ARCHIVES OF ACOUSTICS
Vol. **40**, No. 4, pp. 513–525 (2015)

DOI: 10.1515/aoa-2015-0051

Music Mood Visualization Using Self-Organizing Maps

Magdalena PLEWA, Bożena KOSTEK

Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Narutowicza 11/12, 81-233 Gdańsk, Poland; e-mail: mplewa@sound.eti.pg.gda.pl, bokostek@audioacoustics.org

(*received August 19, 2015; accepted September 30, 2015*)

Due to an increasing amount of music being made available in digital form in the Internet, an automatic organization of music is sought. The paper presents an approach to graphical representation of mood of songs based on Self-Organizing Maps. Parameters describing mood of music are proposed and calculated and then analyzed employing correlation with mood dimensions based on the Multidimensional Scaling. A map is created in which music excerpts with similar mood are organized next to each other on the two-dimensional display.

Keywords: music mood, music parameterization, MER (Music Emotion Recognition), MIR (Music Information Retrieval), Multidimensional Scaling (MDS), Principal Component Analysis (PCA), Self-Organizing Maps (SOM), ANN (Artificial Neural Networks).

1. Introduction

The motivation behind the presented study is that the increasing amount of music available online causes a growing need of new ways of organizing and searching through music libraries. Mood of music is one of the important cues that are used for music exploration. Self-Organizing Maps (SOMs), a tool that is strongly related to human perception, were used to map music pieces into model of emotions. Authors of the presented paper focused on finding audio parameters related to mood of music and implemented SOMs to map music excerpts described by a large feature vector into the 2-dimensional representation of emotions.

Musicologists indicate a few elements of a musical piece – melody, rhythm, agogics, articulation, and dynamics – that are important in analysis and they form foundations of music. Moreover, it may be said that melody together with rhythm carry 90% of musical information. Rhythm is also an element of a piece determining musical style, which may be valuable in Music Information Retrieval (MIR). The rhythmic structure together with melody patterns retrieved from audio signal carry information about the genre of a musical piece, thus both are highly correlated. Moreover, music can be defined in terms of descriptive characteristics such as aesthetic experience, preference, mood or emotions. Huron (2000) assumes that the four most useful characteristics items are: style, emotion, genre, and similarity. However, some music analysts argue that style and genre are to some extent interchangeable expressions. It is also said that a long list of genres is a result of artists' interest to introduce new genres. Moreover, classifications are often arbitrary and encompass sub-genres that belong to different styles or genres. One of the features, which can be useful and intuitive for music listeners, is "mood" (Casey *et al.*, 2008). Even if it seems to be the easiest way to describe music for people who are non-experts, it is very difficult to find an exact correlation between physical features and perceived impression.

It should be noted that music mood recognition constitutes an important part of MIR. Automatic genre/style and similarity classification within MIR have been exploited quite thoroughly in recent years. This is visible not only in the literature sources but also in music services and applications (ISMIR conferences; Cook, 2002; Kostek, Czyzewski, 2001; Kostek, 2013; 2014; Kostek, Kaczmarek, 2013; Mostafavi *et al.*, 2013; Papaodysseus *et al.*, 2001; Rauber *et al.*, 2002a; Raś, Wieczorkowska, 2010; Rumsey, 2011; 2014; Tzanetakis, Cook, 2002; Wieczorkowska *et al.*, 2011; AMAZON, ITUNES, PANDORA). Recently, music mood recognition becomes a thorough subject of research studies and analyses (Cruz *et al.*, 2007; Drossos *et al.*, 2015; Kim *et al.*, 2010; Laurier *et al.*, 2009; Markov, Matsu, 2014; Lu *et al.*, 2006; Panda, Paiva, 2011; Zentner *et al.*, 2008). This area of research studies is called Music Emotion Recognition (MER) and aims at recognizing emotions contained in audio signals (Panda, Paiva, 2011; Plewa, Kostek, 2013; Rauber, Frühwirth, 2001; Rauber *et al.*, 2002a; 2002b; MUSICOVERY). There are also approaches that involve advanced computational methods, e.g. regression approach, Support Vector Machines (SVM) or fuzzy logic (Lima *et al.*, 2012; Trochidis *et al.*, 2011; Zentner *et al.*, 2008) which are used for automatic mood assigning. Results achieved in the mentioned research projects do not exhaust the subject. First of all, an improvement in automatic efficacy is sought (Lima *et al.*, 2012; Zentner *et al.*, 2008) especially as the outcomes of the automatic mood recognition are usually only slightly better than 60–70%. Moreover, subjective studies, which concentrate on assigning appropriate labels corresponding to music features are also needed to find relationship between these descriptors and features derived objectively. The interest towards this particular direction is motivated by music networking services in which users tend to listen to music pieces that reflect their emotions.

Music Mood Recognition is based on the basic definitions of perception. Lewis defined qualia: "There are recognizable qualitative characters of the given, which may be repeated in different experiences, and are thus a sort of universals; I call these qualia" (Lewis, 1929). The discovery of the relationship between the measurable content of the physical world and human perception seems to be a fundamental problem in Music Mood Recognition (Kostek, 2011). One of the conclusions of the mood description research was that tempo seems to be an essential feature determining mood of music (Cruz *et al.*, 2007), so this specific characteristic should be taken into consideration. Since it is possible to extract rhythm and tempo from music (Hevner, 1936) (or it is known if one has an access to the MIDI notation), it is important to discover and verify how significant the correlation between tempo and specific mood descriptors is.

It is to remember that both subjective descriptors and features describing mood are multidimensional, thus another important issue is presentation of these dimensions. Thus one of the aims of this paper is to automatically obtain a graphical representation of mood of music. In this study we have used the MDS (Multidimensional Scaling) analysis-based graphical representation obtained in the previous phase of experiments (Plewa, Kostek, 2013). In some related work one can see that the basis for creating similarity between songs is often acquired from listening tests. Novello *et al.* proposed a web-based listening experiment that assesses the perception of inter-song similarity optimizing stimulus coverage and time of experiments. The experiment used 78 song excerpts selected from 13 genres and involved 78 participants. To discover the background of the participants' perceptual space they used Multidimensional Scaling Analysis (MDS) and quadratic discriminant analysis to search for axes that maximize the separation of the excerpt classes (Novello *et al.*, 2011). However, collecting similarity data from listeners is time consuming, and the MDS analysis – even though often applied to analyze similarity – cannot be used as the main similarity representation (Trochidis *et al.*, 2011). However, in the study of Trochidis and his collaborators (2011) it was shown that the emotion processing mechanism is quite similar for musicians and non-musicians resulting in the same low-level spectral and temporal features correlated with arousal and high level contextual features correlated with valence dimension.

The MDS technique was also used to unravel hidden relationships between the musical styles. The MDS was calculated in this case based on two alternative metrics, namely: the average mutual information and the fractal dimension. The results reveal significant differences in musical styles, demonstrating the feasibility of the proposed strategy and motivating further development towards a dynamical analysis of musical sounds (Lima *et al.*, 2012). However, such methods need well-defined data. Another way to use MDS in music similarity analysis is item-to-item collaborative filtering. This is based on the notion that people who listen to song A, will also listen to song B. Music space may also be evaluated by social tagging. However, in such a case an assignment of tags should be controlled in some way, otherwise the tags may differ for a particular song. To identify relevant tags one may use Principal Component Analysis (PCA), especially when aiming at reducing the dimensionality. However, an apparent way to transform automatically a high-dimensional space to a few-dimensional (even two-dimensions) representation is using Self-Organizing Map (SOM). That is why another goal of the presented work is attempt to organize music dataset employing SOMs.

In this paper, first the previous stage of research consisting in listening tests focused on assigning mood to song excerpts is recalled. Then it is shown which objectively derived parameters are behind the mood model obtained in listening tests. Correlation between MDS dimensions and mood parameters is calculated and therefore a set of features strongly related to mood is created. The main aim of this paper is however to employ PCA technique to reduce the dimensionality of music feature space and then to use it as an input to SOM for visualizing a map of songs. Examples of visualization of the obtained results are shown in Sec. 5. Section 6 briefly recapitulates the results and gives some concluding remarks.

2. Mood of music

2.1. Music mood recognition

MER research studies have not determined any main or right model of the music mood. Contrarily, a great variety of mood models are constantly being explored and devised in psychological and musicology studies. Numerous studies on mood classification reveal different conceptual findings, however they led to the conclusion that mood description can be assigned to one of the following two approaches: dimensional or cluster description.

The dimensional approach focuses on mood identification based on positioning it in the space of several mood-dimensions. Particular dimensions represented by axes are named correspondingly to human perception of mood or emotions. Thayer created a two-dimensional model Valence/Arousal (Thayer, 1989). Axes divide the plane into quarters, which correspond to the following moods: **contentment** (low arousal, high valence), **depression** (low arousal, low valence), **anxious/frantic** (high arousal, low valence) and **exuberance** (high arousal, high valence). The authors of this paper refer to this model in the study presented here.

The categorical approach bases on the clusters of mutually exclusive categories (Mirex, 2009; Hevner, 1936). For example Mirex is a set of adjectives organized in five clusters (categories), Hevner lists 67 adjectives grouped into eight mood clusters, in Schubert's model (Schubert, 2003) 46 affective adjectives are combined into nine clusters according to their position on the two-dimensional Thayer's model. Hierarchical clusters allowing grouping individual songs into sub-clusters or clusters into super-clusters are also known.

Finding the right mood representation is the first step towards the automation of mood recognition. To describe a music piece with subjective tags related to mood description, an expert needs about 20–30 minutes of work (Casey *et al.*, 2008). For larger databases it may result in an enormous workload. That is one of general reasons for creating systems enabling to automatically assign mood to music.

2.2. Multidimensional scaling experiment

The MDS technique was used to unravel hidden relationships between mood of different pieces of music. Multidimensional scaling (MDS) was originated in the area of psychology. It takes a pairwise set of distances and forms an m-dimensional map. The key problem raised in this field is to recognize "underlying dimensions" that would explain similarities or dissimilarities observed by subjects. In the study of Borg and Groenen (2007) the authors stated that MDS application in psychology is often based on direct similarity judgments by the subjects. Noteworthy is that similarity may concern diverse subjects. MDS applied to psychological data enables discovering dimensions that would in a meaningful way explain rules of the perception. This method requires data, which contain direct similarity judgments by respondents. Reconstruction of distances between objects by placing objects in the multidimensional configuration is essential for the MDS concept. One may simply assume Euclidean metric for this purpose. The Euclidean distance d_{ij} between points i and j in m-dimensional space $\mathbf{X}$ is calculated from:

$$d_{ij} = \sqrt{\sum_{a=1}^{m} (x_{ia} - x_{ja})^2}, \tag{1}$$

where a is the number of the element (from 1 to m).

MDS enables to find function f, which maps the proximities p_{ij} – similarities obtained from subjects- into corresponding distances d_{ij} in the MDS space $\mathbf{X}$:

$$f(p_{ij}) = d_{ij}(\mathbf{X}). \tag{2}$$

Since (depending on the data) the exact representation does not always exist, there is a need to define a value, which reflects the goodness of representations or the desired precision (error) of the map. To build an optimum representation, the Multidimensional Scaling algorithm minimizes a criterion called *Stress* or Kruskal's normalized *Stress-1* criterion is observed:

$$\textit{Stress-1} = \frac{\sum\limits_{(i,j)} [f(p_{ij}) - d_{ij}(\mathbf{X})]}{\sum d_{ij}^2(\mathbf{X})}. \tag{3}$$

The closer the stress is to zero, the better the representation.

Multidimensional Scaling was used in various applications related to sound and music. Trochidis *et al.* (2011) applied MDS to analyze similarity within the wide area of western music. Wagenaars *et al.* used Multidimensional Scaling test to determine optimum values of compression in music (Wagenaars *et al.*, 1986), while Małecki (2013) found three dimensions that determine perception of similarity between acoustics of different rooms. MDS was also applied to emotional responses to music (Bigand *et al.*, 2005). Bigand *et al.* (2005) stated that the 3-dimensional space is needed to provide a good representation of emotions, with arousal and emotional valence as the primary dimensions. There are quite a few differences between Bigand's *et al.* research and results presented in this work therefore conclusions derived from both studies may be different.

In the previous study of the authors the MDS experiment was conducted to collect the similarity data

Table 1. List of the music tracks used in the MDS experiment (marked in grey). All of the 15 songs are mapped using SOM representation.

No	Genre	Artist	Album	Title
1	Jazz	Kenny G	Paradise	Malibu Dreams
2	R&B	Central Line	The Funk Essentials 1222 Collection And More	Walking Into Sunshine
3	Pop	The Clash	Combat Rock	Should I Stay Or Should I Go
4	Pop	Tom Jones	Reloaded: Greatest Hits	Kiss
5	Alternative Rock	Pearl Jam	Ten (Legacy Edition)	Black (Remasered 2008)
6	Pop	Sting	Fields Of Gold – The Best Of Sting 1984–1994	Fields Of Gold
7	Rock	Aerosmith	Big Ones	Rag Doll
8	Classical	Sir Landon Ronald	The Elgar Edition: The Complete Electrical Recordings of Sir Edward Elgar	Coronation March Op: 65 (1993 Digital Remaster)
9	Alternative Rock	Hey Champ	Star	Cold Dust Girl
10	Pop	Jennifer Lopez	Love (Deluxe Version)	Charge Me Up
11	Pop	Erykah Badu	Live	Tyrone (Extended Version)
12	Rock	Faith No More	This Is It: The Best of Faith No More	Epic
13	Alternative Rock	Green Day	21 Guns EP	21 Guns (Album Version)
14	Jazz	Eliane Elias	Light My Fire	My Cherie Amour
15	Hard Rock & Metal	Slayer	Seasons In The Abyss	War Ensemble

for the MDS analysis (Plewa, Kostek, 2013). The set of the music tracks used in the experiment is listed in Table 1. Songs for MDS were chosen from the set that was previously evaluated by listeners in the listening test. The MDS set was constructed to contain songs with very similar and very different mood of music and is marked in grey in Table 1.

Similarity data obtained from the MDS experiment were averaged. The MDS (1D) data representation was constructed in MATLAB using Kruskal's normalized *Stress-1* criterion. *Stress-1* factor reached 0.01. All of the 15 songs from Table 1 are mapped onto SOM visualization map in further analysis. The obtained MDS map is presented in Fig. 1.

Dimensions achieved from the MDS analysis correspond to DIMENSIONS with corresponding labels "Calm" and "Joyful". This can lead to the conclusion that Thayer's model is appropriate for describing mood of music. One of the axes (Dimension 1) can be interpreted as Valence ("Joyful", positive or negative content) and the Dimension 2 as Arousal ("Calm", energetic content). These results are coherent with studies that applied Thayer's model to mood in music (Kim *et al.* 2008; 2010; Lu *et al.*, 2006; Plewa, Kostek, 2013; Schmidt, Kim, 2010; Thayer, 1989).

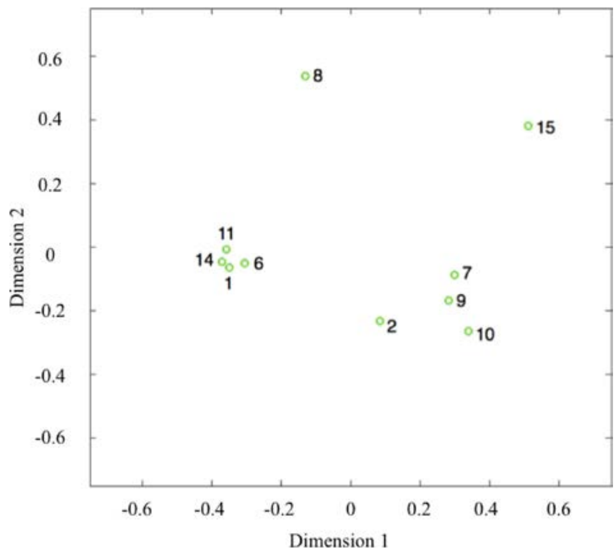

Fig. 1. MDS representation based on direct similarity judgments.

3. Self-organizing maps

Self-Organizing Map is an unsupervised neural network that can be used to organize objects with unclear relations between each other into map-type representation. The topological relations between objects are preserved as detailed as possible. These properties have led to implementation of SOMs in experiments related to mood of music. SOMs are strongly related to human perception, which is adequate for music cognition. Specifics of the task match assumptions and approach of the method are briefly described in the following Section.

3.1. Basics of Self-Organized Maps

Mathematically, SOM (Self-Organizing Map) is defined as an unsupervised neural network providing a mapping from a high-dimensional space to a few-dimensional (in most of cases two-dimensional $\mathbf{K}\times\mathbf{L}$) representation (Rojas, 1996; Ultsch, 2003). SOM consists of the 2-dimensional grid of neurons, with a weight vector related to each unit. SOM is forced by vector x_1, $x_2,\ldots$, and activation $y_1, y_2, \ldots, y_m$ for each neuron unit for the presented object is calculated (Fig. 2). This type of networks was introduced by Kohonen (1982; 1984; Kohonen, Honkela, 2007).

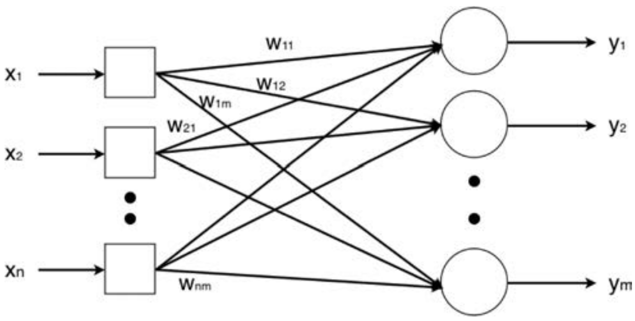

Fig. 2. Schema of the SOM network.

The Euclidean distance d between weight vector of the unit and input is commonly used as the activation function. The weight vector of the unit that achieved the highest activation is selected as a "winner" and is recalculated to resemble as close as possible the presented input vector. Moreover, the weight vectors of units in the neighborhood of the winner are modified accordingly, but not as strong as the "winner" (MATLAB; Rauber, Frühwirth, 2001; Rauber *et al.*, 2002a). The winning neuron a is selected from the $\mathbf{K}\times\mathbf{L}$ network consisting of i elements, according to the following relation:

$$d\left(x, w^{(a)}\right) = \min_{1\le m\le K\times L} d\left(x, w^{(m)}\right), \qquad (4)$$

where d is a measure of distance between n-dimensional vector $\mathbf{x}$ and the weights vector $\mathbf{w}$ of the output vector in $\mathbf{K}\times\mathbf{L}$ space, $w^{(m)}$ is a weight corresponding to neuron with index m. This rule is called Winner Takes All (WTA) and refers to hard competition, where only unit with the highest activation is trained. SOM is forced by n-dimensional signal $x^{(j)}$, where j is iteration in the learning process (index of the element in the learning sequence). The winning unit a, where a indicates the index of the neuron, is updated according to the rule:

$$w_i^{(a)(j+1)} = w_i^{(a)(j)} + \eta^{(j)}\left[\left(x_i^{(j)} - w_i^{(a)(j)}\right)\right] \qquad (5)$$

i indicates the index of the element (of the n-dimensional input vector), $\eta^{(j)}$ is speed of learning in j-th step and belongs to [0,1], $w^{(a)(j)}$ is weight of neuron a in j-th step of the learning.

Winner Takes Most (WTM) concept implies solution that the weight vectors of units which number m belong to the neighborhood N_a of the winner a are modified accordingly, but not as strong as the "winner" (MATLAB; Rauber, Frühwirth, 2001; Rauber, 2002a; Tadeusiewicz, 1993).

$$\forall_{m\in N_a} \; w_i^{(m)(j+1)} = w_i^{(m)(j)} + \eta^{(j)} h\left[x^{(j)} - w_i^{(m)(j)}\right], \qquad (6)$$

where N_a is a set of units adjacent of the winning unit a, and h is the neighborhood function of neuron a. Neighborhood function can vary from simple to complex functions. Both speed of learning η and neighborhood function h are changing during learning process and monotonically decrease to avoid compensation at the final stage of the learning process.

The topology of the output layer and can be arranged on i.e. a rectangular, hexagonal or random lattice (Fig. 3). That determines the number of connections of a single neuron. Useful extensions include using toroid grids where opposite edges are connected.

In brief, SOM training may be described according to two main rules:

- Competitive learning: the prototype vector most similar to the data vector is modified so that it is even more similar to it. This way the map learns the position of the data cloud.
- Cooperative learning: not only the most similar prototype vector, but also its neighbors on the map are moved towards the data vector.

The learning process can vary, depending on the architecture of the network, default weight values and a training set. The topology of the output layer can be arranged on i.e. a rectangular, hexagonal or random lattice. That determines the number of connections of a single neuron. Interpretation of the SOM results cannot be assumed *a priori*. Meaning of the particular areas of the map can be specified after the analysis of individual cases.

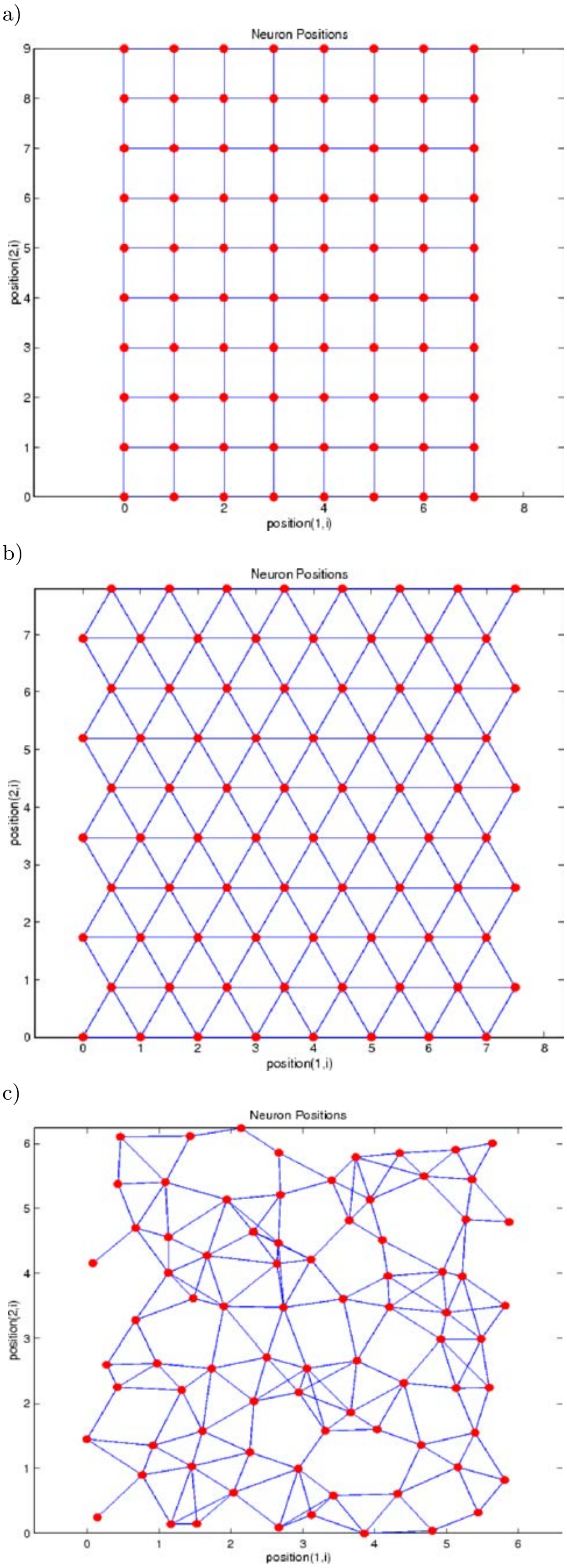

Fig. 3. Examples of Self-organizing Map topologies: a) rectangular, b) hexagonal, c) random. Circles represent neurons and blue lines represent connections between units.

3.2. Self-Organized Maps applied to music

SOMs are strongly related to human perception and are dedicated to complicated tasks, where rules may not be clear, i.e. cluster analysis, creation of models and mapping features. This approach seems to be also natural for music cognition. Thus Self-Organizing Maps are used to organize library systems as well as music libraries (Pampalk, 2001; Pampalk *et al.*, 2002; Rauber, Frühwirth, 2001), also while taking into account music genre (Rauber, 2002a). Furthermore, SOMs help to understand various problems related to sound and music perception. The aim of the work of Palomäki's and his collaborators (Palomäki *et al.*, 1999) was to simulate human perception of spatial sound. He applied self-organizing maps to the evaluation of spatial discrimination of real and virtual sound sources. SOM was trained with localization cues computed using a binaural model. Tuzman (2001) created a system for reduction of impulsive noises based on SOM. Barbedo *et al.* (2005) proposed a Cognitive Model for Objective Assessment of Audio Quality. Their system maps previously extracted parameters into an estimate of the subjective quality. Very common application of SOMs in Music Information Retrieval is to create a 2-dimensional representation either of music set, music database or particular samples.

Rauber and Frühwirth (2001) from Vienna University of Technology proposed a SOM-enhanced JukeBox (SOMeJB) system (Frühwirth, 2001; Frühwirth, Rauber, 2001) to organize their music database analogically to the text library. The classification is mostly content- and genre-based. A system that automatically organizes any music collection according to music similarity was presented by Rauber *et al.* (2002a). The system introduced consisted of 2-dimensional SOM representation that could be generated for any music set. More complex variation involved Growing Hierarchical Self-Organizing Maps (GHSOM) with a 3-layer architecture (Rauber *et al.*, 2002b). GHSOM was fed with 1200 psychoacoustic loudness and rhythm descriptors. It is worth notice that the organization does not follow clean "conceptual" genre styles but rather reflects the overall sound similarity.

All of the various examples listed above show that Self-Organizing maps can be a very useful tool within the area of Music Information Retrieval, thus the attempts of SOM implementation to MER are well founded.

4. Feature extraction

Parameters extracted from audio signal play a role of the "objective" descriptors of music. The task of the researcher is to determine relation between musical features or characteristics and particular sets of parameters. Within the area of music cognition, the choice of the parameters can be based on the analysis that explores the connection between listening tests and values of the parameters. Using too many parameters not related to particular characteristic, i.e. mood of music, can cause the situation where relevant parameters are covered under "the noise" of unrelated ones. Therefore a selection of parameters to be included in the feature vector was performed.

4.1. Input set of parameters

The starting point of the feature vector (FV) content creation for the purpose of automatic mood recognition was examination of previous studies performed in MIR by the authors and their collaborators. Resulted from them was FV applied to two databases, namely ISMIS (Kostek *et al.* 2011) and SYNAT (Hoffmann, Kostek, 2014; Kostek, Kaczmarek, 2013; Kostek *et al.*, 2013; Plewa, Kostek, 2013, Rosner *et al.*, 2014), thus its content may be treated as very thoroughly analyzed. Moreover, the same FV was used in the ISMIS'2011 conference in music competition (Kostek *et al.* 2011), in which more than 100 teams participated, thus it may also be treated as a kind of benchmarking. ISMIS is a database of approx. 1300 music excerpts of high quality audio excerpts, collected and divided into six music genres. On the other hand, the SYNAT database is a collection of 52532 pieces of music described with a set of descriptors obtained through the analysis of mp3-quality recordings. For the SYNAT database, the analysis band is limited to 8kHz. The database stores 173-feature vectors, which in majority are the MPEG-7 standard parameters (109). The vector has additionally been supplemented with 20 Mel-Frequency Cepstral Coefficients (MFCC), 20 MFCC variances and 24 time-related 'dedicated' parameters. The SYNAT database was realized by the Gdansk University of Technology (GUT) (Hoffmann, Kostek, 2014) and music was collected from the Internet. The vector includes parameters associated with the MPEG-7 standard, melcepstral (MFCC) parameters and is supplemented by the so-called dedicated parameters which refer to temporal characteristic of the analyzed music excerpt. Full list of parameters was shown in the earlier study (Kostek *et al.*, 2013). Since MPEG-7 features and MFCC are commonly adopted in rich literature on this subject, thus they are not presented here in detail. The list of parameters includes: Spectral Flatness Measure (SFM), Spread Spectrum Audio, Audio Spectrum Envelope (ASE), Spectral Centroid, Temporal Centroid, Root Mean Square (RMS)-related parameters, etc.; their names along with abbreviations are included in Table 2.

Table 2. The list of parameters within the SYNAT music database.

No.	Parameter	Abbreviation
1	Temporal Centroid	TC
2	Spectral Centroid	SC
3	Spectral Centroid variance	SCV
4–32	Audio Spectrum Envelope for particular bands	ASE (1–29)
33	ASE average for all bands	ASE_M
34–62	ASE variance values for particular bands	ASE_V (1–29)
63	averaged ASE variance	ASE_MV
64	average Audio Spectrum Centroid	ASC
65	variance of Audio Spectrum Centroid	ASC_V
66	average Audio Spectrum Spread	ASS
67	variance Audio Spectrum Spread	ASS_V
68–87	Spectral Flatness Measure for particular bands	SFM (1–20)
88	SFM average value	SFM_M
89–108	Spectral Flatness Measure variance for particular bands	SFM_V (1–20)
109	averaged SFM variance	SFM_V
110–129	Mel-Frequency Cepstral Coefficients for particular bands	MFCC (1–20)
130–149	MFCC variance for particular bands	MFCCV (1–20)
150	number of samples exceeding RMS	THR_1RMS_TOT
151	number of samples exceeding 2×RMS	THR_2RMS_TOT
152	number of samples exceeding 3×RMS	THR_3RMS_TOT
153	mean value of samples exceeding RMS, averaged for 10 frames	THR_1RMS_10FR_MEAN

Table 2. [Cont.]

No.	Parameter	Abbreviation
154	variance value of samples exceeding RMS, averaged for 10 frames	THR_1RMS_10FR_VAR
155	mean value of samples exceeding 2×RMS, averaged for 10 frames	THR_2RMS_10FR_MEAN
156	variance value of samples exceeding 2×RMS, averaged for 10 frames	THR_2RMS_10FR_VAR
157	mean value of samples exceeding 3×RMS, averaged for 10 frames	THR_3RMS_10FR_MEAN
158	variance value of samples exceeding 3×RMS, averaged for 10 frames	THR_3RMS_10FR_VAR
159	peak to RMS ratio	PEAK_RMS_TOT
160	mean value of the peak to RMS ratio calculated in 10 subframes	PEAK_RMS10FR_MEAN
161	variance of the peak to RMS ratio calculated in 10 subframes	PEAK_RMS10FR_VAR
162	Zero Crossing Rate	ZCR
163	RMS Threshold Crossing Rate	1RMS_TCD
164	2×RMS Threshold Crossing Rate	2RMS_TCD
165	3×RMS Threshold Crossing Rate	3RMS_TCD
166	Zero Crossing Rate averaged for 10 frames	ZCR_10FR_MEAN
167	Zero Crossing Rate variance for 10 frames	ZCR_10FR_VAR
168	RMS Threshold Crossing Rate averaged for 10 frames	1RMS_TCD_10FR_MEAN
169	RMS Threshold Crossing Rate variance for 10 frames	1RMS_TCD_10FR_VAR
170	2×RMS Threshold Crossing Rate averaged for 10 frames	2RMS_TCD_10FR_MEAN
171	2×RMS Threshold Crossing Rate variance for 10 frames	2RMS_TCD_10FR_VAR
172	3×RMS Threshold Crossing Rate averaged for 10 frames	3RMS_TCD_10FR_MEAN
173	3×RMS Threshold Crossing Rate variance for 10 frames	3RMS_TCD_10FR_VAR

To start with the whole parameter set has been taken into consideration but the data analysis leads to a conclusion that only some of them might be useful for mood recognition. To determine parameters, which are the most significant to mood description, the correlation analysis was applied. Correlation between MDS dimensions and mood parameters was calculated and therefore a set of features related to mood was created. Significance of correlation was determined according to t-student (with significance equal to 0.025 ($t_{.975}$)). Finally the feature vector describing mood of music consisted of 79 parameters correlated with Dimension 1 and 9 correlated with Dimension 2, listed in Table 3. Correlation reached up to 0.92 for Dimension 1 and 0.82 for Dimension 2. Due to the length of the feature vector correlated with Dimension 1, only beginning and the end of the sequence is listed. Remaining parameters numbered from 8 to −69 include mainly ASE-, SFM- and RMS-based parameters.

Table 3. Set of parameters used for Dimension 1 (related to "Calm" according to MDS experiment and Dimension 2 (related to "Joyful" according to MDS experiment (Plewa, Kostek, 2013)) description. Denotations are the same as in Table 2. Additionally. MEAN, V and FR correspond to the parameter mean, variance values and the number of frames of analysis.

No.	Dimension 1		No.	Dimension 2	
	Parameter	Corr.		Parameter	Corr.
1	ASE15	0.92	1	ASC	0.82
2	ZCR	0.92	2	MFCCV4	0.76
3	ZCR_10FR_MEAN	0.92	3	SC	0.75
4	ASE29	0.91	4	MFCCV7	0.69
5	ASE28	0.90	5	MFCC5	0.65
6	SFMV1	0.89	6	MFCC10	0.63
7	SFM15	0.89	7	MFCCV6	0.63
8–69	...	...	8	ASE1	0.63
79	ASEV23	0.63	9	MFCCV8	0.62

4.2. Principal component analysis

Two sets of chosen parameters related to mood of music consisted of parameters that were strongly correlated to particular dimensions representing mood of music are considered. Principal Component Analysis was applied to achieve possible most orthogonal dimensions (Smith, 2002) and reduce redundant data.

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system (Jolliffe, 2002). The new variables, called the principal components, are defined as linear functions of the original variables and are meant to be new, orthogonal dimensions (Smith, 2002). Components cannot be directly interpreted, although their loading by specific features can be estimated. If the first few principal components account for a large percentage of the information included in data, they can be used to simplify subsequent analyses (Jolliffe, 2002). It is worth noting that Principal Components Analysis is included in various software such as XLStat (2015), SIMKA-P (UMETRIX, 2015), MATLAB (2015) and is commonly used as a tool for data reduction, also in area of music technology, where large sets of data are very frequently encountered.

Małecki (Małecki, 2013) applied Principal Component Analysis to reduce features describing acoustics of the sacral objects. Kaminsky (1995), using PCA, reduced 80 elements vector to 3 components (covering 88.8% of total variation) for the purpose of Nearest Neighbor (kNN) and Artificial Neural Network (ANN) classification of musical instrument sounds.

Principal Components Analysis was applied to two sets: one consisting of 79 parameters related to Dimension 1 and the second consisting of 9 parameters related to Dimension 2. All of the PCA calculations were performed using MATLAB (2015). The following results were received from the Principal Components Analysis:

- For Dimension 1 ("Calm") 7 components are sufficient to contain 99% of information,
- For Dimension 2 ("Joyful") 6 components are sufficient to contain 99% of information.

As a result vector describing Dimension 1 was shortened to 7 components and Dimension 2 to 6 components.

5. Results

The correlation between subjective mood description and parameters reached up to 0.92 for Dimension 1 and 0.82 for Dimension 2. Components achieved from PCA were treated as the SOM input (7 components for Dimension 1 and 6 for Dimension 2).

SOM analyses were performed for various topographies and sizes of the neural network. For a 2-dimensional SOM (2D SOM) representation, the best results were achieved for the grid topology with network dimensions of 5×5. In this case, the feature vector consisting of 13 elements describing both dimension, Calm and Joyful, was used. These settings enabled to achieve quite good representation in one of the dimensions, but did not succeed in another. An example of 2D SOM representation is shown in Fig. 4. Songs are placed on neurons with the highest activation.

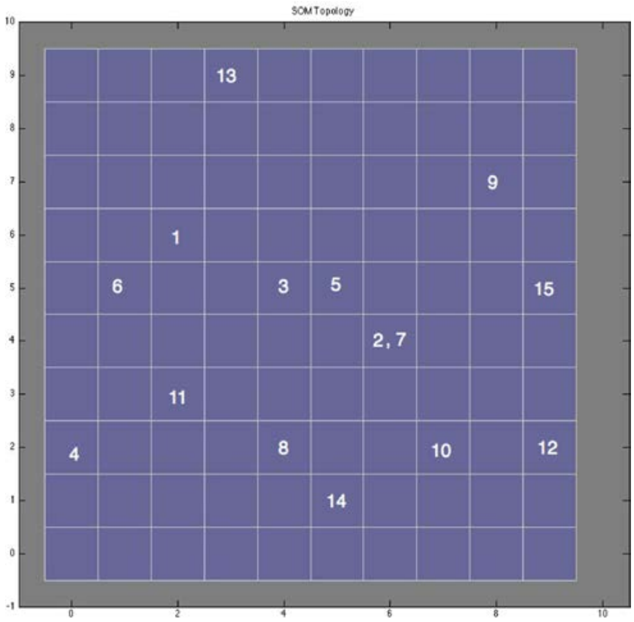

Fig. 4. An example of 2D SOM representation of 15-element music set. Numbers represent particular songs, listed according to Table 1. Studies of the particular cases allow observing quite good results in one of the dimensions.

Due to not satisfying results of 2D representations and promising trends according to one of the dimensions, two separate 1-dimensional SOM networks were constructed. Two vectors were created: one related to Dimension 1 (7 PCA components) and one to Dimension 2 (6 components). This allowed achieving a good representation for Dimension 1 ("Calm"), shown in Fig. 5. Only song labeled with no. "14" (marked in the picture with the oval) was not assigned correctly. Location on the "Calm" axis is not accurate. The other elements are placed properly and their positions are coherent with the MDS-based results. Song no. 15 is evaluated as "the least calm" and that is coherent with the subjective evaluation as well as music genre (Hard Rock & Metal).

Contrarily, representation of Dimension 2 ("Joyful") is less accurate but also contains correct assignments and is presented in Fig. 6. Wrongly placed songs (nos. 1, 6 and 9) are marked in the picture with ovals. Songs nos. 1 and 6 were evaluated as very similar in terms of mood, what was properly recognized by both 2D SOM and SOM related to Dimension 2. It is worth

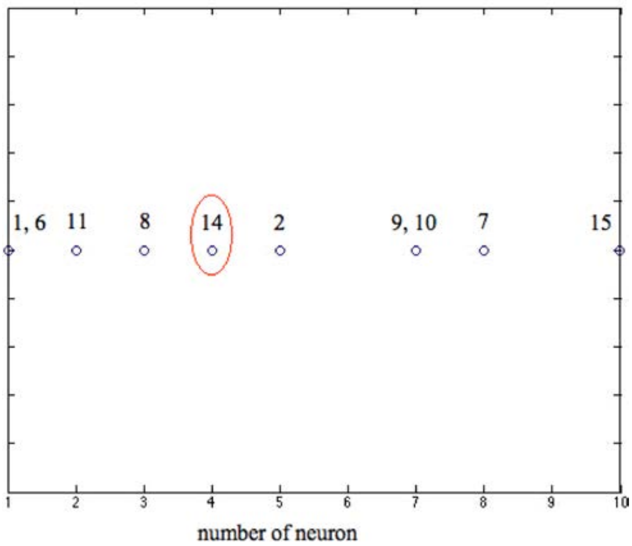

Fig. 5. SOM representation of 10-elements music set for Dimension 1 ("Calm"). Numbers represent particular songs, listed accordingly to Table 1. Song labeled with no. "14" is marked due to the inaccurate location.

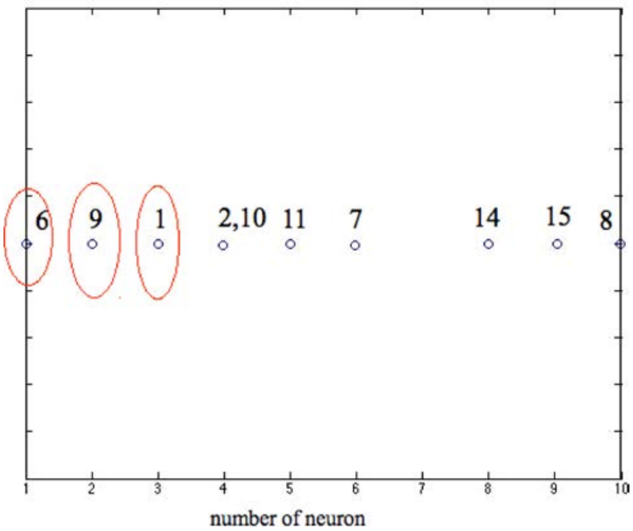

Fig. 6. SOM representation of 10-elements music set for Dimension 2 ("Joyful"). Numbers represent particular songs, listed accordingly to Table 1. Songs located improperly are marked with ovals.

noting that relation between these three songs was properly assigned by 2D SOM.

Even though the accuracy for Dimension 1 is around 90% and for Dimension 2 around 70%, it is much higher than for the 2D representation. At the same time 2D SOM achieved better results for some relations between songs (i.e. nos. 1,6 and 9). Finally the approach for two separate SOMs was chosen because it seems to be more appropriate for this task.

In addition, a question arises whether it is possible to compare results obtained based on unsupervised (SOM) and classification based on a supervised technique, i.e. Artificial Neural Networks (ANN). For this purpose a larger set of music consisting of 150 songs was used. Mood of music excerpts was described by the listeners who used a web-based survey. Since the results of the listening experiment are not the main focus of the presented study, they are recalled only as a reference point for automatic classification. The ANN-based classification was performed using *nntool* within the Matlab environment (MATLAB). A feed-forward ANN with one hidden layer was trained to classify music excerpts into 4 quadrants of Thayer's VA plane (described in Sec. 2). Feature vectors derived from the SYNAT music database were fed to the input of ANN. Different configurations of ANN were tested and but the best results were obtained for a network with 15 neurons in the hidden layer. Higher accuracy was achieved for two separate networks – one dedicated to each dimension (i.e. Valence and Arousal), due to utilizing one network with 4 outputs. Overall, results of 76% accuracy for Valence and 83% for Arousal were obtained. Even though this outcome quantitatively shows how similar the results are, qualitative interpretation differs. ANN classifies into 4 classes (quadrants of VA plane), while SOM results indicate position on the VA plane more precisely.

It is worth noticing that songs nos. 6 and 9 were misclassified in the process of the ANN-based classification, which was also the case for SOM (Fig. 6). However, listeners and experts evaluated both songs quite confidently, which was also seen in the standard deviation results that were even lower than for other music pieces. This can lead to the conclusion that some very significant (and specifically music piece-related) features may be missing in the FV.

6. Conclusions

Although high correlation was obtained between subjective labeling of mood and objective descriptors in the carried out experiments, it is arguable whether assigning emotion label is sufficient to describe mood of music, and secondly whether this may be treated as ground truth in music mood recognition. This is because one's emotions caused by music may easily be mixed with mood intentionally ascribed to a piece of music by a musician.

Principal Component Analysis was very suitable for selecting variables that contain significant information, which resulted in dimension reduction of data. Moreover, the data pre-processing stage is appropriate for SOM analyses. It may also be observed that two separate SOMs, each dedicated to particular dimension, are more efficient than 2D representation.

Distribution of objects on the SOM representation is coherent with the MDS visualization. In all of cases, dimension related to "Calm" is easier to describe and determine than dimension related to "Joyful". This descriptor is much more subjective and probably some more complex rules have to be defined to determine the distribution of songs along this dimension. Overall, achieved results can lead to a conclusion that SOM is a powerful tool to visualize music dataset basing on mood of music distribution.

The ANN-based classification was tested on a bigger music set but some songs were misclassified by both methods (ANN and SOM). Accuracy of the ANN-based classification reached up to 76% for Valence and 83% for Arousal, but it refers only to quadrants of VA plane. Contrarily, SOM results can return Valence and Arousal values more precisely but are more difficult to interpret.

The SOM-based approach to Music Mood Recognition will be developed on a larger music set. Since the chosen topology and number of neurons in SOMs are dependent on the size of the dataset, different SOM configurations will further be tested.

Acknowledgments

This work was partially supported by the grant no. PBS1/B3/16/2012 entitled "Multimodal system supporting acoustic communication with computers" financed by the Polish National Centre for Research and Development.

References

1. AMAZON, http://www.amazon.com/
2. Barbedo J.G.A., Lopes A. (2005), *A New Cognitive Model for Objective Assessment of Audio Quality*, J. Audio Eng. Soc., **53**, 1/2, 22–31; February.
3. Bigand E., Vieillard S., Madurell F., Marozeau J., Dacquet A. (2005), *Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts*, Cognition & Emotion, **19**, 8, 1113–1139.
4. Borg I., Groenen P. (2007), *Modern Multidimensional Scaling: Theory and Applications*, Springer, Germany.
5. Casey M.A., Veltkamp R., Goto M., Leman M., Rhodes C., Slaney M. (2008), *Content-Based Music Information Retrieval: Current Directions and Future Challenges*, Proc. of the IEEE, **96**, 4, April.
6. Cruz R., Brisson A., Paiva A., Lopes E. (2007), *I-Sounds – Emotion-based Music Generation for Virtual Environments*, Springer-Verlag, Berlin Heidelberg, 769–770.
7. Drossos K., Floros A., Kermanidis K.L. (2015), *Evaluating the Impact of Sound Events' Rhythm Characteristics to Listener's Valence*, J. Audio Eng. Soc., 63, 3, 139–153; March.
8. Frühwirth M. (2001), *Automatische Analyse und Organisation von Musikarchiven* [in German], Wien.
9. Frühwirth M., Rauber A. (2001), *Self-Organizing Maps for Content-Based Music Clustering*, Proceedings of the 12th Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy.
10. Hevner K. (1936), *Experimental studies of the elements of expression in music*, American Journal of Psychology, **48**, 246–268.
11. Hoffmann P., Kostek B. (2014), *Music Data Processing and Mining in Large Databases for Active Media*, Active Media Technology, LNCS, **8610**, 85–95, Springer.
12. Huron D. (2000), *Perceptual and cognitive applications in music retrieval*, Proc. Int. Conf. MIR.
13. ISMIR, Intern. Conference on Music Information Retrieval website (http://ismir2015.ismir.net).
14. ITUNES, https://www.apple.com/pl/itunes/ (accessed August 2015).
15. Jolliffe I.T. (2002), *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY.
16. Kaminsky I. (1995), *Automatic source identification of monophonic musical instrument sounds*, Proceedings., IEEE International Conference on Neural Networks.
17. Kim Y.E., Schmidt E., Emelle L. (2008), *MoodSwings: A collaborative game for music mood label collection*, [in:] ISMIR, Philadelphia, PA, September 2008.
18. Kim Y.E., Schmidt E.M., Migneco R., Morton B.G., Richardson P., Scott J., Speck J.A., Turnbull D. (2010), *Music Emotion Recognition: A State of the Art Review*, 11th International Society for Music Information Retrieval Conference (ISMIR 2010), 255–266.
19. Kohonen T. (1984), *Self Organized Formation of Topologically Correct Feature Maps*, Biol. Cybern., **43**, 59–69.
20. Kohonen T. (1984), *Self-Organisation and Associative Memory*, Springer Verlag.
21. Kohonen T., Honkela T. (2007), *Kohonen Network*, *Scholarpedia*.
22. Kostek B. (2011), *Content-Based Approach to Automatic Recommendation of Music*, 131st Audio Eng. Soc. Convention, October 20–23, New York.
23. Kostek B. (2013), *Music Information Retrieval in Music Repositories*, Chapter 17, [in:] Rough Sets and Intelligent Systems, Skowron A., Suraj Z., [Eds.], 1, ISRL, 42, 463–489, Springer Verlag, Berlin Heidelberg.
24. Kostek B. (2014), *Auditory Display Applied to Research in Music and Acoustics*, Archives of Acoustics, **39**, 2, 203–214, DOI 10.2478/aoa-2014-0025.
25. Kostek B., Czyzewski A. (2001), *Representing Musical Instrument Sounds for Their Automatic Classification*, J. Audio Eng. Soc., **49**, 9, 768–785.
26. Kostek B., Hoffmann P., Kaczmarek A., Spaleniak P. (2013), *Creating a Reliable Music Discovery and Recommendation System*, Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation, Springer Verlag.

27. Kostek B., Kaczmarek A. (2013), *Music Recommendation Based on Multidimensional Description and Similarity Measures*, Fundamenta Informaticae, **127**, 1–4, 325–340, DOI 10.3233/FI-2013-912.
28. Kostek B., Kupryjanow A., Zwan P., Jiang W., Raś Z., Wojnarski M., Swietlicka J. (2011), *Report of the ISMIS 2011 Contest: Music Information Retrieval*, Foundations of Intelligent Systems, ISMIS 2011, Kryszkiewicz M. *et al.* [Eds.], LNAI 6804, 715–724, Springer Verlag, Berlin, Heidelberg.
29. Laurier C., Sordo M., Serra J., Herrera P. (2009), *Music Mood Representations from Social Tags*, Proceedings of the 10th International Society for Music Information Conference, Kobe, Japan, 381–386.
30. Lewis C.I. (1929), *Mind and the World Order*, New York.
31. Lima M.F.M., Machado J.A.T., Costa A.C. (2012), *A Multidimensional Scaling Analysis of Musical Sounds Based on Pseudo Phase Plane*, Appl. Anal., Special Issue (2012), Article ID 436108.
32. Lu L., Liu D., Zhang H.J. (2006), *Automatic mood detection and tracking of music audio signals*, IEEE Trans. Audio Speech Language Processing, **14**, 1, 5–18, Jan.
33. Małecki P. (2013), *Evaluation of objective and subjective factors of highly reverberant acoustic field* [in Polish], Doctoral Thesis, AGH University of Science and Technology, Krakow.
34. Markov K., Matsu T. (2014), *Music Genre and Emotion Recognition Using Gaussian Processes*, 2169–3536 IEEE, **2**, 688–697.
35. MATLAB, Neural Network Toolbox, available at www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf (accessed January, 2015).
36. MIREX (2009), Mood Multi Tag Data Description, http://www.music-ir.org/archive/papers/Mood_Multi_Tag_Data_Description.pdf
37. Mostafavi A.C., Ras Z., Wieczorkowska A. (2013), *Developing Personalized Classifiers for Retrieving Music by Mood*, [in:] New Frontiers in Mining Complex Patterns NFMCP 2013, International Workshop, held at ECML-PKDD 2013, 27th September 2013, Prague, Czech Republic.
38. MUFIN system; http://www.mufin.com/us/ (accessed August 2015).
39. MUSICOVERY system; http://musicovery.com/ (accessed August 2015).
40. Novello A., Mckinney M.M.F., Kohlrauschab A. (2011), *Perceptual Evaluation of Inter-song Similarity in Western Popular Music*, J. New Music Research, **40**, 1, 1–26.
41. Palomäki K., Pulkki, V., Karjalainen M. (1999), *Neural Network Approach to Analyze Spatial Sound*, AES 16th International Conference On Spatial Sound Reproduction, Finland.
42. Pampalk E. (2001), *Islands of Music, Analysis, Organization, and Visualization of Music Archives*, Master Thesis, Vienna Technical University.
43. Pampalk E., Rauber A., Merkl D. (2002), *Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Map*, Proceedings of the International Conference on Artificial Neural Network, 871–876.
44. Panda R., Paiva R.P. (2011), *Using Support Vector Machines for Automatic Mood Tracking in Audio Music*, 130th Audio Eng. Soc. Convention, Paper No. 8378, London, UK, May 13–16.
45. PANDORA – Internet Radio: http://www.pandora.com
46. Papaodysseus C., Roussopoulos G., Fragoulis D., Panagopoulos A., Alexiou C. (2001), *A New Approach to the Automatic Recognition of Musical Recordings*, J. Audio Eng. Soc., **49**, 1/2, 23–35; February.
47. Plewa M., Kostek B. (2013), *Multidimensional Scaling Analysis Applied to Music Mood Recognition*, 134 Audio Eng. Soc. Convention, May 4–7, Paper No. 8876, Rome.
48. Raś Z., Wieczorkowska A. (2010), *Advances in Music Information Retrieval, Series: Studies in Computational Intelligence*, Vol. 274, Springer-Verlag, Berlin Heidelberg 2010, ISBN: 978-3-642-11673-5.
49. Rauber A., Frühwirth M. (2001), *Automatically Analyzing and Organizing Music Archives*, [in:] Proceedings of the 5. European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001), Sept. 4–8 2001, Darmstadt, Germany, Springer Lecture Notes in Computer Science, Springer.
50. Rauber A., Pampalk M., Merkl D. (2002a), *Content-based Music Indexing and Organization*, Proc. 25 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 02), 409–410, August 11–15, Tampere, Finland.
51. Rauber A., Pampalk M., Merkl D. (2002b), *Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structuring of Music by Musical Styles*, Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR 2002), 71–80, October 13–17, Paris, France.
52. Rojas R. (1996), *Neural Networks: A Systematic Introduction*, Berlin, Springer.
53. Rosner A., Schuller B., Kostek B. (2014), *Classification of Music Genres Based on Music Separation into Harmonic and Drum Components*, Archives of Acoustics, **39**, 4, 629-638, DOI: 10.2478/aoa-2014-0068.
54. Smith L. (2002), *A tutorial on Principal Components Analysis*, available at http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf (accessed May, 2015).
55. Schmidt E.M., Kim Y.E. (2010), *Prediction of time-varying musical mood distributions using Kalman filtering*, Proceedings of the 2010 IEEE International Conference on Machine Learning and Applications, Washington, D.C.: ICMLA.
56. Schubert E. (2003), *Update of the Hevner adjective checklist*, Perceptual and Motor Skills, **96**, 1117–1122.

57. Rumsey F. (2011), *Semantic Audio: Machines Get Clever with Music*, J. Audio Eng. Soc., **59**, 11, 882–887, April.

58. Rumsey F. (2014), *About Semantic Audio*, J. Audio Eng. Soc., **62**, 4, 281–285, April.

59. Tadeusiewicz R. (1993), *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa.

60. Thayer R.E. (1989), *The Biopsychology of Mood and Arousal*, Oxford University Press.

61. Trochidis K., Delbé. C., Bigand E. (2011), *Investigation of the relationships between audio features and induced emotions in Contemporary Western music*, SMC Conference.

62. Tuzman A. (2001), *Wavelet And Self-Organizing Map Based Declacker*, Paper No. 1959, 20th International Conference: Archiving, Restoration, and New Methods of Recording.

63. Tzanetakis G., Cook P. (2002), *Musical genre classification of audio signal*, IEEE Transactions on Speech and Audio Processing, **10**, 3, 293–302.

64. Ultsch A. (2003), *U*-Matrix: A tool to visualize clusters in high dimensional data*, Department of Computer Science, University of Marburg, Technical Report Nr. 36:1–12.

65. UMETRIX (2015), http://umetrics.com/sites/default/files/books/sample_chapters/multimega_parti-3_0.pdf, access 10.06.2015.

66. Wagenaars W.M., Houtsma A.J., Van Lieshout R.A. (1986), *Subjective Evaluation of Dynamic Compression in Music*, J. Audio Eng. Soc., **34**, 1/2, 10–18; February.

67. Wieczorkowska A., Kubera E., Kubik-Komar A. (2011), *Analysis of Recognition of a Musical Instrument in Sound Mixes Using Support Vector Machines*, Fundamenta Informaticae, **107**, 1.

68. XLStat (2015), http://www.xlstat.com (accessed June 2015).

69. Zentner M., Grandjean D., Scherer K. (2008), *Emotions evoked by the sound of music: Characterization, classification, and measurement*, Emotion, **8**, 494–521.