is satisfying. The second form is *SOM* (Self Organizing Map) neural network. It consists of two dimensional table of neurons. SOM projects high dimensional features vectors (input neurons) onto this two dimensional table of neurons. The map seeks to preserve the topological properties of the input space vectors. This means that SOM plots the similarities of the data by grouping similar data items together. SOM is a competitive network. The way that SOM goes about organizing themselves is by competing for representation of the samples of input data. Neurons are also allowed to change themselves by learning to become more like samples in hopes of winning the next competition. The learning process makes the weights organized themselves into a map that represents similarities of input vectors (in our case feature vectors that express emotions in a piece of music).

## 4. Experimental study

The aim of the experiments described in this section was to evaluate the efficiency of the component methods applied in each step in the system (Fig. 1)and efficiency the system as a whole. The tests were performed with the usage of 104 pieces of music recorded in MIDI format. In order to obtain credible results songs representing various kinds of music, performed by different authors, were used. The following kinds of music were collected: *Pop*, *Rock*, *Dance*, *Blues*, *Jazz*, *Classic*, *Hip Hop* and *Techno*.

The segmentation method was tested first. Its aim was to find parts of songs that potentially contain the same emotional charge. Some songs have very complex music structure that manifests itself in a high frequency of accidental notes. This is conductive to many but short parts that are difficult to classify. Therefore the maximal number of segments *N* has been assigned and set to 8. The minimal duration *T* of one segment was set to 1second. The results of segmentation are shown in Table 1. As it can be seen 104 songs are split up to 11 classes of accuracy with values from 1-101, where 1 means the highest precision. It is worth to notice that more than 80% of songs were clustered faultless (they do not have accidental notes). Depending on the conformity with a music scale four groups were distinguish. The number of songs in each group is shown in Table 1, as well.

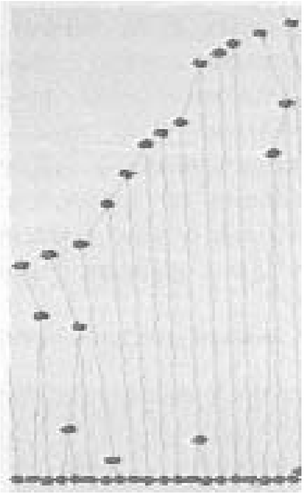


**Figure 3. Dendrogram visualises the aggregating process of clusters**

| $N$ | 5 | 12 | 20 |
|---|---|---|---|
| simple linkage $L_S$ | 138 | 84 | 63 |
| average linkage $L_A$ | 115 | 49 | 24 |
| complete linkage $L_C$ | 58 | 23 | 14 |

**Table 2. The standard deviation – $S_N$ for three methods of clustering linkage, *N* – a number of clusters**

Next experiments refer to the clustering method. Hierarchical way of data organization can be characterized by possibility of the clusters number control. The goal was to compare the clusters linkage methods as regards of evenly splitting music segments. Experiments were made for the number of clusters $N \in \{5, 12, 20\}$. The results in term of standard deviation ($S_N$) are shown in Table 2. Clusters with the most homogenous structure are obtained by complete linkage (the smaller $S_N$ the more homogenous clusters). It is the result of the fact that a distance between clusters are assigned by two the most remote elements from both groups. This gives more significance for elements that are far away from centrum and simultaneously it causes that linked elements have similar features. The simple linkage that joins two clusters containing two the closest elements gives the worst results. This linkage is conducive to chaining creation, i.e. one cluster. This can be easily observed in dendrogram in Fig. 3 with clearly seen slanting line suggesting an addition of new elements in incremental way. The Table 3 shows the number of segments for various methods of cluster linkage for $N = 5$.

The hierarchical clustering method groups and discovers relationships between objects. Human has to evaluate the quality of obtained solution. Therefore in the next experiment the consistency of automatic clustering with human evaluation is compared. It is measured as formation of consistent clusters containing segments with the same emotional charge and distinguishing elements with another characteristics. In practice it is limited to evaluation of elements scattering in clusters. Other metrics, for instance purity, need expert knowledge referring to the presumable placement of songs. System does not posses any knowledge about the shape of emotional space therefore application the last metrics was not possible.

| $C$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $L_S$ | 315 | 8 | 15 | 2 | 1 |
| $L_A$ | 15 | 273 | 41 | 8 | 4 |
| $L_C$ | 106 | 101 | 124 | 8 | 2 |

**Table 3. The number of segments in clusters for N=5 for three methods of clustering linkage, *C* – cluster index, $L_S$, $L_A$, $L_C$ – simple, average, complete linkage, respectively**

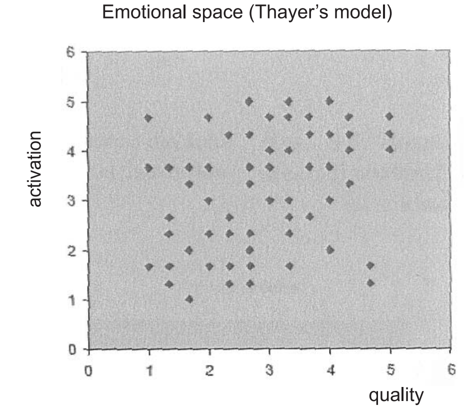


**Figure 4. Emotional space of segments on the basis of users' evaluation**

The applied test procedure is composed of the following steps:

1. the choice of an emotional taxonomy and a random set of songs from data basis;
2. evaluation of a samples by human beings according to the assumed taxonomy;
3. an automatic assignment of emotional state to the samples (the results of our system performance);
4. the comparison of the results i.e evaluation of clusters obtained automatically from the system and those advisable by human beings.

The system does not use taxonomy of emotion. It defines areas by searching similar songs in an emotional charge. A human in opposite to machine has to know a point of reference i.e. a semantics. To realize this the Thayer's model has

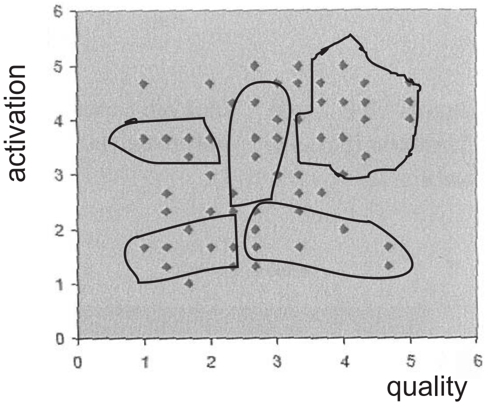


**Figure 5. Averaged users' evaluation and placement of clusters for N=20**

been borrowed. As we have mentioned it expresses emotion by two components *activation* and *quality*. Both coordinates are in the range [0,6].The lowest value equal to 1 assigns minimal quality or minimal activation. The highest value is equal to 5.

70 segments were chosen randomly from the set which was obtained after segmentation. These segments were evaluated by independent three users taking into account their *quality* and *activation*. These values were averaged and the results are shown in Fig.4.

As it can be noticed a piece of music are evenly spread over the space, but more accurate analysis gives us information that quality and activation grow up in the same way. This suggests that emotions with high *quality* also have high *activation*.

The same segments were automatically clustered. The results for the number of clusters $N \in \{5, 12, 20\}$ were plotted on the diagram shown in Fig. 5 and analysed. For N=5 songs were grouped in small groups. The obtained results can be evaluated as satisfying locally only. Some areas alternate with each others and they do not create consistent parts. The best results were obtained for $N = 20$. They are illustrated in Fig. 5, where only 5 the most numerous clusters are visualised. Analysing the segments localization in the emotional space we can see a tendency to form clusters placed on different parts in emotional space. Most of them create consistent integral parts. As it can be seen the whole space is covered by the songs in emotional space.

The same segments of music were also processed by SOM network which consists of *50x60* neurons. Originally, the clusters have been visualised on the colour map. In Fig. 6 the results are shown in black and white only. Therefore clusters originally assigned by the same colours have the same labels in this figure. The clusters create compact areas. This result confirms the hypothesis that the extracted features contain information about relationships responsible for the moulding emotional charge in music.

## 5. Conclusion

The results of experiments prove that a musical structure has features that effect on the human emotions. The obtained clusters gather pieces of music with similar emotional charge, their span depends on the tree depth (dendrogram depth) assumed as a current level of details. High influence on a system evaluation has the number of people taking part in the experiments (users evaluating emotions in music), so we plan to increase their number in the nearest future. The experiments have shown that some parts of music were perceived in various way by human beings. In the future in order to obtain consistency of the system the model of user that would personalise his reception will be created.

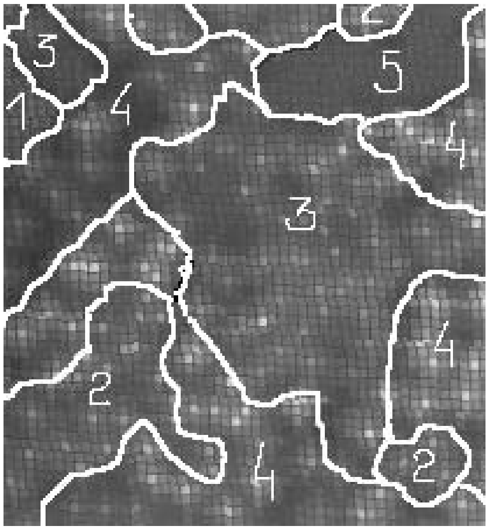


**Figure 6. Visualisation of song segments on SOM neural network**

## References

[1] Audio network: Atmospheric music and mood music from the audionetwork library. Internet: http://www.audiolicense.net/t3_atmosphere.asp, 28.03., 2008.

[2] The experience project: Mood and meaning music search: Music for life. Internet: http://www.experienceproject.com/music_search.php, 28.03., 2008.

[3] N. Hiraga, R. Matsuda. Graphical expression of the mood music. In *IEEE International Conference on Multimedia and EXpo*, volume 3, pages 2035 – 2038, 2004.

[4] M. Li, T. Ogihara. Content -based music similarity search and emotion detection. Technical report, University of Rochester, 2003.

[5] D. Lu, L. Liu and H. Zhang. Automatic mood detection from acoustic data. In *ISMIR2003, Fourth International Conference on Music Information Retrieval*, 2003.

[6] D. Lu, L. Liu and H.-J. Zhang. Automatic mood detection and tracting of music audio signals. *IEEE Transaction on Audio, Speech and Language processing*, 14(1):5–18, 2006.

[7] J. Posner, J. Russel and B. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development and psychopatology. *Development and Psychopathology*, pages 715–734, 2005.

[8] M. Zhu, Y. Kankanhalli. Key-based melody segmentation for popular songs. In *17th International Conference on Pattern Recognition (ICPR'04)*, volume 3, pages 862–865, 2004.