GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRONICS, TELECOMMUNICATIONS AND INFORMATICS

The author of the PhD dissertation: Magdalena Plewa
Scientific discipline: Telecommunications

## DOCTORAL DISSERTATION

Title of PhD dissertation: **Automatic Mood Indexing of Music Excerpts based on Correlation Between Subjective Evaluation and Feature Vector**

Title of PhD dissertation (in Polish): **Automatyczna indeksacja bazy muzycznej na podstawie korelacji pomiędzy oceną subiektywną nastroju utworu muzycznego a wektorem parametrów**

| | |
|---|---|
| Supervisor<br><br>*signature* | Second supervisor<br><br>*signature* |
| prof. dr hab. inż. Bożena Kostek | |
| Auxiliary supervisor<br><br>*signature* | Cosupervisor<br><br>*signature* |
| | |

Gdańsk, year 2015

# DOCTORAL DISSERTATION

## Automatic Mood Indexing of Music Excerpts based on Correlation Between Subjective Evaluation and Feature Vector

Magdalena Plewa

SUPERVISOR

prof. dr hab. inż. Bożena Kostek

Gdańsk, 2015

**Acknowledgments**

I would like to thank my supervisor **Prof. Bozena Kostek** for giving me the opportunity to focus on a very specific topic that is very interesting to me and important for my personal experience of music. I would also like to express my gratitude for solving all potential issues related to traversing the whole country as well as supporting my professional involvement, internships and travelling ideas, all of which enabled me to broaden my horizons and gain various experiences.

I want to thank my colleagues at Multimedia Systems Department for all the helpful discussions on the issues related to my Ph.D. work. I also would like to thank **Dr Mateusz Bień** from Music University in Kraków for his support with visualization tools, **Dr Jakub Pierzchała** for rewarding debates related to the graphical interface and **Dr Paweł Małecki** for technical support. At the same time I would like to express gratitude to all the listeners who participated in numerous subjective tests as well as all the musicians who contributed their music and performances during recording sessions.

I would like to thank my mentors and colleagues at the University of Stavanger and the Banff Centre for sharing their points of view and inspiring discussions.

I would like to thank **Annalie Wilson** for her interest in the topic and for polishing my English.

I am very grateful to **Prof. Jan Adamczyk** for making things happen and making me believe that everything is possible.

I am grateful also to **my parents** for their support during all the stages of my education and for their understanding and respect for my choices.

Special thanks to my partner, **Szymon Piotrowski**, for his patience and support.

The present thesis 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.

**ABSTRACT**

With the growth of accessible digital music libraries over the past decade, there is a need for research into automated systems for searching, organizing and recommending music. As mood of music is considered as one of the most intuitive criteria for listeners, this work is focused on an approach based on the emotional content of music and its automatic recognition. An overview of audio signal parametrization was carried out, with the main focus on features related to music characteristics. In addition, a novel analysis of single instruments tracks versus mix, aimed at mood of music recognition, was performed. Moreover, original parameters describing rhythmic content in different frequency ranges were proposed. Research presented in this work contains a series of experiments related to models and description of emotions in music. As a result a graphical model dedicated to the subjective evaluation of mood of music was proposed and created. A music set consisting of 154 excerpts from 10 music genres was evaluated in the listening experiment. Achieved results indicated a strong correlation between subjective results and objective descriptors and on that basis a vector of parameters related to mood of music was created. Automatic mood recognition employing SOM and ANN and was carried out. The comparison between outcomes achieved from both methods and subjective evaluation was performed and that led to the conclusion that all methods returned coherent results. The accuracy of automatic classification was satisfactory as it outperformed literature-based results, and its success was particularly notable considering the subjective character of the analyzed material.

**STRESZCZENIE**

Gwałtowny przyrost liczby bibliotek muzycznych (oraz ich zawartość sięgająca aktualnie milionów utworów muzycznych) łatwo dostępnych przez Internet spowodował rozwój badań w kierunku automatycznych systemów wyszukiwania, organizacji i rekomendacji muzyki. Nastrój muzyki jest uważany za najbardziej intuicyjne kryterium opisu muzyki przez słuchaczy, dlatego też w prezentowanej pracy skoncentrowano się na organizacji muzyki w kontekście zawartych w niej emocji. Przeprowadzono przegląd parametrów fonicznych ze szczególnym uwzględnieniem deskryptorów opisujących strukturę muzyczną. Wykonana została też nowatorska analiza ścieżek poszczególnych instrumentów w porównaniu do całości miksu ukierunkowana na emocje zawarte w poszczególnych ścieżkach. Na tej podstawie zaproponowane zostały oryginalne parametry opisujące zawartość rytmiczną w poszczególnych pasmach. W ramach pracy została przeprowadzona seria eksperymentów dotyczących modeli i opisu nastroju muzyki, wynikiem czego był autorski graficzny model emocji dedykowany do subiektywnej notacji emocji zawartych w muzyce. Baza 154 utworów muzycznych należących do 10 gatunków została poddana subiektywnym testom odsłuchowym mającym na celu określenie nastroju przypisanego do utworów. Uzyskane wyniki posłużyły do analizy korelacyjnej i uzyskania wektora parametrów opisujących nastrój muzyki. W procesie automatycznego rozpoznawania nastroju muzyki wykorzystano algorytmy SOM oraz ANN. Porównanie wyników uzyskanych z obu algorytmów oraz testów subiektywnych pokazało, że są one spójne. Dokładność automatycznego rozpoznania nastroju została uznana za satysfakcjonującą, a nawet przewyższająca wyniki uzyskane przez innych badaczy. Wynik ten jest zadowalający, biorąc również pod uwagę subiektywny charakter analizowanego zagadnienia.

**ADALINE** - ADAptive LiNear Element

**ADSR** - Attack Decay Sustain Release

**AFF** - Audio Fundamental Frequency

**AH** - Audio Harmonicity

**AMT** - Auditory Modeling Toolbox

**ANN** - Artificial Neural Network

**AP** - Audio Power

**ASB** - Audio Spectrum Basis

**ASC** - Audio Spectrum Centroid

**ASE** - Audio Spectrum Envelope

**ASF** - Audio Spectrum Flatness

**ASP** - Audio Spectrum Projection

**ASS** - Audio Spectrum Spread

**AW** - Audio Waveform

**BOG** - Bag Of Frames

**FL** - Fuzzy Logic

**FFS** - Forward Feature Selection

**GEMS** - Geneva Emotional Music Scale

**GGM** - Gaussian Mixture Models

**GHSOM** - Growing Hierarchical Self-Organizing Maps

**HMER** - Hierarchical Music Emotion Recognition

**HMM** - Hidden Markov model

**HR** - *Harmonic Ratio*

**HSC** - Harmonic Spectral Centroid

**HSD** - Harmonic Spectral Deviation

**HSS** - Harmonic Spectral Spread

**HSV** - Harmonic Spectral Variation

**ICA** - Independent Component Analysis

**ID3** - Metadata container most often used in conjunction with the MP3 audio file format

**KNN** - K Nearest Neighbours Algorithm

**LAT** - Log Attack Time

LIST OF SYMBOLS AND ABBREVIATIONS

**LPC** - Linear Prediction

**MARSYAS** - Music Analysis Retrieval and Synthesis for Audio Signals

**MDS** - Multidimensional Scaling

**MER** - Music Emotion Recognition

**MFCC** - Mel-Frequency Cepstral Coefficients

**MIDI** - Musical Instrument Digital Interface

**MIR** - Music Information Retrieval

**MPEG-7** - standard is a set of standardized tools to describe multimedia content

**MUSHRA** - MUltiple Stimuli with Hidden Reference and Anchor

**MTBF** - Modified Time-Based Features

**NMF** - Non-Negative Matrix Factorization

**PCA** - Principal Component Analysis

**RMS** - Root Mean Square

**SC** - Spectral Centroid

**SCR** - Skin Conductivity Response

**SFM** - Spectral Flatness Measure

**SOM** - Self-organizing Map

**SSD** - Statistical Spectrum Descriptors

**STFT** - Short Time Fourier Transform

**SVD** - Singular Value Decomposition

**SVM** - Support Vector Regressor

**SVR** - Support Vector Machine

**SYNAT** - is a database of 52532 pieces of music hosted at Gdansk University of Technology Multimedia System Department

**TBF** - Time-Based Features

**TC** - Temporal Centroid

**ULH** - *Upper Limit* of *Harmonicity*

**VA** - Valence/Arousal

$\boldsymbol{\sigma_x}$ - standard deviation of signal x

$\boldsymbol{\mu_x}$ - mean value of signal x

$\boldsymbol{\rho}$ - correlation coefficient

LIST OF SYMBOLS AND ABBREVIATIONS

***Stress-1*** - value, which reflects error of the map obtained from Multidimensional Scaling

**φ** - activation function of neuron in artificial neural network

$w^{(m)}$ - weight corresponding to neuron m

**η** - speed of learning of Self-Organized Map

***h*** - neighborhood function of neuron in artificial neural network

$N_a$ - *set of units within a* neighborhood of neuron in artificial neural network

***μ(x)*** - membership grade in fuzzy logic

# LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

# LIST OF TABLES

LIST OF TABLES

# TABLE OF CONTENTS

# 1 INTRODUCTION

The need of music accompanies people from thousands of years [14]. It provides a means by which people can share emotions, intentions, and meanings despite different cultures or languages [200]. At the same time music perception is not only strongly influenced by individual background and preferences, but also has deep roots in social and cultural trends. Music itself is strongly associated with perception. Isaac Newton was the first to point out that light is colorless and "The waves themselves are not colored." [174]. According to that statement, color is the interpretation of physical phenomena by the human brain, based on complex processes. A very similar mechanism can be observed with sound and music perception - music has to be perceived inside our brains. Melody, rhythm, timbre or any other subjective attributes start to exist when the human perception system and the human brain interact.

In music perception studies many different classifications and systems that describe music components are defined. Levitin [174] observes that from the listener's perspective there are seven major elements of music: loudness, pitch, melody, harmony, rhythm, tempo, and meter. These components are significant for discussion related to emotions included in music.

The traditional approach to studying music emotion perception consists in subjective tests, in which a number of listeners evaluate a given music excerpt, and then these results are analyzed using statistical processing. Therefore the area of psychoacoustics supports the researchers with a tool to evaluate all aspects related to music, however this process is very lengthy and arduous, and does not always return reliable results. Therefore, there is a need for automatization in this field, especially as music emotion evaluation/annotation becomes one of the very important topics, also music industry salient, evolving into Music Emotion Recognition (MER) [50].

Music Information Retrieval (MIR) [271] is a multidisciplinary field of research studies, which embraces musicology, psychology, music performance, signal processing, audio signal parametrization (e.g. MPEG-7 was invented for the needs of MIR), artificial intelligence methods and others topics. The main goal of MIR is to find information about music by engaging intelligent, automated processing, automatic music description and evaluation of the proposed solutions [50]. MIR is highly involved in recommendation

systems and many recent studies are dedicated to this topic [142,230,270,330]. Due to an enormous amount of music that is reachable online, a new approach based on more specific targeting is observed rather than continuous extension of accessible content only [57,198]. Different systems are based on various strategies from collaborative filtering [96,162,297], through metadata and lyrics information [113,116,295] to the content-based approach [237,250,331,339].

Metadata-based content search is the most common, as well as very powerful, method of organizing music databases [50]. It is used by many music download services and has reached a degree of success with them. However, there are disadvantages of this approach as it is extremely difficult to maintain consistent expressive metadata description. It is estimated that it takes about 20–30 minutes per track of one expert's time to enter the metadata [236], which incurs an enormous cost. On the other hand, in the content-based music description, information including digital audio signal is retrieved. Content-based methods are not being developed to replace but to enhance metadata-based systems. Within this approach, music is treated as any other signal but dedicated measures are defined to describe values that are relevant to the topic. Low-level audio features are measures of audio signals that contain information about a musical piece and music performance [50]. These descriptors encompass not only the desired information but also intercorrelated factors due to the difficulty of precisely measuring just a single aspect of music. This refers to the whole concept of signal parametrization and finding signal descriptors that contain information about specific aspects of music.

The beneficiaries of developing methods for music searching are students and researchers dealing with trends in music, musicologists, people monitoring trends in this field, as well as music industry-interested parties. There are three main groups of recipients: those involved in the music industry (producers, labels), end users (customers listening to music, personalized media), and professionals (musicians, producers, teachers, researchers, musicologists, lawyers, etc.). Studies related to the MIR include both contemporary and archival collections.

Mood is one of the features that is useful and intuitive for listeners when describing a piece of music [50]. However, even if it seems to be the easiest way to qualify music for people who are non-experts, it is very difficult to find an exact correlation between physical features and perceived mood, which is necessary to make the annotation process automatic.

# 1 INTRODUCTION

Mood and emotion are closely related affective phenomena. In order to distinguish, in this context, emotional content of music and music-evoked emotions, it is called here **mood of music**. In the literature terms "**music emotions**" and "**mood of music**" are often used interchangeably. **However, for the purpose of the presented dissertation both of them describe mood and emotions included in music, as opposed to music cognition studies, where "emotions to music" are the listener's personal experience of feelings evoked by music.** This distinction might seem to be unnecessary, but it is introduced to clarify the nature of studies that involve only musical content analysis without measuring additional external factors. In this work emotions of the listener are not analyzed, but a **study of how listeners would judge the mood of a particular musical fragment is performed**. Discussion aims towards rules, which could predict how the listener would describe musical content by a specific mood.

The relationship between music and emotions has been a subject of some studies in the last decades, where researchers tried to identify the influence of certain musical attributes such as tempo, mode, rhythm and others on the human subject's perception. This content-based approach leads to the field where mood and emotions are examined as components of the musical signal. Professionals describe music with many sophisticated expressions, while the perception process takes place in the brains of people who are not educated in music. Therefore we should try to discover a relationship between signal (in this case music denoted as an audio signal in the digital domain) and the listener's perception of the mood of music.

Recently, music mood recognition has become a thorough subject of research studies and analyses within MIR [55,131,168,188,194,235,330,331,339,349]. This area of research studies is called **Music Emotion Recognition (MER)** and aims at recognizing emotions contained in music signals [114,234,243,254-256]. A considerable part of the research involved in MER is based on tags, semantics and lyrical content. Another approach is to use advanced computational intelligence methods (i.e. Support Vector Machines, Support Vector Regression, Gaussian Mixture Model, Ranked Attributes Tree and many others) aimed at automatic recognition of mood of music [104,182,237].

Despite decades of research in this area, the results obtained in the works listed above leave space for further research. Many authors stated in their final conclusions that the issue has not been completely solved [104,168,188], neither has mood perception been

fully understood. Therefore, in the research study undertaken by the author, methods from Self-Organizing Maps Artificial Neural Networks supported by fuzzy logic-based data processing to model characteristics of human perception are to be applied. Otherwise, it might be perceived as an inconsequence to analyze subjective attributes (emotions) in an objective domain (audio signal features) only, using a very strict (crisp) relationship. Therefore, fuzzifying the boundaries of mood-related data in the process of automatic recognition by the ANN algorithm increases automatic mood classification accuracy.

Some attributes of artificial intelligence computational methods are more similar to characteristics of the human learning process, perception and cognition. Each person has a different "starting point" that can be related to personal attributes such as cultural background or sensitivity to detail. At the same time, everyone has their own scene of music that they know or appreciate, that they have associated with a particular moment, that they remember well and that induces particular emotions. It is possible to imagine a situation in which two listeners would have totally opposite reaction to the same piece of music. Therefore it is groundless to expect a very high consistency of the results between listeners and extreme accuracy of the prediction system. Moreover, not all terms related to emotions or mood have the same meaning to listeners. These are the reasons why the research study presented by the author is focused mostly on creation of a mood model that would be easily understood and user-friendly, as well as on the computational methods that are close to a human's reasoning and perception. One of the partial objectives of this work was to create an original model of mood dedicated to subjective evaluation of emotional content of music. The main assumptions were that the model has to be intuitive for users and compatible with a dimensional model used for mood prediction process. In the course of the study several methods based on the human's perception were employed. Multidimensional Scaling was used to determine the dimensions underlying perception of mood of music. Self-Organizing Maps were employed to map a set of music according to the emotional content obtained in an unsupervised manner and Artificial Neural Networks, trained in a supervised way, were used for classification.

**Aims of the study**

The main aim of the presented thesis is to **introduce** a framework of automatic organization of music based on emotions (see Fig. 1.1). Due to the fact that interpretation of

music depends on the preference and experience of an individual person, as well as the fact that averaged results might be to some extent confusing, it was decided to **create** a complete mapping, which organizes songs in relation to each other and not according to the preferences of a single person. That is why one of the goals is to employ intelligence computational methods based on human perception such as Self-Organizing maps, ANNs or fuzzy logic-based data processing and compare the results obtained with subjective mood evaluation. An additional aim is to analyze tracks containing single instruments and determine whether separation can improve automatic description of mood of music. Another additional aim is to link the results of the correlation analysis between parametric description of a music file to the nature of the mood that can be assigned to a particular piece or excerpt thereof. This is due to the fact that the objective description of the mood of a song is the missing link in the organization of large, distributed music databases. Another partial objective is to create an intuitive model that is used to describe mood of music, extending concepts known in the literature sources. The proposal of this novel model is one of the contributions of the present thesis. Another additional aim is to show, that the dimensional-based approach enables better automatic mood recognition. The stages of analysis executed in the course of this dissertation are presented in Fig. 1.1.

Therefore, as a result of this doctoral dissertation the following three theses are expected to be proved:

**1. It is possible to find parameters describing a musical excerpt, which are highly correlated with subjective mood labeling results.**

**2. Self-organizing maps (SOMs) or artificial neural networks (ANNs) trained employing designed feature vectors can effectively be applied to the automated indexing of mood of musical excerpts.**

**3. Annotations of mood of music achieved by subjective assessments and classifying based on both supervised and unsupervised learning can be coherent.**

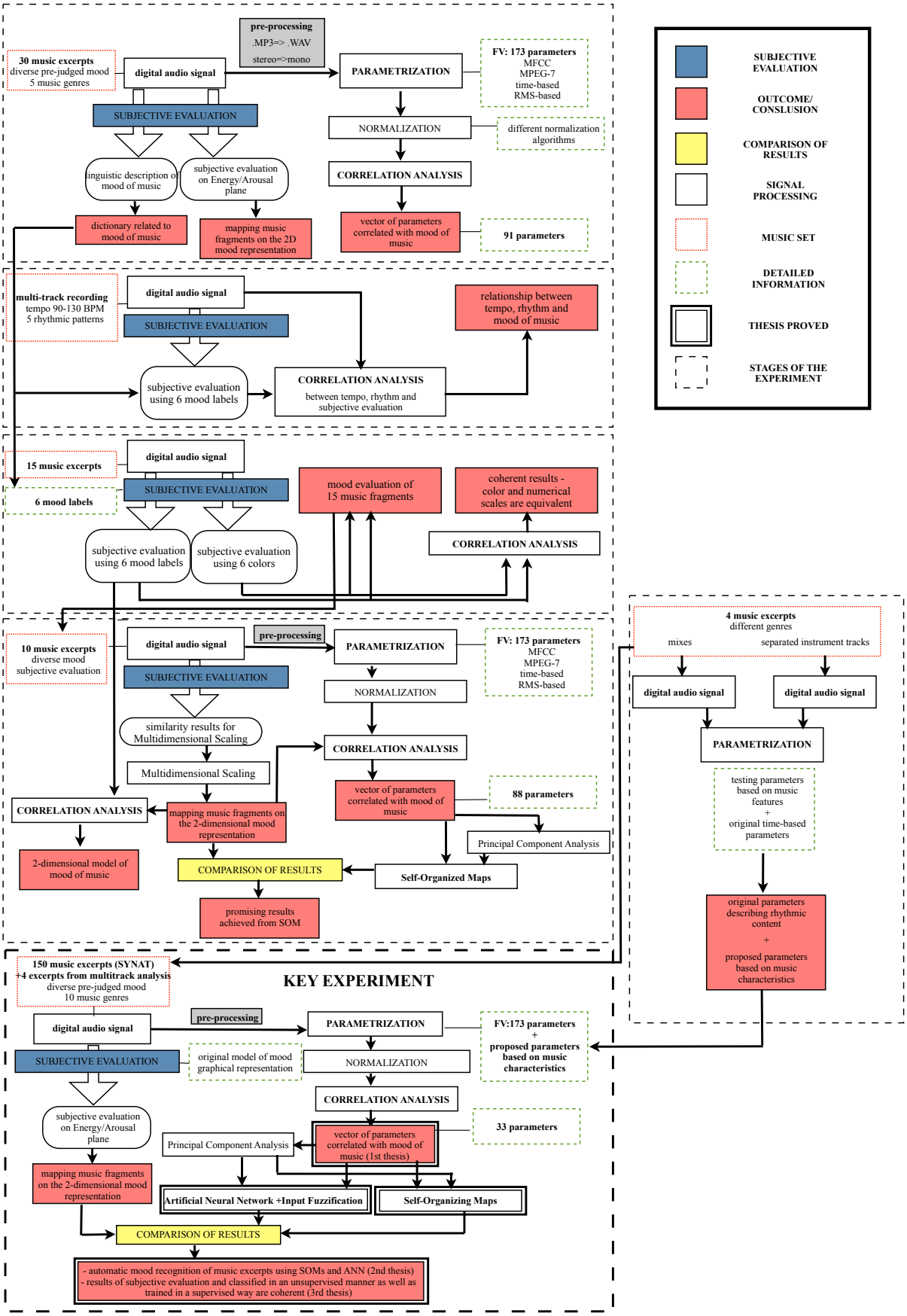



*Figure 1.1 Stages of analysis executed in the course of the present dissertation*

The organization of the presented dissertation is shown in Fig. 1.2, in which all chapters are introduced along with their content. The next two chapters (Chapters 2-3) provide the theoretical background of the work, including perception and evaluation of music, especially in the context of emotions, as well as issues relevant to Music Information Retrieval and an overview of previous studies carried out in the Music Emotion Recognition area. Chapters 4-5 introduce tools used in the course of the present thesis. In Chapter 4 an overview of audio signal parametrization, supplemented with the proposed analysis of tracks of separated instruments and new parameters describing rhythmic content of music, introduced by the author, are shown. Chapter 5 includes theoretical background related to Multidimensional Scaling, Artificial Neural Networks, Self-Organized Maps and fuzzy logic. In Chapter 6 preliminary experiments are executed, and analyses of results and verification of methods are investigated and compared to corresponding data in the literature. Chapter 7 provides information about the key experiment, where a proprietary model of emotions is introduced, and experimental results and their analysis, outcomes of various methods including Self-Organizing Maps, and ANNs utilizing fuzzified input are presented. In addition, a visualization tool, utilized for designing an interface, enabling us to read the results of the listening tests and automatic mood recognition in an intuitive way, is described. Chapter 8 presents conclusions, summarizes the main findings, and discusses perspectives of further research.

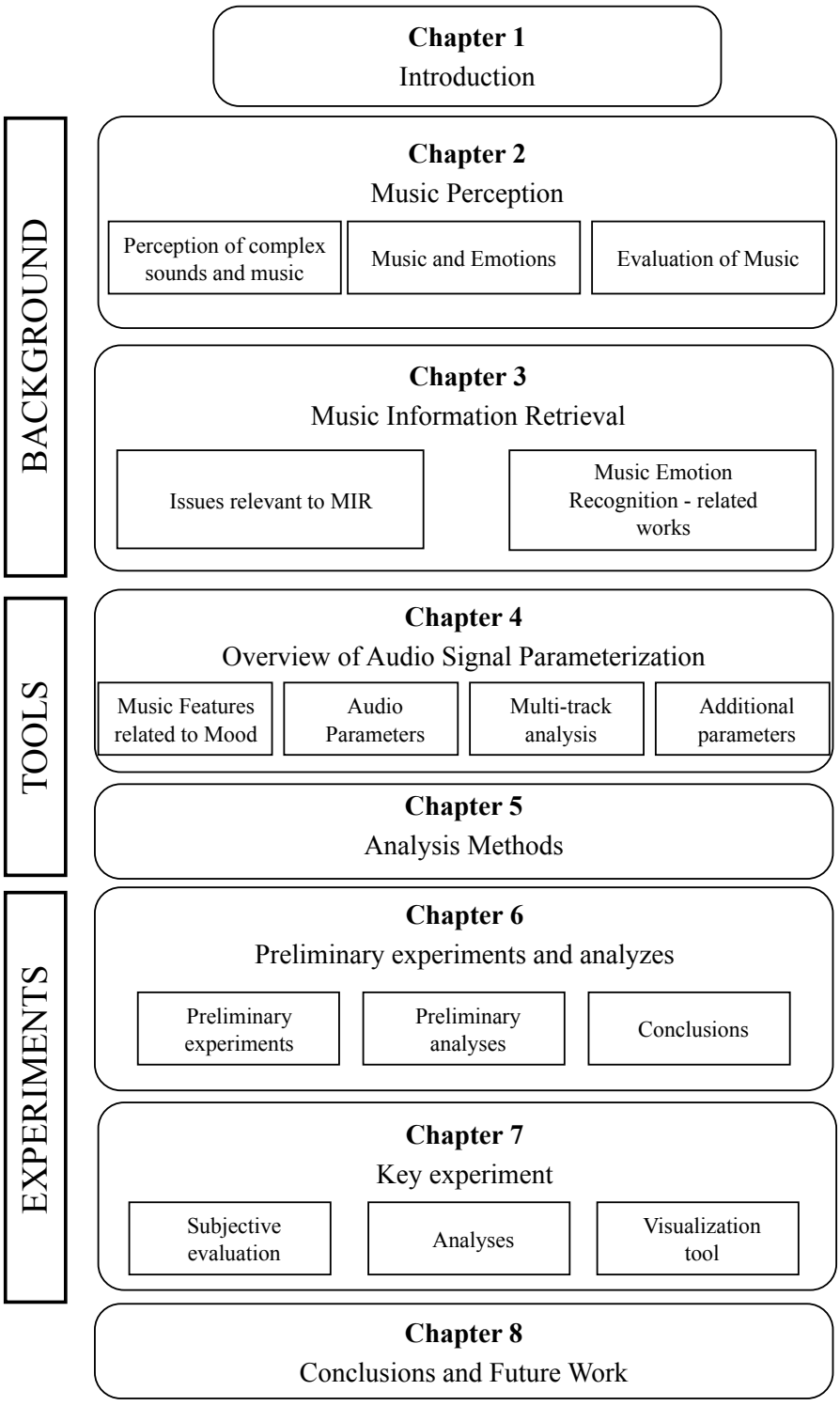



*Figure 1.2* *Organization of the thesis. Chapters are presented along with their content*

# 2 OUTLINE OF MUSIC PERCEPTION

Paraphrasing Zimmerman's statement: "Music is not just sound" [352] by saying that music is not sound *alone*, this reinforces the notion that music should be considered as art and that it has a strong relationship with culture. Moreover, as music expresses the composer's intentions to communicate an idea or feelings, it has an emotional impact on listeners, thus it should be interpreted personally. Those are some of the reasons that any generalization on music can be very misleading. It is a fact that music is a sequence of sounds organized in structures (even though these structures might be based on very irregular or even random order) and accordingly, the composer Edgard Varèse's famous quote says: "Music is organized sound." [266].

It is well known that music is an important medium of communication and provides a means by which people can share emotions, intentions, and meanings [21]. Schopenhauer formulates his opinion on music in the following way: "The inexpressible depth of music, so easy to understand and yet so inexplicable, is due to the fact that it reproduces all the emotions of our innermost being, but entirely without reality and remote from its pain... Music expresses only the quintessence of life and of its events, never these themselves."

As said before, there is no doubt that music expresses emotions intentionally, at least to some extent. As with any art, the listener can interpret music according to artists' intentions or in agreement with individual feelings. Whatever the effect, the influence of music is so strong that it is hard to disagree with Sacks [266]:

"Perhaps it is not just the nervous system, but music itself that has something very peculiar about it—its beat, its melodic contours, so different from those of speech, and its peculiarly direct connection to the emotions."

Music perception is not only strongly influenced by individual background and preferences, but also has deep roots in social and cultural trends.

There is not a single theory that explains all of the relationships mentioned above. In many studies researchers from field of psychology [240,266,290] and musicology [252-349] are trying to explain the marvel and magnificence of this phenomenon. Even though, there are obviously a lot of unfilled spots in our knowledge in human perception overall, it is worth trying to name and describe at least the main principles. In this chapter, some rules of music perception and especially emotion in music are referred to. Then, the methods of

subjective evaluation of music are described, with particular focus on mood of music but not the emotions of the listener. A wider discussion on mood of music and emotions evoked by music is performed in Section 2.4.

## 2.1 PERCEPTION OF COMPLEX SOUNDS AND SEQUENCES OF AUDITORY EVENTS

Considering the attributes of sound there is a need to distinguish between physical characteristics of sound and attributes perceived by the listener. In the physical domain sound can be simply described by amplitude, frequency, temporal attributes and position in space. On the perception side there are essential descriptors such as loudness, pitch, temporal attributes, space [209]. To form a simplified rule, they correspond with physical features as follows:

- loudness - level/amplitude,
- pitch – frequency,
- temporal attributes – onset time, duration, ADSR envelope,
- space – direction, distance, reflections from the surrounding environment, width, depth,
- timbre – relation of amplitude/frequency/temporal atributes.

Nevertheless, these relationships are general in their description and they do not describe perception comprehensively in the case of a single stimulus. This work is not primarily devoted to psychoacoustics and perception of a single stimulus. More complex relations, i.e. isophonic curves, variation of pitch with level, etc., are discussed in this context in the literature of the subject [8,69,209], and they are not going to be recalled here. More focus is directed at the perception of more complex sounds, grouping mechanism, perception of structures and finally the reasons why some sounds are perceived as music.

Gestalt principles of perception [291] is one of the theories with a holistic approach to the auditory perception, applicable to both artistic and non-artistic objects [174]. It states that natural mechanisms of grouping sounds are motivated by “action” requirements, to support recognition of objects or situations. The Gestalt grouping mechanism underlies our ability to analyze a sound environment, thus different sounds mean different sources, similar sounds – the same source. Gestaltists described many of the factors that govern perceptual organization (Fig. 2.1). No single rule always works, but it appears that the rules

can generally be used together, in a coordinated and probably quite complex way, in order to arrive at a correct interpretation of sound. Most of these rules were originated in visual perception but apply to both: vision and hearing [209]. Gestalt principles of perception are presented in Fig. 2.1 and described in details in the next paragraphs.

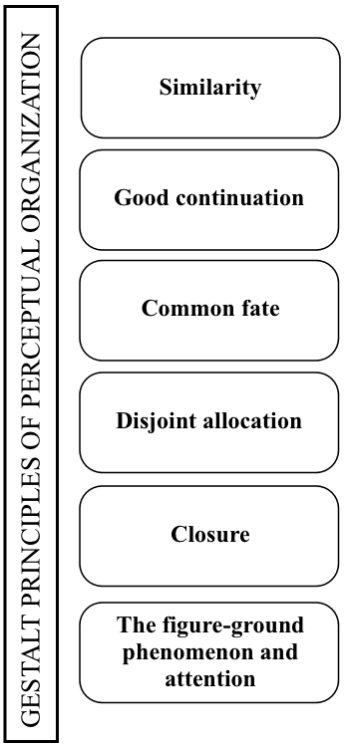



*Figure 2.1 Gestalt principles of perceptual ogranization [209].*

**Similarity**

This rule states that elements are grouped if they are similar. Simple sounds, events with the same frequency or a smoothly changing frequency are recognized as the same source. For more complex sounds timbre and spectral differences seem to be the most important factor. Although similarity does apply not to frequency and timbre only but also to loudness and subjective location, location is a less important grouping cue than pitch, which can easily be presented with "The scale illusion" [69]. The phenomenon occurs when different sounds are presented through headphones to subjects and they choose sounds from both speakers to construct a scale or another pattern. This method seems to be commonly used by composers, i.e. Bach "Well-Tempered Klavier, Prelude 8, Book 2", Tschaykowsky "Sixth Symphony, last movement", who create the illusion of patterns using sound from different sources and registers. The environment noises, reflections and

reverberation can cause confusion with localization and that is one of the reasons why this cue is less important.

**Good continuation**

This principle follows a physical property of sound sources; changes of all characteristics of sound (frequency, intensity, location, or spectrum) tend to be smooth and continuous rather than sudden. Thus, a smooth change in any of these aspects indicates that a signal comes from a single source, whereas a sudden change indicates that sound derives from another source. It also has a consequence that the larger structures override the immediate factors.

**Common fate**

Even if sounds coming from a single sound source differ in the frequency domain, they usually vary in a highly coherent way. They have a tendency to start and finish together and they change in intensity and frequency together. These phenomena constitute the basis of the principle of the common fate: if the same kinds of changes of two or more components are observed at the same time, then they are grouped and perceived as a part of the same source.

**Disjoint Allocation**

This rule states that a single element in sound can only be assigned to one source at a time. For certain types of stimuli, the perceptual organization may be ambiguous, since there might be more than one way to interpret the input. When a given element might belong to one of the number of streams, the observation may alter depending on the stream within which that component is included. In other words, if an element is close to the sequence it is absorbed by this sequence and it is perceived as a part of it.

**Closure**

This rule is based on adaptation of the auditory system to situations when a given source may be temporarily masked by other sounds. The masked sound tends to be perceived as continuous. Two sounds very close in time but separated in frequency are very unlikely to come from the same source.

**Attention and figure-ground phenomenon**

It seems that the auditory system is not capable of attending to every aspect of the auditory input. Certain parts or aspects of sound are selected for conscious analysis. Complex sound is divided into streams, and the auditory system focuses primarily on one

stream at a time. A stream that stands out perceptually is attended, while the rest of the sound is less prominent and discarded. This separation into attended and less important streams is called by Gestaltists "figure-ground phenomenon". It is worth noticing that a decision of direct attention towards one stream over a number of them, does not depend only on information available in the acoustic signal. Other sources of information or knowledge, such as i.e. interest in a particular conversation, etc., may also be involved. This process activates quite complex high-level auditory processing.

There are a few more contributions that are not constituted as Gestalt principles of perceptual organization but are important for the auditory perception and are present in Gestalt works [174,209]. Tendency to favor symmetry and analogy helps to constitute streams. This is one of the cues for the brain to interpret particular sequences as melodies. The main basis of grouping is pitch grouping. It is supported by other high-level auditory processing mechanisms such as memory and pattern recognition.

With regard to music perception Gestaltists wondered how it is possible that a melody (a set of specific pitches) could retain its recognizability, even when all of its pitches are changed, scores are played with a different instrument and in a different tempo [174]. They were interested in the problem of configurations, that is, how it is that elements come together to form wholes; how objects that are qualitatively different from the sum of their parts, and cannot be understood in terms of their parts, remain still recognizable despite major changes. Even though they did not come up with a satisfying answer, their statements contribute to our understanding of how objects in the visual and auditory world are organized.

Albert Bregman analyzed the auditory system from the perspective of information used to separate objects [42]. He defines ‘source’ as a physical entity that produces sound waves. ‘Stream’ is the perception of a group of elements that occur simultaneously or/and successively, and that is considered as a coherent whole, and is perceived by being produced by one source. It is a very unlikely situation that we perceive sound that is produced by only one source, thus the auditory system is capable of distinguishing between sources and separating frequency components to appropriate sources. These mechanisms are strongly related to masking, which is a phenomenon that has enormous influence on music composing, arrangement, mixing and production process [139,140]. Bregman stated that the auditory system is based mostly on simultaneous onsets. Simultaneous events are

being grouped and this factor - more relevant than pitch and time - is the significant factor in auditory grouping. Timbre, spatial location, and amplitude are also important cues for object separation. Bregman observed that one instrument playing notes with different amplitude is creating different streams. At the same time, one instrument playing notes far from each other can be perceived as different streams. J. S. Bach used these phenomena to create polyphony even using single instruments. It is easy to observe notes from different registers perceived as parallel melodies in organ Toccata and Fugue in D minor. In Cello Suites musicians are trying to emphasize the single voice polyphony created with characteristic articulation [183] after a Notebook for Ann Magdalena Bach [16]. Even though amplitude and frequency create two different and quite independent mechanisms, they both interfere in conscious and unconscious perception.

## 2.2 MUSIC PERCEPTION

As quoted at the beginning of this Chapter, music is being considered as an organized sound. At the same time Sloboda [291] brings to attention another, more illusory aspect of music experience: "The principal end-product of music listening activity is a series of fleeting, mental images, feelings, memories and anticipations.". He underlines that the attention to music is the most important factor while listening. During any kind of analysis it is important to remember that composers write for listeners, not for analysts and researchers. Listeners "see the bigger picture" and analyze the structures of music unconsciously. In music, a single sound is not isolated and it stands in significant relation to the others. Repeated sequences become more and more familiar so are easier to recognize [174]. Contour memory supports recognition of melody even if it is transposed, or played in different tempo.

Attention in music listening is partially influenced by the state of the listener at the moment of listening. Still the question of whether the subject went through musical training or is a novice is more important. It is especially apparent when polyphony and different layers of musical compositions occur [309]. Factors such as frequency, temporal and informational masking, play an enormous role in this process. Listeners also tend to proceed in a particular mode of listening, i.e. either analytical, where they de-construct the signal into components, or synthetical, where relations between elements are more

important. Musical training supports memory, which causes easier recognition of patterns, intervals, analogies, and variations [174].

The important part of musical training is hard work and has two significant aspects. One of the important parts of music training is learning directly and automatically from the perception. Its implication is that the knowledge learnt is the knowledge to perceive better [338]. The most mysterious and complex part is linking physiological responding and perception within our brains. Perception process is still not fully understood but some of the features that determine music reception are discovered. Music as all more complex auditory signals is organized by the brain into structures. This process depends on the timbre, pitch, and temporal attributes of the components [40,112].

In music perception studies many different classifications and systems that describe music components exist. Levitin [174] observes that from the listener's perspective there are seven major elements of music: loudness, pitch, melody, harmony, rhythm, tempo, and meter. It is worth mentioning that these features apply to music but not to single, extracted sounds. At the same time, [47] indicates slightly different elements of a musical piece – melody, rhythm, agogics, articulation, and dynamics. According to the second approach, it may be estimated that melody together with rhythm carry 90% of musical information [47]. Jones [124] listed only four main elements: harmony, melody, rhythm and tempo. On the other hand Peretz [240] mentioned tonal structure, melody, rhythm and articulation. Friberg [92] shows the following features as relevant for music mood analysis: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality and rhythm. Other common features, not included in that list, are, for example, mode or a musical form [168].

Elements proposed in the literature and cited above are listed and organized into groups as shown in Fig. 2.2. This ordination is used in further paragraphs to describe in detail the main components of music. Even though particular elements are assigned to specific groups, they all interact with each other. For example agogics is part of dynamics but also influences timing. Melody is closely related to phrase, but articulation to timbre. These are only some of the relations and they differ from performance to performance. Hence the assumption can be made that all elements are complementary and not exclusive.

Even though music notation is widely described in the literature of the subject, the main concepts are recalled here. It is relevant to the topic because some of the notation elements are used directly or provide basis for features used in MIR such as tempo (BPM - Beats Per

Minute), rhythmic patterns, chromagram and many others. The relationships between formal notation and representations used in MIR such as spectrogram are shown to explain the connection between these systems. Different music cultures use different music systems. Because the presented dissertation concerns Western Music, systematics and notation included in subsequent paragraphs are related to this music tradition.

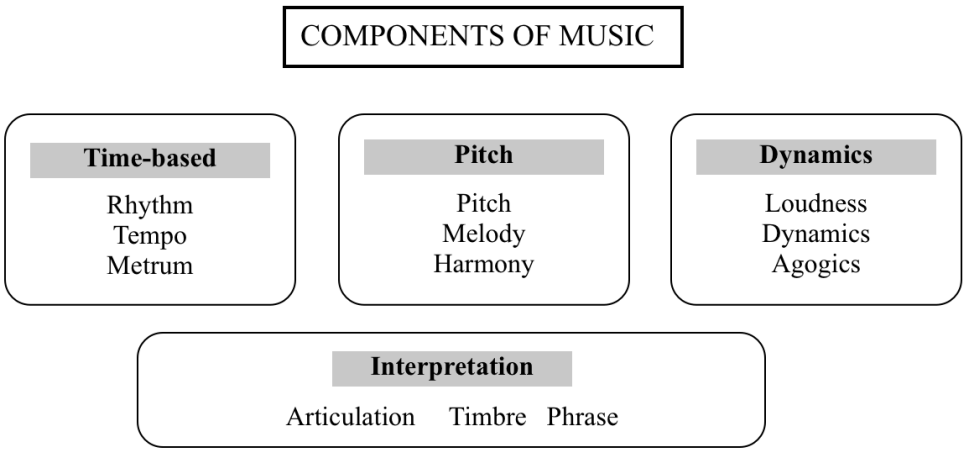



*Figure 2.2 Components of music compiled from various works [124,47,174,240]*

### 2.2.1 Time-Based Components

The ability of humans to perceive metric **rhythm** was a subject of investigation of many researchers. It is said that the way the brain processes rhythmic information depends on musical education. Some experimental results indicate that musicians process rhythmic patterns hierarchically whereas music non-professionals accomplish this task in a sequential manner [59]. In a natural environment musical sequences are mostly regular. Sturges [300] assumed that a perceptual mechanism is 'designed' to use rhythmic features of sound events. It allows prediction of upcoming events and sequences. Time organization of musical sequences conforms to particular characteristic orders and depends on pauses, accents, previous sounds, etc. [209]. Rhythmic music is based on beats, which can physically be characterized as distinct energy bursts in time. Pulse clarity estimates, on a large time scale, how clearly the underlying pulsation in music is perceived and can therefore be regarded as a measure for the underlying periodicity of music [165].

In music **rhythm** is a part of the melody structure, and **tempo** is the speed controller of the melody structure, therefore **rhythm** and **tempo** are deeply connected. Scholes [276] defined **rhythm** as the order and the proportion of durations. Pistone [241] described

**tempo** as the characteristic of motion execution with respect to measures as well as melodic, harmonic, rhythmic, or dynamic cues.

Structure is perceived as "**rhythm**" when a periodical sequence of time events can be discerned as segments, i.e. certain events must be recognized as starting points of periods [305].

Listeners have a preference for a certain range of tempi, typically centered around 120 BPM (Beats Per Minute) [206]. However it is clear that, when dealing with music, not every piece is always perceived in **tempo** of approximately 120 BPM. BPM units are described in the Section dedicated to notation.

**Tempo and rhythm detection**

Automatic estimation of the temporal structure of music, such as musical beat, tempo, **rhythm**, and **meter**, is not only essential for the computational modeling of music understanding but also useful for MIR. Temporal properties estimated from a musical piece can be used for content-based querying and retrieval, automatic classification, automatic drum accompaniment [158,337] music recommendation, and playlist generation [76,283]. If the **tempo** of a musical piece can be estimated, for example, it is easy to find musical pieces having a similar **tempo** without using any metadata. The difficulty of beat tracking depends on how explicitly the beat structure is expressed in the target music: it depends on temporal properties such as **tempo** changes and deviations, rhythmic complexity, and the presence of drum sounds.

Research on tempo and beat tracking was conducted in various fields of interest [86,272]. The task of computational **rhythm** retrieval is complex and it consists of a few stages. The simplified approach to this task may be reduced to retrieving the sequence of onset times and/or durations of sounds from the musical data – this process is called quantization. In another approach, the time signature is retrieved on the basis of musical content; in this class of methods usually the period of time is found, which divides the stream of sounds into repeating fragments. This task may additionally be combined with phenomenal accent retrieval in such a way that the phase of phenomenal accentuations in a piece is found. If the accentuations found line up with locations where humans tap the foot to the melody, it may be concluded that the bar lines are found – the rhythmic level of the size equal to the meter is thus acquired. The next complication is to retrieve metric **rhythm**, i.e. the hierarchic structure of related rhythmic levels. Existing metric **rhythm** research

usually focuses on retrieving low rhythmic levels – usually to the level of a bar - those methods are sufficient to emulate human perception of the local **rhythm**. High-level perception is required for drum players, thus the computational approach needs to retrieve the hypermetric structure of a piece. If it reaches high rhythmic levels such as phrases, sentences and periods, automatic drum accompaniment applications can be developed.

**Rhythm** is also often an element of a piece determining musical style, which may be valuable in music retrieval. The rhythmic structure together with patterns retrieved carry information about the genre of a piece, thus both are highly correlated.

**Rhythm** is described using rhythmic notes. The position of the note indicates the onset of the music event and the length of time that a note is played is called **note duration**, which is determined by the type of note (Fig. 2.3). Additional marks such as rests, dots and others are used to completely describe the time sequence of sounds.

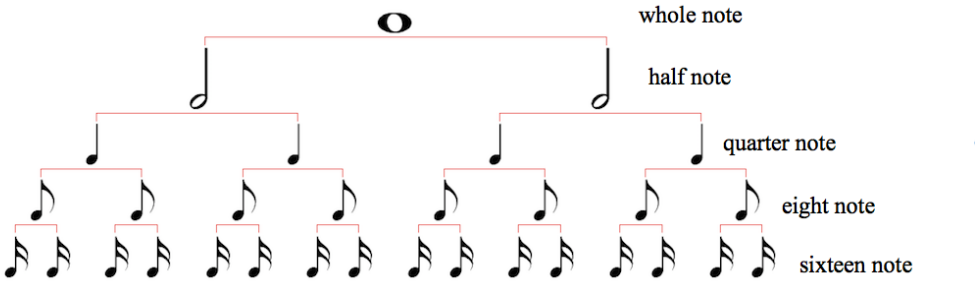



*Figure 2.3* *Types of note indicate the duration time. Basic note types are presented above: whole note, half note, quarter note, eight note and sixteen note*

**Meter**

Scholes [276] defined the **meter** of music as its general structure. It refers to the patterns of accents heard in regularly recurring measures of stressed and unstressed beats at the frequency of the music pulse.

Each time signature can be classified into a certain **meter**. Meter is described by a pair of numbers. The lower number indicates the units (according to the note duration shown in Fig. 2.3) and the higher refers to the number of beats in a measure. A measure (bar) is a segment of time corresponding to a specific number of beats and is indicated by vertical bar lines (Fig. 2.4). Meters are divided into duple, triple, quadruple and odd meters. They indicate not only the "capacity" of a bar but also a beat that should be followed by the performer. Examples of notation of different meters with beat suggestions are presented in Fig. 2.4. An example of rhythm notation along with a corresponding spectrogram is shown in Fig. 2.5.

### Tempo

Tempo in music is described in Beats Per Minute (BPM) units. It indicates how many quarter notes should occur within a minute. It is commonly shown as in Fig. 2.6. In the tradition of classical music, tempo is determined by Italian tempo markings. A list of tempo markings with approximate BPM is shown in Tab. 2.1. In this notation some of the tempo descriptors are dedicated to particular types of music pieces. They also include additional information about the character of a music piece (i.e. Marcia moderato is dedicated to marches). The spectrograms of the same rhythm performed in tempo 120 BPM and 240 BPM are presented in Fig. 2.7. A simple conclusion is that a piece performed twice as fast, lasts two times shorter. But it also has other consequences such as a change in the decay time of particular notes and interaction with reverb (if applicable).

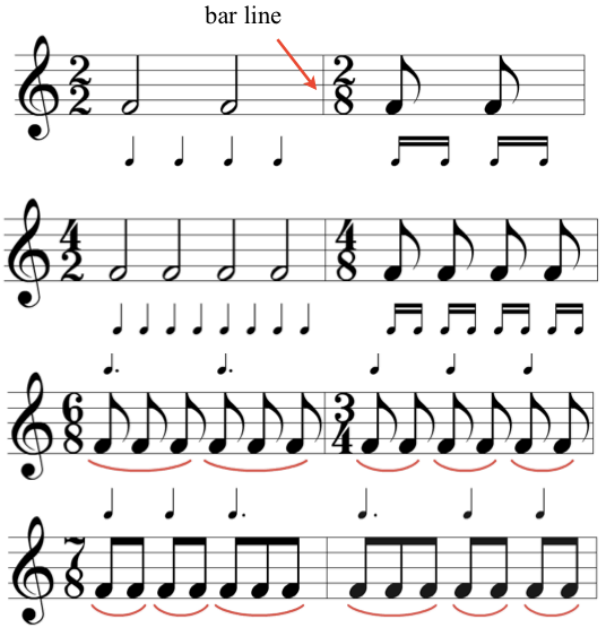



*Figure 2.4 An example of notation of different duple, triple, quadruple and odd music meters along with the grouping interpretation. Smaller notes indicate the beat suggested for the performance*

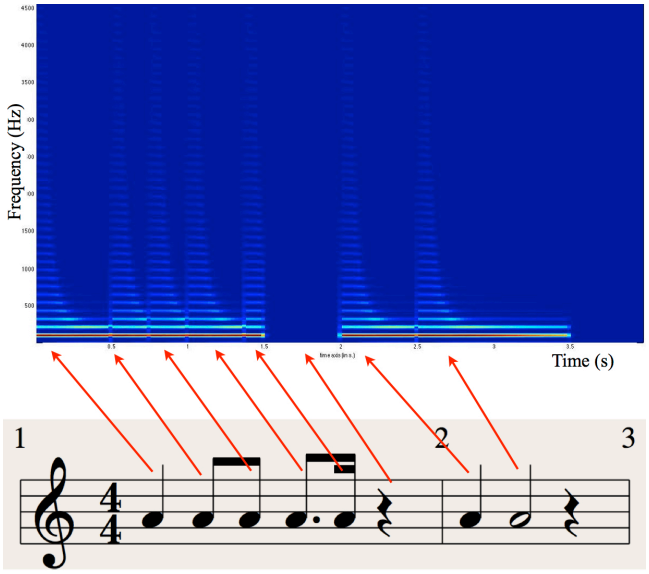



*Figure 2.5* *An example of rhythm notation with the corresponding spectrogram*

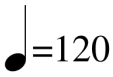



*Figure 2.6* *Symbolic annotation of tempo 120 BPM*

*Table 2.1* *Music tempo from slowest to fastest*

| **Tempo** | **Approximate BPM** |
|---|---|
| Larghissimo | 24 |
| Grave | 25–45 |
| Largo | 40–60 |
| Lento | 45-60 |
| Larghetto | 60-66 |
| Adagio | 66-76 |
| Adagietto | 72-76 |
| Andante | 76-108 |
| Andantino | 80-108 |
| Marcia moderato | 83-85 |
| Andante moderato | 92-112 |
| Moderato | 108-120 |
| Allegretto | 112-120 |
| Allegro moderato | 116-120 |
| Allegro | 120-168 |
| Vivace | 168-176 |
| Vivacissimo | 172-176 |
| Allegrissimo | 172-176 |
| Presto | 168-200 |
| Prestissimo | >200 |

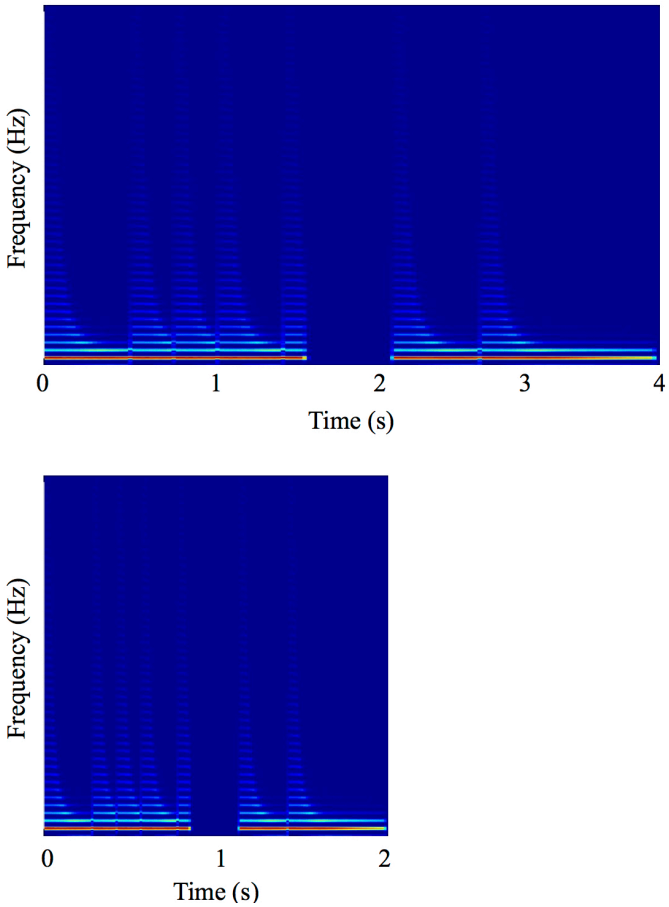



*Figure 2.7 Spectrograms of the same rhythm performed in tempo 120 BPM and 240 BPM*

### 2.2.2 Pitch Components

**Pitch** is a psychological concept, related both to the actual frequency of a particular tone and to its relative position in the musical scale. It functions only as a subjective sensation and is strongly related to perception. Pitch is related to the frequency or rate of vibration of a string, column of air, or other physical source. It is also influenced by the duration time, level and other factors. These dependences are widely examined and described in the psychoacoustic literature Different research works exist in which authors aimed to determine the relationship between the pitch of a pure tone and its frequency [8,207]. In the presented work, the debate related to pitch is focused on the musical context, including i.e. Bark scale. Stevens *et al*. [299] proposed mel-scale. Because mel-scale is commonly used in MIR, it is described in more detail.

Mel-scale is a scale of pitches acquired from listeners' judgments, where pitches have to be equal in distance from one another. The reference point is perceptual pitch of 1000 Mels

to a 1000 Hz tone (for level 40 dB above the listener's threshold). The relation between the Mel and Hertz scales is presented in Fig. 2.8

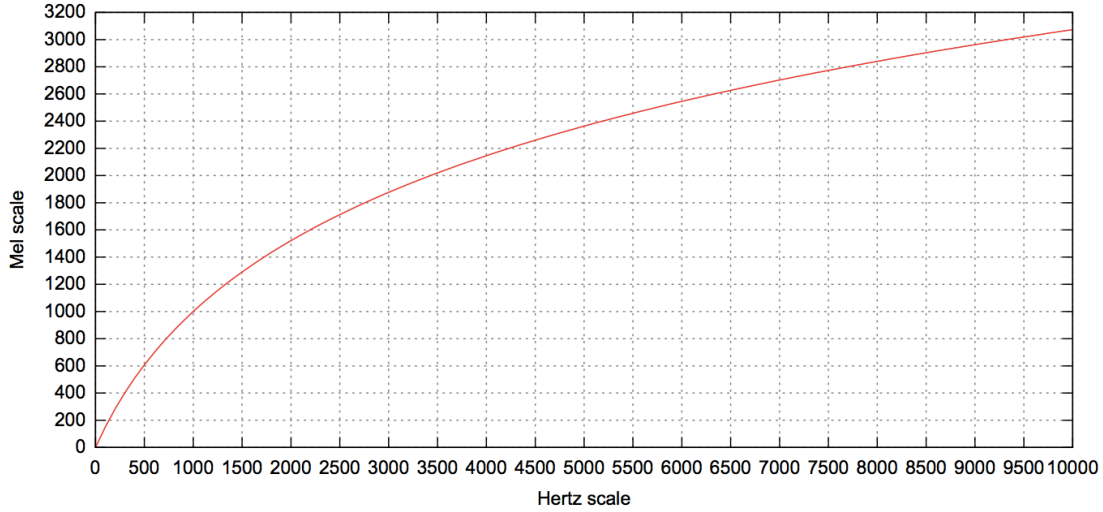



*Figure 2.8 Relation between Hertz and Mel pitch scales [323]*

The relation between the Hertz and Mel scales is given by the following formula [229]:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (2.1)$$

where $f$ is frequency in Hz.

In music, pitch, in most cases, is related to the fundamental frequency of sound. Western Music is based on the chromatic scale, where the frequency range is divided into semitones, which are the smallest defined distances between two sounds on the scale. Further discussion is based on equal temperament tuning [46], which is the most common tuning in popular music. Different systems such as Pythagorean tuning, meantone temperament, well temperament and others are described in the literature of the subject [46,284].

Each semitone (halftone) consists of 100 cents, and 12 semitones build an octave. The frequency of two frequencies a semitone apart is calculated as follows:

$$f_2 = f_1 \sqrt[12]{2} \qquad (2.2)$$

The steps of a music scale in equal temperament tuning along with its representation on the piano keyboard (which is the most common representation of the scale in Western Music) and ranges of a few musical sources are presented in Fig. 2.9. Sounds with a smaller frequency are called low and close to the top of the scale (bigger frequency) - high. All these assumptions are valid only in Western Music. In different music cultures not only scales and

relations between scale steps are defined differently, but even the terms high and low are culturally relative - the Greeks talked about sounds in the opposite way because the stringed instruments they built tended to be oriented vertically [46,161].

A musical **scale** is a subset of the theoretically infinite number of pitches, and every culture selects these based on historical tradition. The most common scales in Western Music are major and minor (Fig. 2.10). Sounds in between are considered mistakes unless they are used intentionally for expression or interpretation. Western music theory recognizes three minor scales and each has a slightly different flavor. Also other scales are commonly used in particular music genres or styles. Blues music generally uses a five-note (pentatonic) scale that is a subset of the minor scale, and Chinese music uses a different pentatonic scale [174].

Each note of a scale has a special name, called a scale degree. A basic scale with scale degrees description is shown in Fig. 2.11.

**Melody** (musical line) is an organized sequence of music pitches. Contour describes the overall shape of a melody that the listener perceives as a single entity. Dowling [70] and others have shown that contour is the most salient musical feature for infants. Peretz [240] discovered that the right hemisphere of the brain contains a contour processor that in effect draws an outline of a melody and analyzes it for later recognition, and this is dissociable from rhythm and meter circuits in the brain.

Melodies can consist of one or more musical phrases or motifs that are important cues for the interpretation [174]. Melody can be performed as a single line (monophony), along with other melodies (polyphony) or as a part of harmony.

Relation between consecutive notes of the melody is described by intervals, which consist of an integer number of semitones [46]. The list of intervals within an octave is presented in Tab. 2.2.

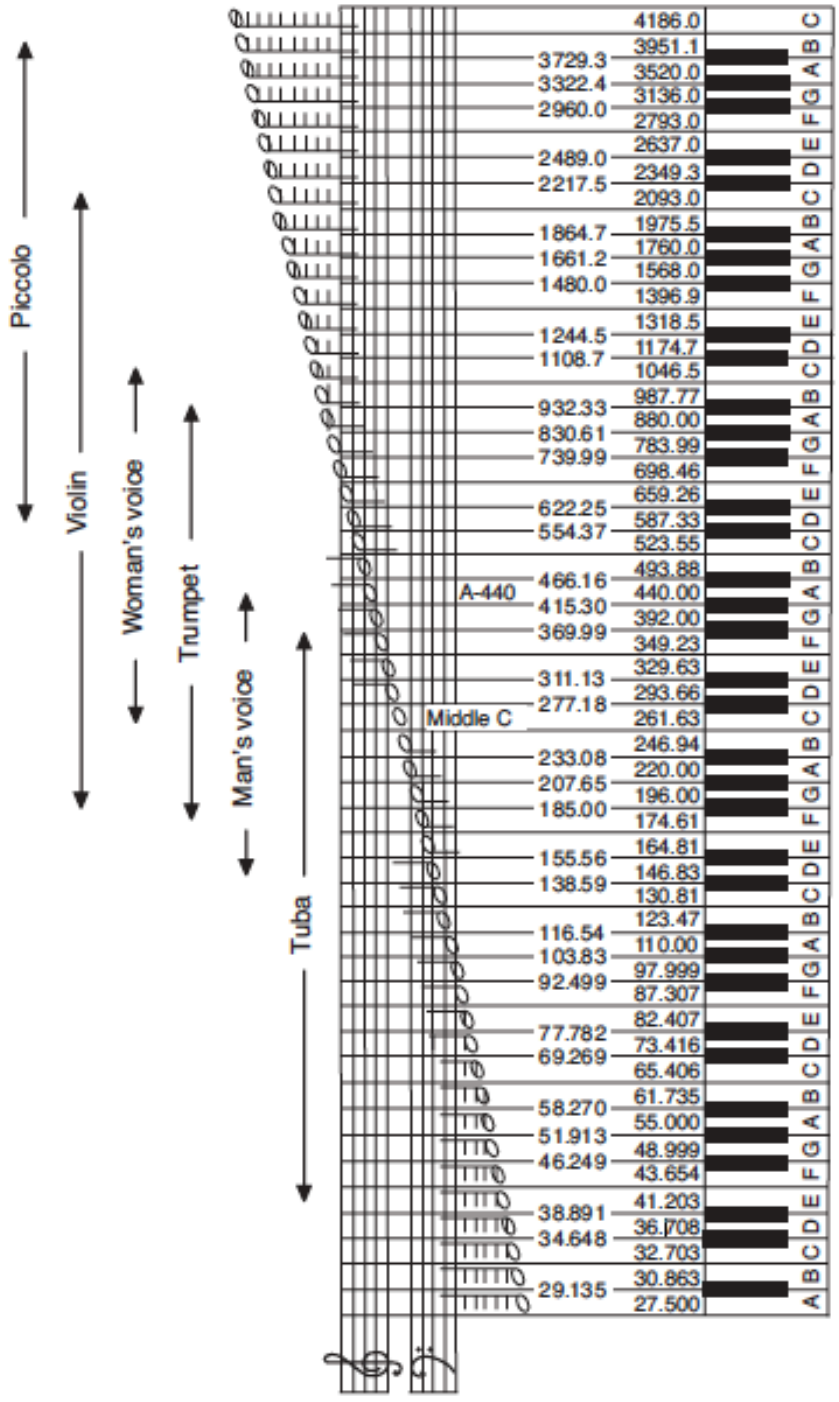



*Figure 2.9* *Music scale used in Western Music. Pitches are presented along with corresponding piano keys and the frequency range of a few common music sources [174]*

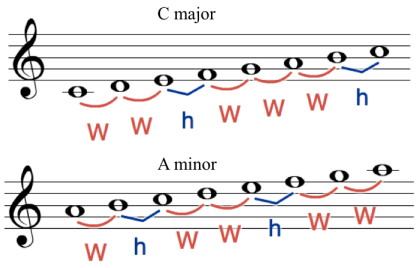



*Figure 2.10* *Major and minor music scales, "w" indicates a distance of a whole tone (2 halftones) and "h" a halftone*

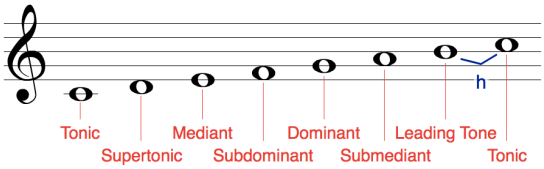



*Figure 2.11* *A music scale with scale degrees description*

*Table 2.2* *The list of intervals used in Western Music along with corresponding distance in semitones*

| Distance in semitones | Interval name |
|---|---|
| 0 | unison |
| 1 | minor second |
| 2 | major second |
| 3 | minor third |
| 4 | major third |
| 5 | perfect fourth |
| 6 | tritone |
| 7 | perfect fifth |
| 8 | minor sixth |
| 9 | major sixth |
| 10 | minor seventh |
| 11 | major seventh |
| 12 | octave |

An example of a **melody** given as scores along with the spectrogram is presented in Fig. 2.12.

Monophony is the simplest of textures, consisting of a single melody without accompanying lines or harmony. This may be realized as one note at a time, or with the same note duplicated at the octave. Within the context of the Western musical tradition, monophony is usually used to describe the music of the late Middle Ages and Renaissance. Polyphony is a texture consisting of two or more simultaneous lines of independent melody. Baroque forms such as the fugue, which are great examples of polyphony, are usually

described as counterpoint because of the special rules regarding the composition of additional lines [68].

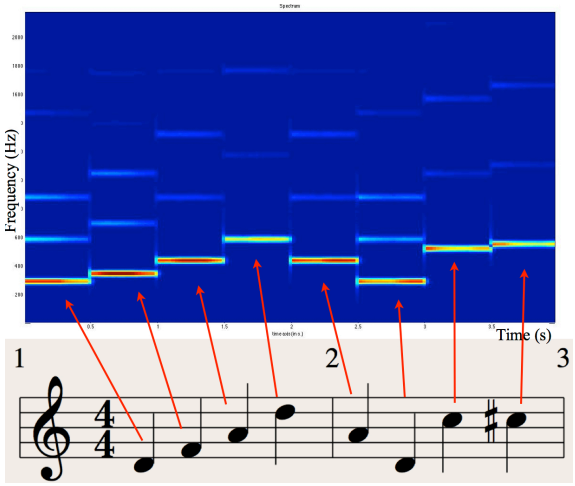



*Figure 2.12 Spectrogram and score notation of an exemplary melody.*

Polyphony and counterpoint are at the root of harmony in Western Music. Harmony is based on relationships between the pitches of different tones. **Harmony** in music is the use of simultaneous sounds or subsequent notes that are perceived as a figure. Tonal contexts set up by pitches lead to expectations for what will come next in a musical piece. Harmony refers to a parallel melody to the primary one or to a chord progression (the clusters of notes). The **harmony** is a system that involves chords and their construction and chord progressions and the principles of connections between them. **Harmony** is also strongly related to scales (major, minor and variations) as well as to music **key**, which often defines the scale of a music piece and is based on the scale steps presented before (Fig. 2.11). A sequence of basic chords (major, minor and seventh chords) with equivalent spectrogram is presented in Fig. 2.13. There are several parameters describing harmonic content of music signal, i.e. chroma, key consonant, dissonant, harmonic strangeness, chroma eccentricity [24]. They are briefly described in Chapter 4, as they are used in music parametrization.

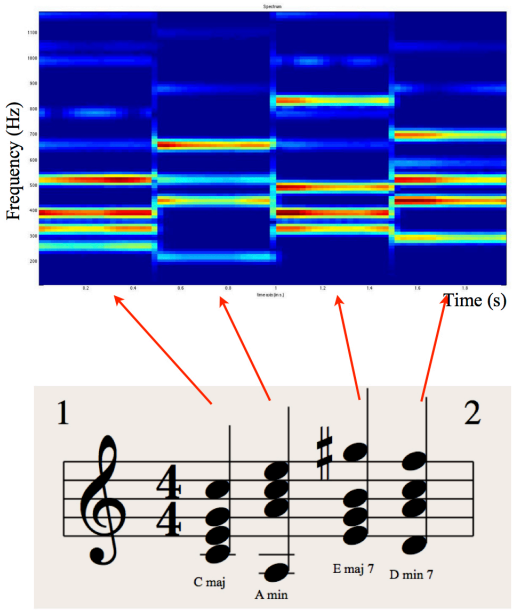



*Figure 2.13 A sequence of chords: C major, A minor, E major 7 and D minor 7 with equivalent spectrogram.*

### 2.2.3 Dynamic Components

**Loudness** is a purely psychological construct that relates to the physical amplitude of sound, which is represented by Sound Pressure Level (SPL). It is commonly linked with Root Mean Square (RMS [dB]) although this simplification does not include non-linear attributes of hearing. Therefore different loudness scales (i.e. phons, sones and others) were introduced to include the knowledge of auditory perception. The relationship between the sones and phons scales is presented in Fig. 2.14 [209].

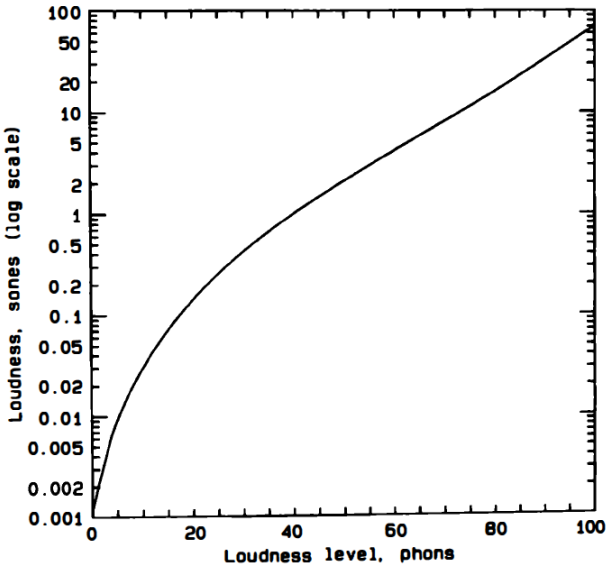



*Figure 2.14 The relationship between loudness in sones and loudness level in phons for a 1 kHz sinusoid [209]*

Other scales were defined for new technologies, broadcasting and sound engineering purposes. To name a few of them: LU (Loudness Units), LUFS (Loudness Units relative to Full Scale), LKFS (Loudness, K-weighted, relative to Full Scale) [263,306] and many others.

On the other hand, in music notation loudness (sometimes called dynamics) is described by written or printed musical notation. Dynamics are relative and do not refer to specific volume levels. The list of most common dynamic indications (from softest to loudest) is included in Tab. 2.3.

*Table 2.3 The list of common dynamic indications from softest to loudest*

| Dynamic | Abbreviation | Translation |
|---|---|---|
| pianississimo | *ppp* | very very soft |
| piano pianissimo | *pp* | very soft |
| piano | *p* | soft |
| mezzo-piano | *mp* | moderately soft |
| mezzo-forte | *mf* | moderately loud |
| forte | *f* | loud |
| fortissimo | *ff* | very loud |
| fortississimo | *fff* | very very loud |

Another cue for dynamics of music performance is **agogics**, which is defined as an element that determines the dynamics of the music piece. It was introduced by Riemann [262] and is strongly connected with tempo and interpretation. Accents that are emphasis placed on particular notes are also an important part of dynamics. An example of accent marks are shown in Fig. 2.15.

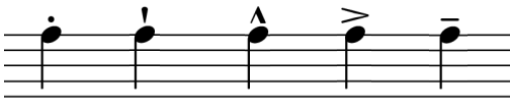


*Figure 2.15 Exemplary accent marks. From left to right: staccato, staccatissimo, marellato, marcato and tenuto*

Music notation contains many other descriptors or suggestions that strongly influence dynamics, pitch or tempo and also more sophisticated elements of performance. Additional, commonly used terms such as legato, portato, terms related to bow technique, glissando and many others are part of regular music education and can easily be found in the literature [13,37,130].

### 2.2.4 Interpretation

The performance practice is based not only on the written cues given by a composer but also on the music tradition or school and the individual interpretation. Some of the composer's suggestions are not strictly defined and leave space for interpretation (i.e. dynamics, performance of accents, tempo). Interpretation is the name of the process where the performer is deciding how to perform music that has previously been composed. What is more important, interpretations of different musicians performing the same music can vary widely. Variation includes quantitative values such as tempo, number of repetitions [13] and other sonic saddle elements such as **timbre** (soft, harsh, delicate, etc.), drive of the rhythm, climax of the **phrase**, **articulation** and many others [141,334].

Apart from the technical performance point of view, the most important part of interpretation is the emotions and mood included in music and the performance. The world famous cellist, Pablo Casals said: "Don't give notes. Give the meaning of the notes." [141]. All of the technical descriptions in this section are just tools to raise the meaning and emotion from music. Interpretation and emotional content of music is a topic of continuous discussion between composers, musicians, researchers and finally listeners. But according to the famous saying: **"Writing about music is like dancing about architecture"**. Therefore we have to just agree that some of the secrets of the magic of music will stay in the domain and sound will never be possible to explain verbally.

### 2.2.5 Other Cues for Music Perception

Music perception depends on many factors. Some of them are social or cultural while others are highly individual and rely on personal preferences, music and sound training, detection thresholds and many other conditions. A few examples of dissimilarities in music perception are presented in next paragraphs.

In Western Music tradition listeners tend to associate major scales with happy or triumphant emotions, and minor scales with sad or defeated emotions [174]. This has deep roots in Medieval sacral music, where particular scales and chords were used to emphasize particular feeling or emotions. At the same time, in Arabic music various scales containing microtones are dedicated to particular types of music pieces (i.e. for funerals, weddings, salvation, etc.). In both cases scales are used to obtain a similar result but the means, here a set of pitches, are very different.

Social factors include the popularity of a particular form of music (i.e. motet in Renaissance or piano concert in classicism) [326], appurtenance to a certain subculture that is connected to a particular music genre, and to the accessibility of music resources. These conditions can affect either positive or negative associations with a music piece, regardless of the music content.

The personal aspect of music perception is even more complex. It is affected by social and cultural background, individual associations with particular situations or context, completed music and sound related training, familiarity with a certain kind of music, the situation of the music listening and many others. Due to the vastness of the problem, only a few selected factors are to be discussed here.

The important part of musical training is perceptual learning. One category of perceptual learning, discussed by Eleanor Gibson [95] is differentiation – an ability improved by exposure to stimuli. Gibson analyzed these aspects from the perspective of the sound engineer. To distinguish and control numerable parameters such as single instruments sound, space, tuning, effects and many others, the sound engineer must learn to differentiate features in music unperceivable to others.

Castro and Lima [51] investigated how age and musical experience influence emotion recognition in music. They concluded that years of music training are correlated with recognition accuracy. This is coherent with the fact that learning music can emphasize the perception of structures underlying musical emotions. Castro and Lima showed that age and musical expertise slightly affect how listeners recognize emotions in music. Therefore these findings add to the set factors that indicate individual differences in emotion processing.

Identifying performance mistakes is strongly related to completed music training. Pluta [248] showed that intonation discrimination is highly dependent on musical training and can be improved by dedicated exercises. Trained musicians, composers and especially conductors can follow lines of the whole symphonic orchestra and detect a mistake of a single instrument. Primarily the auditory system focuses on a single stream but thanks to exercises people involved in music and sound can distinguish very small changes in multiple aspects of sound at once.

Cultural differences also strongly influence the perception of mood of music. In numerous studies the mood of music from different cultures was examined i.e. English and

Chinese songs [115,344], Greek music [146], American music [66]. Also different cultural groups were asked to evaluate the same music [66,115] and the cultural dependence of perceived mood was studied.

Some recipients of art prefer one field and master the perception craft by a long-term exploration. Also people exist who are more sensitive to music than other forms of art. A single piece of music can affect listeners in various ways depending on their character, earlier experiences and their preference. While some people cry listening to *Symphony of Sorrowful Songs* by Górecki, others find it boring and not inducing any feelings or meaning. At the same time, Rick Ross (hip-hop artist and producer) has fans who find his music interesting and entertaining and opponents who perceive it as meaningless and aggressive.

All of the factors discussed above are related to music. On the other hand, various psychological studies show differences in emotion processing [60,103,143]. This is another condition that can cause other significant differences in personal perception of the mood of music. Also listening conditions are sometimes considered as an important part of the perception of the mood of music. Watson and Mandryk [331] attempted to model musical mood from audio features supplemented by listening context.

These few examples show to what extent music perception is subjective and dependent on many factors and preferences. That said, everyone agrees that music carries essence and meaning, helps relaxing, support simple activities (i.e. work, jogging), and finally expresses emotions [174]. Emotional content is one of the main reasons why music is present in cultural life for ages [21,35].

## 2.3 MUSIC AND EMOTIONS

It is difficult to think of music without including a prominent role of the emotions. Most of psychological research has been focused on the listener's emotions induced by music. It is important to distinguish between the emotions of the listener and the mood or emotions of the music. This distinction is crucial because it determines the approach and methodology of the research [179]. At the same time Song and collaborators [295] examined the relationship between perceived and induced emotional responses to music. Analysis of their results did not reveal significant differences in ratings between perceived emotion and induced emotion. Moreover, the results indicated that, regardless of the

discrete type of emotion experienced, listeners' ratings of perceived and induced emotion were highly positively correlated.

In the presented dissertation, the author is focused on the content of musical signal, that is the mood of music. Nevertheless, origins of models of emotions and terminology are in the music cognition studies; therefore major points of this field are discussed.

In psychology studies Sloboda was one of the first researchers who introduced 'music cognition' as a field of science in the 1980s [290]. He showed later that 'cognition' and 'emotion' might be connected more strongly than expected [291]. He showed that emotional responses to music require cognition. One of his questions was whether musical emotions can be predicted based on the context. Some scholars have noticed that musical emotions may be 'too subjective' to be predictable [97]. The answer to that was the statement: "emotions to music can never be predicted from musical characteristics alone." [291]. Many survey studies were performed by Sloboda and his collaborators [290,291] [126,127] and they come to the conclusion that music emotions can be predicted to some extent from information about the context, supplemented by external factor descriptors. Fifteen predictors were featured in a discriminant analysis, five for each of the main factors (i.e., listener, music, situation). The analysis focused on predicting three common emotion categories in a representative sample – happy-elated, sad-melancholic, and nostalgic-longing. Results showed that emotions could be predicted with an accuracy of 70%. This outcome suggests that music emotions are not too subjective to be modeled. However, the prediction is not perfect, even though this analysis included many predictors not only in the music but also related to the listener, and the situation [126,127]. Studies on emotions evoked by music are limited because of important issues about the definition and measurement of emotion. "Emotions are relatively brief, intense, and rapidly changing reactions to potentially important events (subjective challenges or opportunities) in the external or internal environment - often of a social nature - which involve a number of subcomponents (cognitive changes, subjective feelings, expressive behavior, and action tendencies) that are more or less 'synchronized' during an emotional episode." [127].

Different theories recognize various factors that influence emotions induced by music, e.g. [126,278,279,349]. The most comprehensive attempt to delineate various mechanisms that underlie musical emotions is the BRECVEM model proposed by Juslin and Sloboda [127]. They propose seven mechanisms (besides cognitive appraisal) through which music

might induce emotions: Brain stem reflexes, Rhythmic entrainment, Evaluative conditioning, Contagion, Visual imagery, Episodic memory, and Musical expectancy. These seven psychological mechanisms include information from various domains, mostly outside music.

## 2.4 MOOD OF MUSIC

The relation between music and emotions has been the subject of some studies in the last century, where researchers tried to identify the influence of certain musical attributes such as tempo, mode, rhythm and others on the human subjects perception. This content-based approach leads to the field where mood and emotions are examined as components of the musical signal. In order to distinguish, in this context, emotional content of music, it is called **mood of music**. Yang and Chen [343] distinguish "emotions expressed by music" from "emotions induced by music". In the literature terms "**music emotions**" and "**mood of music**" are used alternatively. **For the purpose of the presented dissertation both of them describe mood and emotions included in music, as opposed to music cognition studies, where "emotions to music" are a listener's personal experience of feelings evoked by music.** This distinction might seem to be unnecessary, but it is introduced to clarify the nature of studies that involve only musical content without measuring additional external factors. In this Section, particular components and attributes of music are examined with respect to the mood of music. As stated in Section 2.2, discussion related to music content is based on Western Music tradition.

It is not always obvious what the composer meant. Contrarily, in the case of opera, such notions are clear according to the music history, opera tradition and libretto. In other situations, there is a lot of space for interpretation.

Huron [117] points out that since the preeminent functions of music are social and psychological, the most useful characterization would be based on four types of information: genre, emotion, style, and similarity.

[43] recognized the following music features that influence the mood of music: harmony, tempo, loudness, timbre, rhythm diversity. Levitin [174] discussed the relationship between particular elements of music and mood of music. One of the important means of expression is timbre. After Levitin [174]: "Composers use timbre as a compositional tool [...] to express particular emotions, and to convey a sense of atmosphere

or mood. Scriabin, Ravel but also Steve Wonder, Paul Simon - Music as painting, melodies are equivalent of shape and form, and the timbre is equivalent to use of color and shading".

Scherer and Oshinsky [279] showed that faster tempos were positively associated with happiness and negatively associated with sadness. Husain et al. [118] found that a mode manipulation affected mood, but not arousal, whereas a tempo manipulation affected arousal, but not mood (for models of mood see Section 2.5). Moreover, for arousal, a significant interaction between mode and tempo revealed that the impact of tempo was stronger for the major mode than it was for the minor mode. A similar interaction was also observed for ratings of musical enjoyment, such that tempo had more effect on enjoyment ratings in the major mode than it did in the minor mode. Observations of Levitin [174] are coherent with these conclusions. In his opinion, tempo is a major factor in conveying emotion. Songs with fast tempos tend to be regarded as happy, and songs with slow tempos as sad. The impact of rhythmic content is often underestimated. As an example, syncopation is a very important concept that relates to expectation, and ultimately to the emotional impact of a song. The syncopation catches us by surprise, and adds excitement. Moreover, it is widely known that listeners tend to associate major scales with happy or triumphant emotions, and minor scales with sad or defeated emotions [174]. Also very tiny changes in loudness and dynamics have a profound effect on the emotional communication of music.

Much research has focused on emotional response to a single element. Webster and Weir [332] have explored the interactive effects of mode, texture and tempo in a single experiment. Research on the interactions between musical elements is important because most music is a complex, often dynamic combination of musical elements. The results of the experiment point in the direction of complex sets of rules, where combinations of characteristics entail a particular mood of music (Fig. 2.16).

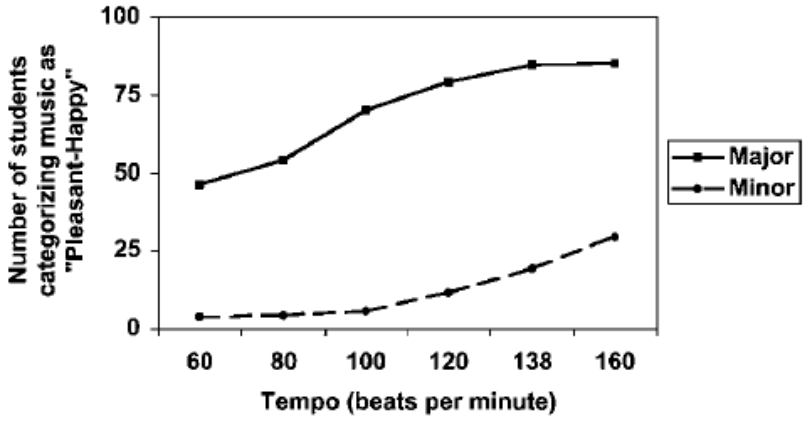



*Figure 2.16* *Number of participants (out of 83) categorizing five musical phrases as "Pleasant-Happy" as a function of mode and tempo [332]*

## 2.5 MOOD MODELS

Music Emotion Recognition (MER) research has not yet determined any main or right model of the music mood. Contrarily, a great variety of mood models are constantly being devised in different psychological studies. Numerous studies on mood recognition reveal different findings and conclusions [168,288,335,343]. Most of the models brought to mood description can be assigned to one of the following two approaches: dimensional or cluster description.

### 2.5.1 Dimensional Approach

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 simple human perception of mood or emotions.

Thayer created a two-dimensional model Valence/Arousal [308]. Axes divided the plane into quarters (Fig. 2.17), which correspond to the following moods: **contentment** (low arousal, high valence), **depression** (low arousal, low valence), **anxious/frantic** (high arousal, low valence) and **exuberance** (low arousal, high valence).

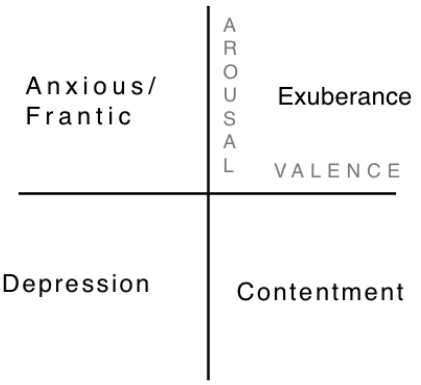



Figure 2.17 *Mood representation in Thayer's model* [*308*].

Russel constructed a system based on a two-dimensional Valence/Arousal model [264] (Fig. 2.18), in which 28 expressions are placed on the mood plane. A schematic representation of these models is shown in Fig. 2.18.

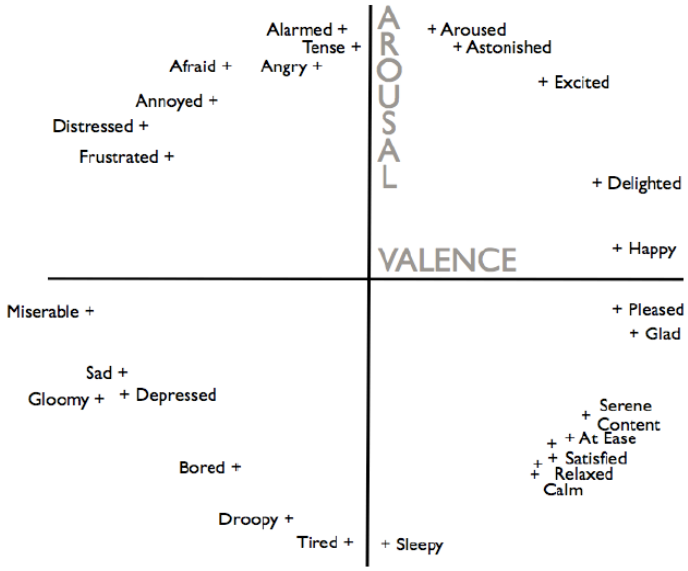



*Figure 2.18 Russell's model of music mood presented on Valence/Arousal plane [264].*

Russel's conception is contrary to conclusions presented in MIREX (Music Information Retrieval Evaluation eXchange) music classification competition [114], where authors show that a song can be simultaneously described with more than one mood tag group.

Russel's circular structure is one of many possible dimensional representations. Tellegen-Watson-Clarck model [307] enhances Thayer's system by adding a second set of axes rotated 45°. Extra axes represent engagement and pleasantness (Fig. 2.19).

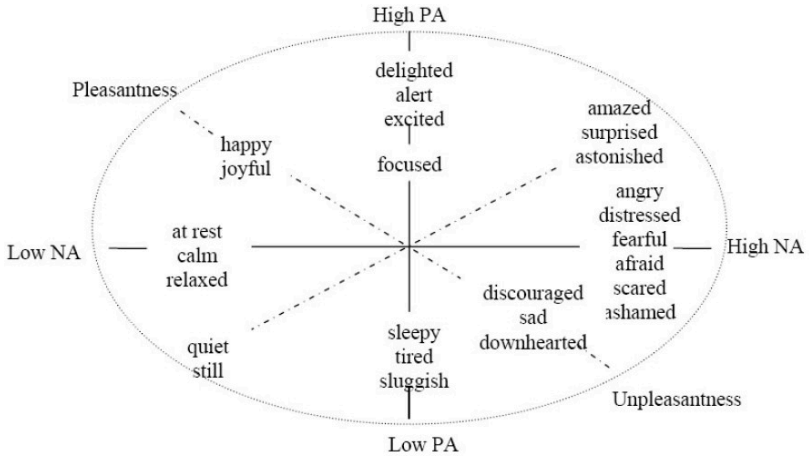



*Figure 2.19 Tellegen-Watson-Clarck mood model [307].*

### 2.5.2 Categorical Approach

Categorical approach is based on categories, groups or clusters, to which particular music pieces are assigned. In specific research the number and meaning of categories are different. A few of them, commonly used in MER, are discussed below.

With the objective to evaluate several algorithms within the same system, MIREX [71,114] organized five clusters of mutually exclusive categories.

Clusters of mood adjectives used in the MIREX Audio Mood Classification task are listed below:

**Cluster 1** passionate, rousing, confident, boisterous, rowdy

**Cluster 2** rollicking, cheerful, fun, sweet, amiable/good natured

**Cluster 3** literate, poignant, wistful, bittersweet, autumnal, brooding

**Cluster 4** humorous, silly, campy, quirky, whimsical, witty, wry

**Cluster 5** aggressive, fiery, tense/anxious, intense, volatile, visceral

One may name two other very common models. The first is Hevner's model, which proposes a list of 67 adjectives grouped into eight mood clusters (Fig. 2.20) [108]. Hevner's adjective checklist was rearranged into ten groups by Farnsworth [86]. The second model is Schubert's [280], in which 46 affective adjectives are arranged into nine clusters according to their position on the two-dimensional Thayer's model (Tab. 2.4).

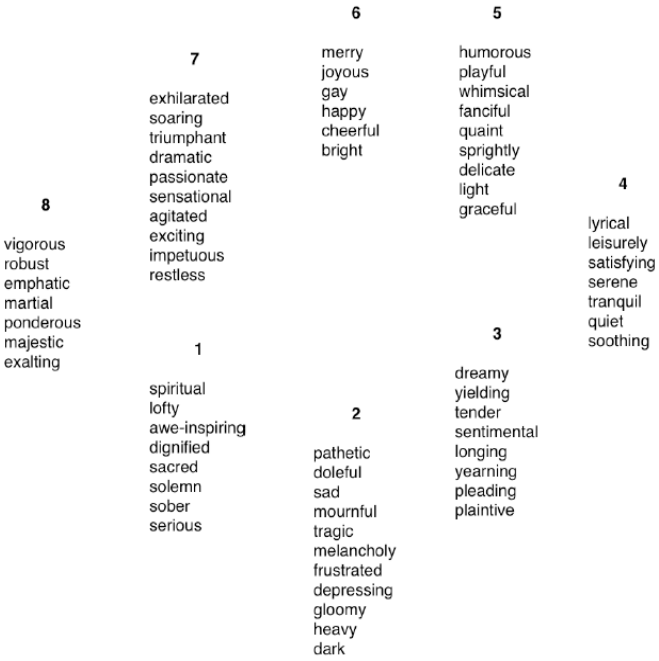



*Figure 2.20 Hevner's model with 67 adjectives grouped into eight clusters [108]*

*Table 2.4 The Nine Emotion Clusters Proposed by E. Schubert in 2003 [280]*

| **Cluster No.** | **Emotions in Each Cluster** |
|---|---|
| 1 | Bright, cheerful, happy, joyous |
| 2 | Humorous, light, lyrical, merry, playful |
| 3 | Calm, delicate, graceful, quiet, relaxed, serene, soothing, tender, tranquil |
| 4 | Dreamy, sentimental |
| 5 | Dark, depressing, gloomy, melancholy, mournful, sad, solemn |
| 6 | Heavy, majestic, sacred, serious, spiritual, vigorous |
| 7 | Tragic, yearning |
| 8 | Agitated, angry, restless, tense |
| 9 | Dramatic, exciting, exhilarated, passionate, sensational, soaring, triumphant |

Xu and Wunsch showed that MIREX clusters might not be appropriate due to some semantic overlap between categories [341]. Moreover, they have shown that both Hevner's and MIREX representations have advantages and limitations when evaluated in a semantic mood space. The authors also found that basic emotions: happy, sad, angry and tender, are very relevant to social networks. Laurier *et al*. proposed a folskonomy representation with four clusters each containing 15 adjectives (Tab. 2.5). The adjectives are very similar to the categories proposed in the main emotion theories [308]. The theory is strongly related to

the two-dimensional model of Russell's concept. Clusters represent the four quadrants of the classical Valence-Arousal representation:

- Cluster 1: angry (high arousal, low valence)
- Cluster 2: sad, depressing (low valence, low arousal)
- Cluster 3: tender, calm (high valence, low arousal)
- Cluster 4: happy (high arousal, high valence)

*Table 2.5 Clusters of mood tags proposed by Laurier et al. [168].*

| **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
|---|---|---|---|
| angry | sad | tender | happy |
| aggressive | bittersweet | soothing | joyous |
| visceral | sentimental | sleepy | bright |
| rousing | tragic | tranquil | cheerful |
| intense | depressing | good natured | happiness |
| confident | sadness | quiet | humorous |
| anger | spooky | calm | gay |
| exciting | gloomy | serene | amiable |
| martial | sweet | relax | merry |
| tense | mysterious | dreamy | rollicking |
| anxious | mournful | delicate | campy |
| passionate | poignant | longing | light |
| quirky | lyrical | spiritual | silly |
| wry | miserable | wistful | boisterous |
| fiery | yearning | relaxed | fun |

The semantic space created by Laurier *et al.* is relevant and coherent with the existing basic emotion systems.

Another method that does not directly belong to approaches mentioned before is a common and well-established test method in emotion-related research. It involves a pictorial rating (Fig. 2.21) system called SAM (Self-Assessment Manikin) [120], which was designed to allow intercultural research or research with children. In some interpretations SAM method is based on the Thayer's model [179] and pictogram ratings are directly re-calculated into positions on the AV plane.

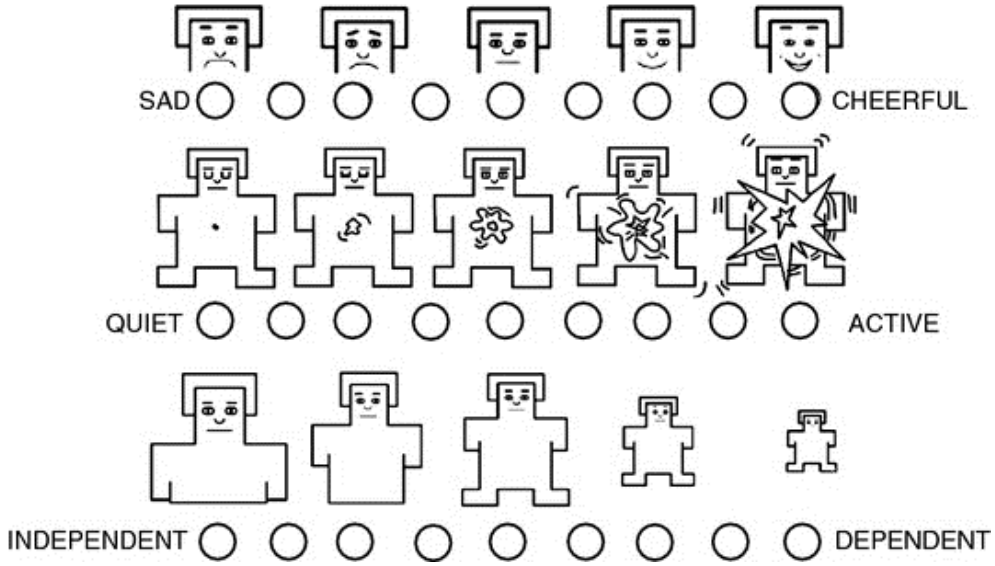



*Figure 2.21 Emotion evaluation system SAM based on pictorial ratings [179]*

The main advantage of the categorical approach is that it is intuitive and easy to use. It also seems to be relatively simple to work with the clusters approach, as the task is "limited" to define the classifications rules. Unfortunately, it is very hard to choose the exact number of clusters. Too many clusters are difficult to use and might be misleading because of a semantic overlap between categories. Using only a few categories may not be sufficient to describe such complicated notions as mood and emotions and may impede classification decisions.

Zenter and collaborators performed a large-scale research study to compile a list of music-relevant emotion terms and to study the frequency of both felt and perceived emotions [349]. On this basis they proposed a 9-factorial model of music-induced emotions. A distinctive feature of their solution is that 40 items are organized to create 9 first-order and 3 second-order factors. Geneva Emotional Music Scale (GEMS), as they called it, was proposed as a measurement tool for musically induced emotions.

Choosing a proper mood representation is the first step towards the correct automation of mood recognition. All of the earlier mentioned research was realized in English. There is a question, however, whether the proposed clusters should be straightforwardly translated into other languages or rather carefully analyzed, assessed, and then rearranged.

Presented mood models are based on different philosophical theories. However they do not take into consideration that mood perception is subjective and people experience emotions in many different ways. Sotiropoulos *et al*. proposed an excellent solution called MUSIPER, which uses objective audio signal features to model the individual perception.

Neural network-based learning allows the determining of a subset of audio features for the specific user. This idea seems to be an answer to reservations related to subjective perception of music and emotions.

## 2.6 SUBJECTIVE EVALUATION OF MUSIC

Music is multidimensional as well as an auditory perception. In order to study the nature of sound, it is possible to measure the physical characteristics of an audio signal in the acoustic or digital? domains. This characterization of the signal does imply how the auditory system will interpret and quantify it. Since the direct measurement of the perception is very complex and expensive [319], listeners are asked to quantify their experience [27]. The subjective evaluation of music is the situation where listeners are asked about their experience related to music perception. The choice of method strongly depends on the attribute of music and sound that is examined. In some fields, where standardization is required (i.e. lossy audio compression algorithms discussed further), particular test procedures have been defined. In ITU-R recommendation BS.1534-2 [122] the MUSHRA test (MUltiple Stimuli with Hidden Reference and Anchor) is introduced to obey the bias while comparing results of different tests. These recommendation defines the design of the experiment, selection of listening panels, test method, attributes, program material, reproduction devices, listening conditions, statistical analysis, presentation of the results of the statistical analyses and contents of test reports. As mentioned before, these tests are designed for the purpose of audio quality assessment, which is a very sensitive area due to the commercial aspect of the algorithms. Nevertheless, this area can be a valuable source of guidelines, especially for how external conditions such as the listening environment, duration of the test, order of the samples and others can affect the results.

Music evaluation is not standardized because almost each experiment varies in terms of the investigated aspect or the model of the examined attribute. Nevertheless, some methods (i.e. AB, ABX, method of adjustments, method of limits etc.) are commonly used [8,27,191,248]. However all these tools are dedicated mostly to the studies where detection of a small threshold is in request, or researchers are trying to determine whether listeners can hear the difference between audio samples in various aspects. Berenzweig and his collaborators [28] discussed different methods of music similarity measures. Various aspects of sound and music require different methods and interface, i.e. evaluation of hall

acoustics [93,191], pitch accuracy [248], localization of the phantom sound source [245]. In most cases, the interface is designed especially for the purpose of the experiments carried out (recently, most of them have been computer-based) and includes questions related to the examined attribute.

### 2.6.1 Subjective Evaluation of Mood of Music

In music emotions evaluation there is no standard established. A survey is a straightforward and very common technique for collecting information about emotional content in music [134]. Each research varies in terms of procedure, mood of music model, music set, interface and other details [202,281,288,311]. Most tests are conducted to evaluate the efficiency of automatic mood description. Other experiments are executed to determine the vocabulary [288] or choose the mood labels [134] or are carried out for other purposes. Skowronek *et al*. [288] proposed a method to obtain a reliable "ground truth" database for automatic music mood classification. The concept is based on a careful selection of musical excerpts and a broad search for proper mood labels. Due to a lack of standardized procedures, this seems to be a reasonable approach. However, it assumes only one model of emotions and set of labels, which disqualifies it as an universal tool. Nevertheless, guidelines related to music selection (different moods, different music genres, pre-judging) are reasonable and well motivated.

Selected listening tests performed within the area of MER are listed in Tab. 2.6 along with a short description of the testing procedure. This shows a background of the experiments executed and reported for the purpose of the presented dissertation.

Even though the above given experiments are based on different procedures, some rules are universal and should be maintained regardless of the test. A list of the main factors that should be considered while designing the listening tests related to mood of music is shown in Fig. 2.22. The chosen mood of music model determines the interface and the main concepts of the testing procedure (i.e. marking on the plane, choosing labels, etc.). A model should be user-friendly and sufficiently profound to cover all expected mood judgments. A group of listeners can affect the results, so they should be appropriate for the particular task, e.g. close to the potential end-users of the music recommendation system.

*Table 2.6 Details of selected listening tests related to mood of music*

| Author(s) | Title | Music set | Subjects | Used mood of music model |
|---|---|---|---|---|
| Schuller et. al. [281] | Determination of nonprototypical valence and arousal in popular music: Features and performances | 2648 pop songs | 4 listeners | 5 discrete labels |
| Trohidis et. al. [311] | Multilabel classification of music into emotion | 593 songs | 3 expert listeners | 6 emotion labels |
| Miller et. al. [202] | Last.fm in numbers | Last.fm | 20 milions/ month | 960,000 free-text tags |
| Jun et. al. [125] | A Fuzzy Inference-based Music Emotion Recognition System | 40 music samples (randomly chosen) | 5 listeners | Arousal/Valence space |
| Laurier et. al. [168] | Music mood representations from social tags | Last.fm | 30 milions | 107 selected tags |
| Skowronek et. al. [287] | A demonstrator for Automatic Music Mood Estimation | 1059 excerpts from 12 music genres | 12 subjects | 12 mood classes |

According to a well-known principle [27], a test should not exceed 15 min. Bachorik *et al*. concluded that most music listeners require about 8 seconds to judge the mood of a song [17], which determines the minimum sample duration time. The set of music should be diverse and should be presented in random order, different for each listener to avoid the bias of the previous sample. The playback system is not crucial but should be reasonable (i.e. headphones) to reproduce the whole frequency range of a musical signal. Respecting these guidelines allows for creating a reasonable test that should give meaningful results. Procedures of experiments conducted in the course of this dissertation are described in details in Chapter 6.

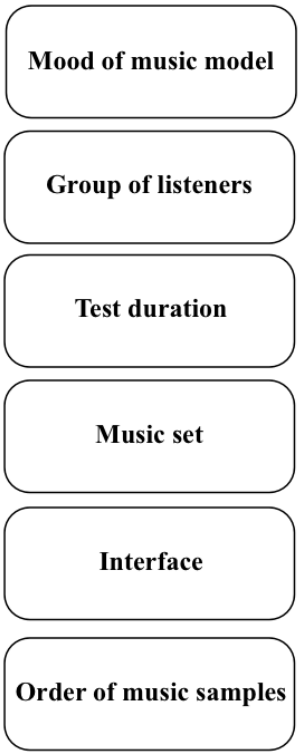



*Figure 2.22 List of main factors that should be considered while designing the listening tests related to mood of music*

# 3 MUSIC INFORMATION RETRIEVAL (MIR)

Music Information Retrieval is a multidisciplinary field of studies, which is related to musicology, psychology, music performance, signal processing, artificial intelligence methods and others. The main goal of MIR is to find information about music by engaging intelligent automated processing, automatic music description and evaluation of the proposed solutions [50]. MIR is highly involved in recommendation systems and many recent studies are dedicated to this topic [142,230,270,333]. Due to an enormous amount of music that is available online, a new approach based on more specific targeting is observed, rather than the continuous increase of accessible content only [57,198]. Different systems are based on various strategies from collaborative filtering [96,162,297], through metadata information [113,167,295] to a content-based approach [237,250,331].

MIR also provides solutions to many tasks such as melody extraction, instrument identification, genre classification and many others. The range of topics included in MIR is described in Section 3.1. In subsequent sections studies related to Music Emotion Recognition (MER) are described and discussed. Various approaches are presented and a comparison of strategies and results is performed. Then artificial intelligence methods employed for MIR and especially MER tasks are invoked.

## 3.1 ISSUES RELEVANT TO MUSIC INFORMATION RETRIEVAL

MIR is organized according to use cases, which determine the type of a query, the ,meaning of the match, and the form of the output. Questions and outcomes can be given in different forms, i.e.: textual information (metadata), musical excerpts, recordings, scores, or music features. The match can be exact or approximate depending on the type of the task. The specificity of questions defines the query type and the choice of the exact or approximate output. High specificity systems identify the exact content of individual recordings, i.e. identifying a particular version of the song, and low specificity systems focus on broad descriptions of music, such as genre. Specificity is divided into three categories: high-specificity systems match instances of audio signal content; mid specificity systems match high-level music features, such as melody, but do not match audio content; and low

specificity systems match global (statistical) properties of the query. Some of the MIR tasks and their specificities are listed in Tab. 3.1 [47].

Table 3.1 Examples of MIR tasks and their specificities

| Use Case | Specificity | Description |
|---|---|---|
| Music Identification | High | Identify a compact disk, provide meta-data about an unknown track |
| Plagiarism detection | High | Identify misattribution of musical performances, misappropriation of music intellectual property |
| Copyright monitoring | High | Monitor music broadcast for copyright infringement or royalty collection |
| Versions | High/Medium | Remixes, live vs. studio recordings, cover songs. Used for database normalization, and near-duplicate results elimination |
| Melody | High/Medium | Find works containing a melodic fragment |
| Identical Work/Title | Medium | Retrieve performances of the same opus number or song title |
| Performer | Medium | Find music by specific artist |
| Sounds like | Medium | Find music that sounds like a given recording |
| Performance Alignment | Medium | Mapping one performance onto another independent of tempo or repetition structure |
| Composer | Medium | Find works by one composer |
| Recommendation | Medium/Low | Find music that matches the user's personal profile |
| Mood | Low | Find music using emotional concepts |
| Style/Genre | Low | Find music within given music style or genre |
| Instrument(s) | Low | Find works with particular instrumentation |
| Music-Speech | Low | Radio broadcast segmentation, Music archives cataloguing |

In MIR, three strategies are most common for solving use cases, that is: metadata, high-level music content description, low-level audio features.

The first approach is based on metadata, which is information encoded and searched like text, and is well-matched to low specificity queries. The second strategy involves high-

level descriptions of music content corresponding to intuitive or expert knowledge about how a piece of music is constructed. The high-level descriptions approach is suited to mid specificity queries. The third method is based on low-level signal properties, which are used for all specificities.

Metadata is the basis of many MIR systems [121,162,297]. Many services exist to provide reliable metadata for existing collections of music, either for end users or for large commercial music collections such as Gracenote [98], Decibel [67] or MusicBrainz [218]. These services provide both factual metadata, namely objective truths about a track (artist, album, year of publication, track title, duration), and cultural metadata, which contains subjective concepts (i.e. mood, emotion, genre, style). The metadata for compressed and uncompressed digital music is often encoded in ID3 tag and can be included in formats such as: MP3, Ogg Vorbis, FLAC, MPC, Speex, WavPack TrueAudio, WAV, AIFF, MP4, and ASF. Casey [50] stated that metadata is the most common method of organizing. This approach is used by many music download services and has achieved a degree of success for them. However, there are disadvantages of this approach as it is extremely difficult to maintain consistent expressive metadata descriptions. Moreover, it is estimated that it takes about 20–30 minutes per track of one expert's time to enter the metadata [236], which incurs enormous costs. Therefore media services are beginning to open up to social exchange information about content to collect metadata and similarities in taste and preferences. Nevertheless, issues related to the consistency of the description remain valid. Metadata cannot solve the entirety of MIR due to consistency issues. Content-based methods are not being developed to replace but to enhance metadata-based methods.

In high-level music content description musical concepts such as melody, harmony, key, rhythmic patterns, tempo, meter and others are used to describe the content of the music. It is a very intuitive approach but extraction from polyphonic recordings, i.e., multiple instruments playing different lines simultaneously, remains difficult to achieve. It is difficult to extract melody, rhythm, timbre, harmony and other features from both audio and symbolic representations such as MIDI files. The goal of the tasks listed above is to encode music into a schema that conforms to traditional Western music concepts, that can then be used to make queries and search music. An automatic extraction of high-level music descriptions has been a subject of extensive research study within the Music Information

Retrieval Experimental eXchange (MIREX) [114]. Many recent works are focused on extracting high-level music features from low-level audio content [73,210,228,268].

The third strategy for content-based music description is to use information included in the digital audio signal. Within this approach, music is treated as any other signal but dedicated measures are defined to describe values that are relevant for the topic. Low-level audio features are measurements of audio signals that contain information about a musical work and music performance [50]. These descriptors contain not only the desired information due to the difficulty of precisely measuring just a single aspect of music. This approach engages the whole idea of signal parametrization and finding signal descriptors that contain information about a specific aspect of music. Issues related to parametrization and groups of parameters commonly employed in MIR and especially in MER are described in detail in Chapter 4.

## 3.2 MUSIC EMOTION RECOGNITION (MER)

Musicologists indicate a few elements of a musical piece – melody, rhythm, agogics, articulation, and dynamics - that are important in analysis and they form the foundations of music. Moreover, it may be said that melody together with rhythm carry 90% of musical informativeness. Rhythm is also an element of a piece determining musical style, which may be valuable in MIR. The rhythmic structure together with melody patterns retrieved from a music piece 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, perception of preference, mood or emotions. Huron [117] 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 in introducing 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" [50]. 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. Recently, music mood recognition becomes a thorough subject of research studies and analyses [134,168,194,188,234,324,330,331,339,349,]. This area of research studies is called **Music**

**Emotion Recognition (MER)** and aims at recognizing emotions contained in music signals [156,254-256,331,339].

MER is based on the basic definitions of perception. Lewis defined his term '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." [50]. The discovery of the relationship between the measurable content of the physical world and human perception seems to be a fundamental problem in MER. Various computational methods and algorithms have been implemented to organize and interpret the content derived from the audio signal for that purpose.

### 3.2.1 Models of Mood Used in MER

Music Information Retrieval or Music Emotion Recognition research studies have not determined any main or right model of music mood. Contrarily, a great variety of mood models are constantly being explored and devised in psychological and musicological studies. Most of them are based on the models from psychology or music perception that are described in detail in Section 2.5. Nevertheless, some representations that came from another background such as social networking or other studies also function in the research [22,43,182,242].

Selected models used in MER studies are listed and briefly described in Tab. 3.2. All of them belong to one of the approaches: dimensional or categorical. Discussion on the various models of mood is presented in Section 2.5.

In their MER studies, Brinker and collaborators [43] examined the relationship between Thayer's Valence/Arousal (VA) model and twelve mood labels (Tab. 3.3). They aimed at obtaining a model with clear moods covering the full range of emotional content. They observed that there is more consistency over subjects for arousal than for valence. As a result, they presented placement of particular moods in the VA plane and this is shown in Fig. 3.1. Finally, Brinker *et al*. [43] chose six mood categories for their MER research (Fig. 3.2).

*Table 3.2 Selected models of mood used in MER studies [20]*

| No. | Description | Approach | Ref. |
|---|---|---|---|
| 1 | Update of Hevner's adjective model (9 categories) | Categorical | [280] |
| 2 | 5 MIREX mood clusters (Passionate, Rollicking, Literate, Humorous, Aggressive) | Categorical | [33,65,114,312,328] |
| 3 | 5 emotions (Happy, Sad, Tender, Scary, Angry) | Categorical | [75,265] |
| 4 | 4 quadrants of the Valence/Arousal (VA) plane (Exuberance, Anxious, Depression, Contentment) | Categorical | [33,330] |
| 5 | 11 subdivisions of the Valence/Arousal plane (Pleased, Happy, Excited, Angry, Nervous, Bored, Sad, Sleepy, Peaceful, Relaxed, Calm) | Categorical | [104] |
| 6 | 12 clusters based on tags | Categorical | [182] |
| 7 | 72 tags from CAL500 dataset | Categorical | [22] |
| 8 | 8 subdivisions of the Valence/Arousal plane | Categorical | [134] |
| 9 | 4 basic emotions (Happy, Sad, Angry, Fearful) | Categorical | [321] |
| 10 | 9 affective dimensions from Asmus (Evil, Sensual, Potency, Humor, Pastoral, Longing, Depression, Sedative, Activity) | Dimensional | [15] |
| 11 | Arousal/Valence plane | Dimensional | [104] |
| 12 | 6 dimensions | Dimensional | [193] |
| 13 | 3 dimensions (Arousal, Valence, Tension) | Dimensional | [75] |

*Table 3.3 Twelve mood labels used in experiment of Brinker and his team [43]*

| Identifier | Music mood label |
|---|---|
| **A** | Sad |
| **B** | Calming/soothing |
| **C** | Arousing/awaking |
| **D** | Powerful/strong |
| **E** | Tender/soft |
| **F** | Cheerful/festive |
| **G** | Carefree/lighthearted/light/playful |
| **H** | Angry/furious/aggressive |
| **I** | Peaceful |
| **J** | Emotional/passionate/touching/moving |
| **K** | Loving/romantic |
| **L** | Restless/jittery/nervous |

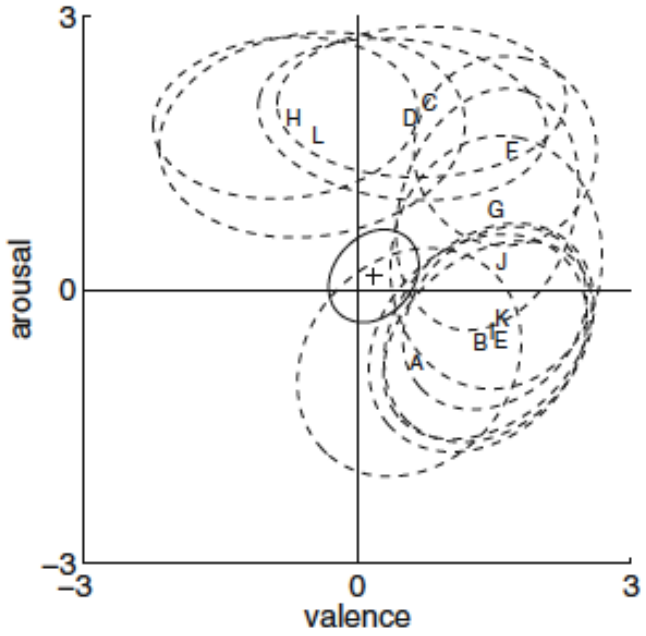



*Figure 3.1* *Locations of music moods in the VA plane, described according to the identifier labels listed in Tab. 3.3. Neutral mood category is indicated by a solid line [43]*

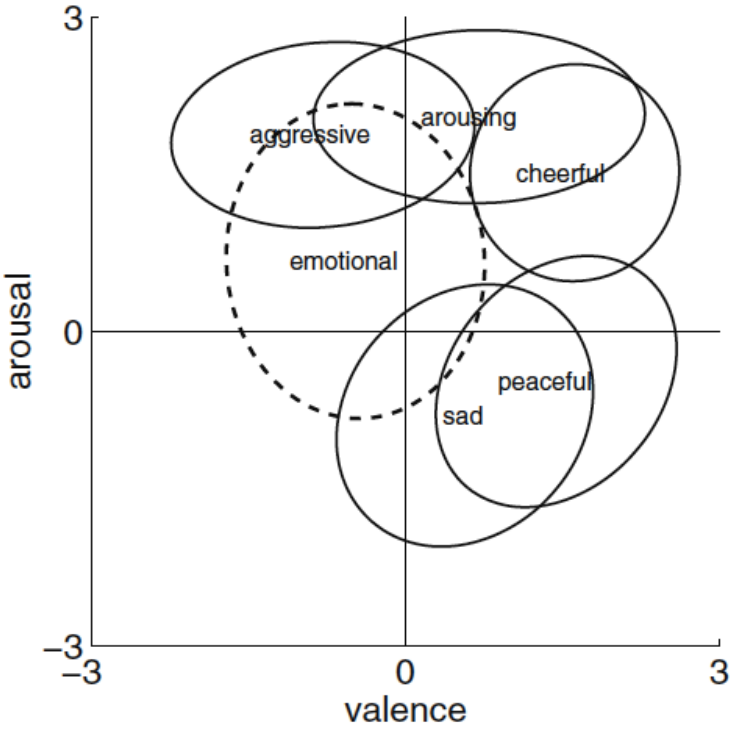



*Figure 3.2* *Five mood categories supplemented by one negated category ("emotional") selected by Brinker et al. [43]*

It is useful to remember that both subjective descriptors and features describing mood are multidimensional, thus another important issue is the these dimensions are presented.

Therefore some studies employing MDS (Multidimensional Scaling) analysis to determine dimensions underlying the music mood perception were performed. In some related work one can see that the basis for determining similarity between songs is often acquired from listening tests. Novello *et al*. [226] 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 (MDS) analysis and quadratic discriminant analysis to search for axes that maximize the separation of the excerpt classes [226]. 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 [310]. In the study of Trochidis and his collaborators 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. MDS was also used in the course of the research presented in this dissertation [243] and is described in Section 6.4.

### 3.2.2 Metadata-based Approach to MER

A considerable part of the research involved in MER is based on tags, semantics and lyrical content. Hu and Downie [113] investigated relationships between genre, artist and mood tags. Their results indicate that the genre-mood and artist-mood relationships are not stable enough to include them in further consideration. Laurier *et al.* [168] analyzed how people tag music by mood. They created a semantic mood space from *last.fm* tags using Latent Semantic Analysis. They performed SOM analysis and presented mood space as well as a tree diagram of the mood tags obtained with a hierarchical clustering approach. Simplified results of their research are presented in Figs. 3.3 and 3.4.

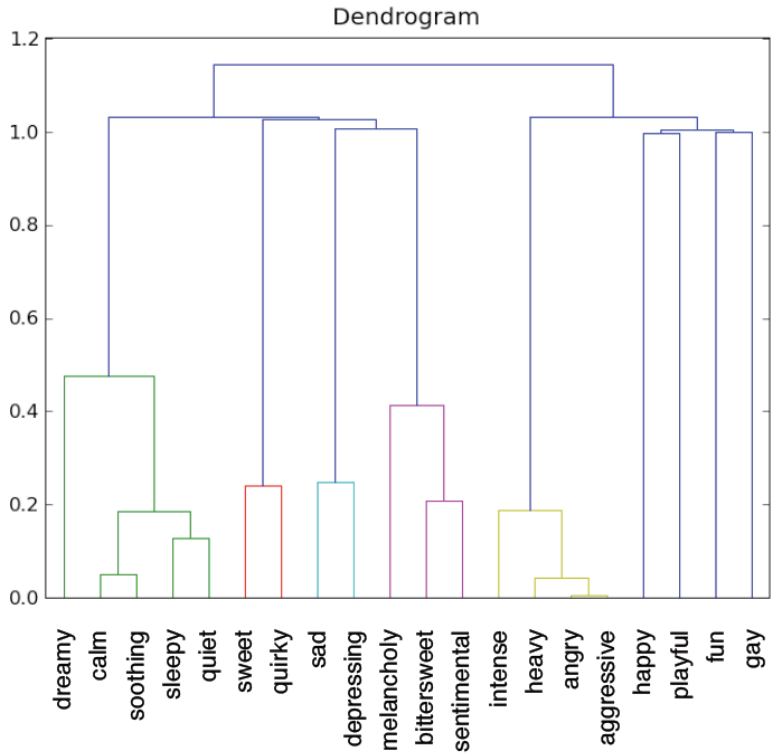



*Figure 3.3* *Dendrogram of the 20 most used music mood tags organized by Laurier et al [168]*

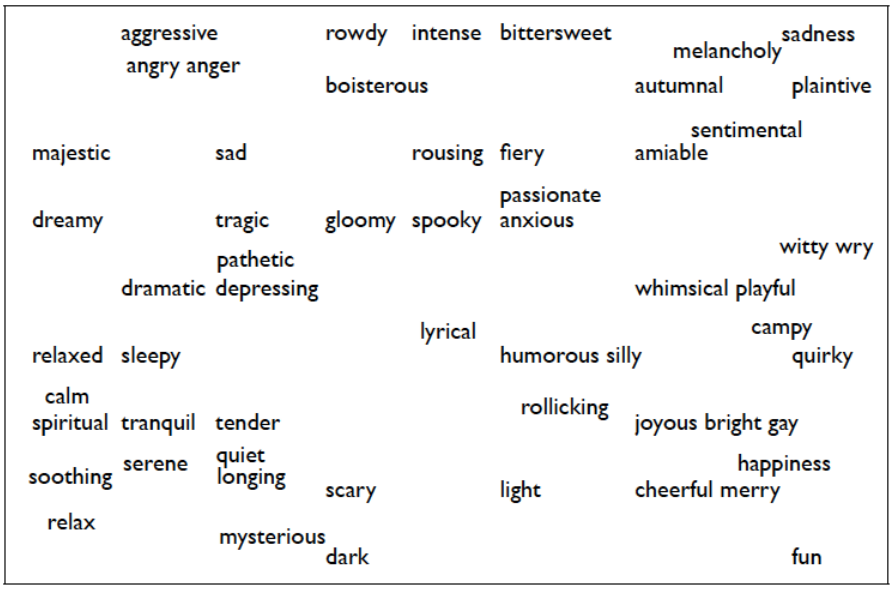



*Figure 3.4* *Self-Organizing Map of the mood tags in the semantic space [168]*

Bischoff and collaborators [33] proposed a hybrid approach that involves both metadata and content-based analysis. Evaluation performed by them shows that both kinds of information are complementary and should be merged for enhanced classification accuracy.

Saari *et el.* [265] investigated the role of audio and tags in music mood prediction. They compared Semantic Layer Projection and tags in terms of mood description accuracy. Their

results show that audio is in general more efficient in predicting perceived mood than tags. They also mapped music tags onto a Valence/Arousal plane and this representation is shown in Fig. 3.5.

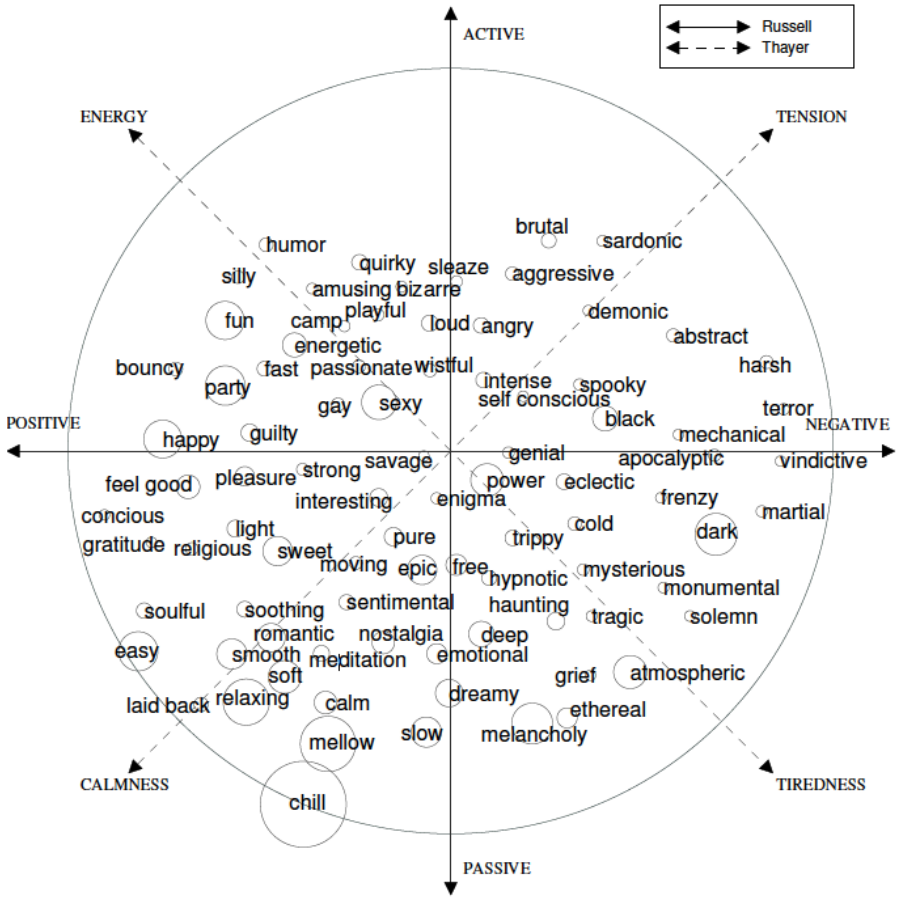



*Figure 3.5* *Mood music tags mapped onto Valence/Arousal plane [265]*

Hu and Yu [116] explored the relationship between mood and creativity in rock lyrics. Their research led to findings that lyrics of negative and sad songs demonstrate higher linguistic creativity than those of positive and happy songs. Hu and collaborators studied the role of lyric text in music mood classification. They tested almost 6000 songs from 18 moods (according to the users' tags) and came to the conclusion that lyrics can support music mood classification in some mood categories but cannot be the main and only clue. On the other hand, van Zaanen and Kanters [346] built classifiers that classify lyrics of songs into moods. The aim of their work was the automatic assignment of mood of music based on lyrics only. The implementation of the tf*idf method improved the results but still was not satisfactory. All discussed examples show that lyrics can supplement the content-based approach but cannot be the only classification cue.

### 3.2.3 Artificial Intelligence Methods Applied to MER

In MER many research studies involve advanced computational methods, e.g. regression approach, Support Vector Machines (SVM), Support Vector Regiression (SVR), or fuzzy logic [181,308,310,322,349], which are used for automatic mood assigning. Numerous researches show attempts to automatically classify music mood. The most popular approach involves the schema that consists of particular stages that are shown in Fig. 3.6. All of the elements are applied to the chosen set of music. First, pre-processing is implemented, then audio analysis is performed and parameters are extracted. The derived vector is fed into the chosen artificial intelligence computational method, which returns classified data. In many cases the classification method results are compared with those obtained in listening tests or with the experts' evaluation, and coherence of the notation is treated as a measure of correctness.

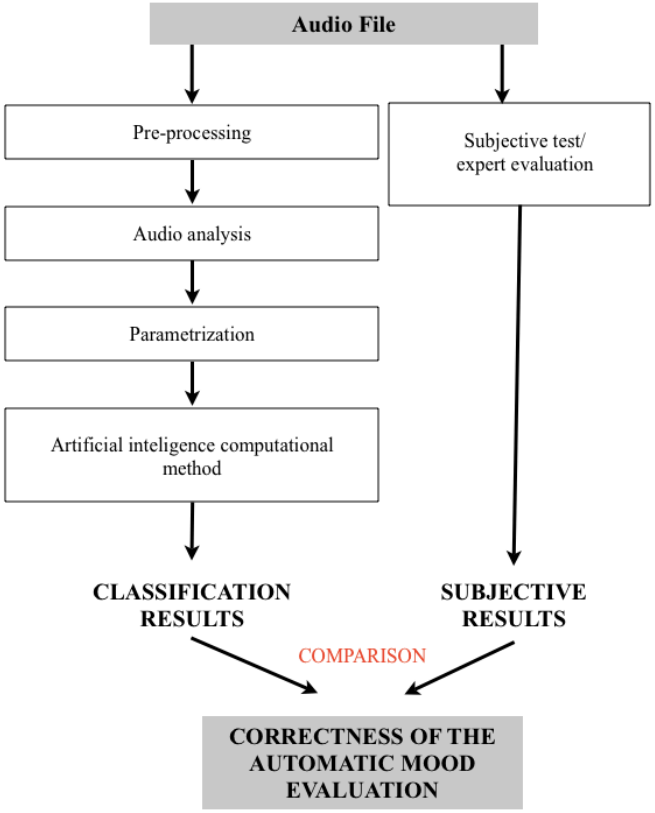



*Figure 3.6 Schema of the research aiming for automatic mood classification*

Results achieved in the projects mentioned before do not exhaust the application of computational methods to MER. The measure of correctness varies for different methods. Basic measures used in works cited in the subsequent section are: accuracy, F measure and

precision. These measures are commonly used and widely described in the literature of the subject and their basic definitions are described i.e. by Everitt and Skrondal [84].

First of all, an improvement in automatic efficacy is sought [181,349], 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 the 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. Li *et al*. [175] used a song database hand-labelled with adjectives belonging to one of the 13 categories and trained SVMs on timbral, rhythmic and pitch features. The authors report a large variation in the accuracy of estimating different mood categories, with the overall accuracy (F score) remaining below 50%. Feng *et al*. [88] used a Back Propagation Neural Network to recognize to which extent music pieces belong to four emotion categories ("happiness", "sadness", "anger", and "fear"). They used features related to tempo (fast-slow) and articulation (staccato-legato), and reported precision of approximately 66%. In multi-label classification, multiple labels are assigned to training examples from a set of disjoint categories. MER was first formulated as a multi-label classification problem by Wieczorkowska *et al*. [335] applying a classifier specifically adapted to this task. Sanden and Zhang [269] examined multi-label classification in the general music-tagging context using in their experiment 21,000 clips from Magnatune (each associated with one or more of 188 different tags). Using statistical distributions of spectral, timbral and beat features they tested Multi-Label *k*-Nearest Neighbors, Calibrated Label Ranking (CLR), Back Propagation for Multi-Label Learning, Hierarchy of Multi-Label Classifiers, Instance Based Logistic Regression and Binary Relevance *k*NN models. The CLR classifier using a Support Vector Machine outperformed all other approaches (they obtained an F-measure of 0.497 and precision of 0.642). However, Calibrated Label Ranking with Decision Trees, Backpropagation for Multi-Label Learning, and Multi-Label k -Nearest Neighbors also performed competitively. The results of selected studies, where learning machines and algorithms such as SVM, SVR, Gaussian Mixture Model (GMM), k-Nearest Neighbours (kNN), Naive Bayes (NB), Multiple Linear regression (MLR), J48 Decision Trees (J48), Dynamic Texture Mixture (DTM) and others were employed, are presented in Tab. 3.4 along with respective references, used music set, group of features, algorithm and accuracy.

The most important features used in listed works i.e. Mel-Frequency Cepstral Coefficient (MFCC), Statistical Spectrum Descriptors (SDD), temporal and others are described in Section 4.3, which is dedicated to audio parametrization.

*Table 3.4 Selected content-based music emotion recognition (MER) systems. Results evaluation described either: [1]- F-measure or [2]- Accuracy. Best reported configurations are indicated in bold*

| No. | Reference | Music set | Features | Machine Learning | Accuracy |
|---|---|---|---|---|---|
| 1 | Lin *et al*. (2009) [182] | 1535 | MARSYAS (436) | SVM | 56%[1] |
| 2 | Han *et al*. (2009) [104] | 165 | key, average energy, temporal, beat interval, harmonic strenght | SVR, **SVM**, GMM | 94.5%[2] |
| 3 | Zhao *et al*. (2010) [351] | 24 (Chinese & Western) | pitch (5), rhythm (6), MFCCs (10), SSDs (9) | Bayesian network | 74.9%[2] |
| 4 | Myint and Pwint [223] | 100 | intensity timbre, rhythm strength, correlation peak, temporal | SVM | 37%[2] |
| 5 | Lee et al. (2011) [170] | 1000 | timbre | SVM | 67.5%[2] |
| 6 | Mann *et al*. (2011) [193] | 114 | RMS, dynamics, spectral centroid, tonality, temporal | SVM | 80-94%[2] |
| 7 | Vaizman *et al*. [321] | 76 (piano+voc) | 34 MFCCs | DTM | 60%[1] |
| 8 | Saari *et al*. [265] | 104 (film music) | 52 (dynamic, rhythmic, pitch, harmonic, timbre) + MFCCs (14) | **NB**, k-NN, SVM | 59.4%[2] |
| 9 | Wang *et al*. [329] | 500 (Chinese music) | lyrics, rhyme | MLR, NB, **SVM**, J48 | 61.5%[1] |

The review of mood recognition presents various, often re-occurring, supervised machine learning techniques. Selected researches are listed along with their references in Tab. 3.5.

*Table 3.5 Selected supervised machine learning techniques applied to MER*

| **Supervised machine learning technique implemented in MER** | **Related references** |
|---|---|
| Support Vector Machines (SVM) | [25,22,175,203,234,273-275,285,314,315] |
| Gaussian Mixture Models (GMM) | [273-274] |
| Support Vector Regression (SVR) | [104,273,345] |
| Neural Networks | [88] |
| Linear Regression and Multiple Linear Regression | [75,273-275] |

It is worth noticing that very common approaches in recent publications are systems based on SVMs and GMMs for solving the classification problem and SVRs and Linear Regression techniques for the regression problem [203,234,285].

Some studies employed more complex processing that involved a few levels of the analysis. For example Pouyanfara and Sameti [250] performed two-level classification for MER tasks and achieved an accuracy rate of 78% for SVM and 87% for their two-level approach for 280 pieces dataset. It is important to note that the model of mood used by them was very simple and consisted of only 4 parts of Thayer's Valence/Arousal plane. Wu and collaborators [339] modeled music emotion recognition as a multi-label multi-layer multi-instance multi-view learning problem. In their approach music is formulated as a hierarchical multi-instance structure, where multiple emotion labels correspond to at least one of the instances with multiple views of each layer. Their Hierarchical Music Emotion Recognition model (HMER) captures music emotion dynamics with a hierarchical structure. On the other hand, Rauber and Frühwirth [254] from Vienna University of Technology proposed a SOM-enhanced JukeBox (SOMeJB) system [89] 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. The system introduced consisted of the 2-dimensional SOM representation that could be generated for any music set. A more complex system involved GHSOM (Growing Hierarchical Self-Organizing Maps) with a 3-layer architecture [256].

Most available studies on emotion recognition in music do not consider the problem of emotional variation throughout a song. In fact, the aim is usually to find the single emotion that best describes an entire song. Nevertheless, the mood of music is not always constant for the whole duration of the piece. That is the genesis of systems that involve mood tracking or dynamic mood recognition. Markov and Matsui [194] introduced a dynamic

structure based on State-Space Models that as a result returns the Arousal/Valence trajectory. Their experiment involved a very small dataset and did not lead to any general conclusions. Lu and collaborators [188] tracked mood changes over time using hierarchical and non-hierarchical frameworks based on Gaussian mixture models (GMM). Their system followed changes between Thayer's four principal mood quadrants in the valence-arousal representation. A similar approach was presented by [234] who aimed to track mood in audio music, specifically its changes over time in terms of Thayer's quadrants [308]. Using results from MoodSwings, an interactive game, Kim and collaborators [133] performed a continuous tracking; an example of their analysis for different parts of a song is presented in Fig. 3.7.

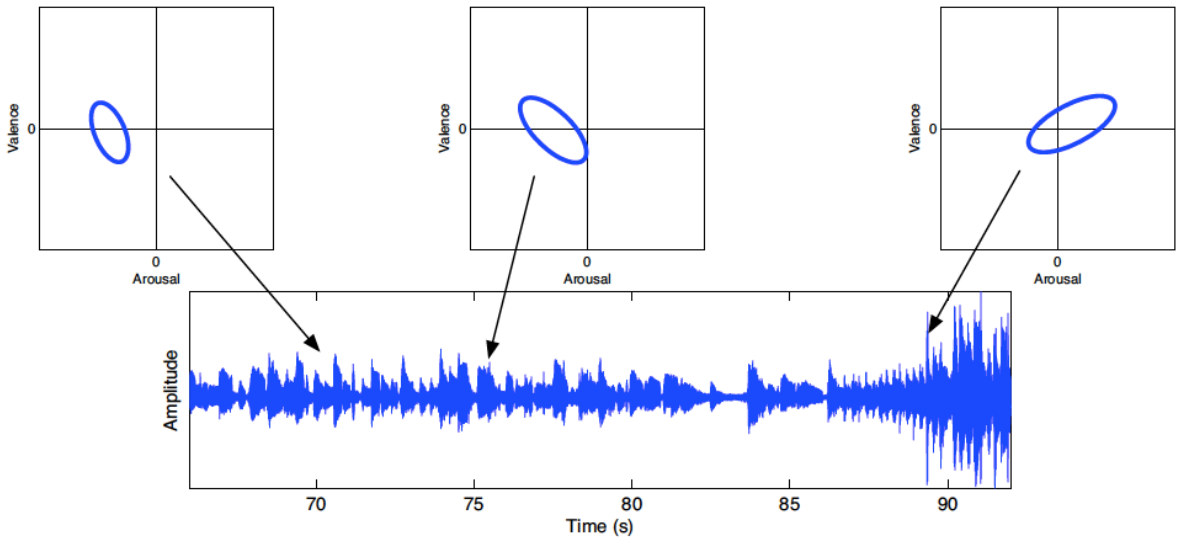



*Figure 3.7* *Mood of music changes in the music clip according to analysis performed by [133]. The ellipses represent the standard deviation of the evaluation*

Yang et al [345] proposed a solution, based on a continuous model using the Thayer's plane. The authors map each song to a point in the plane, using regression and audio features from various frameworks to predict the arousal and valence values. They achieved accuracy of up to 58.3% for arousal and 28.1% for valence.

Caetano and Wiering [49] criticized the common bag of frames (BOF) approach, which encodes audio signals as the long-term statistical distribution of short-term spectral features, commonly used in MER. They believe that BOG has several limitations and the semantic gap is responsible for the limited performance of many MER systems. As an

alternative to mood tracking, they introduced the theoretical framework of a computational model of auditory memory that incorporates temporal information into MER systems.

In a significant number of studies, tempo, rhythm and other time-based music features were considered as an important cue for MER. A detailed description of these features is available in Section 2.2. Tsunoo **[312]** implemented recognition of rhythmic patterns for specific music genres as well as music mood recognition. A simplified schema of his system is presented in Fig. 3.8.

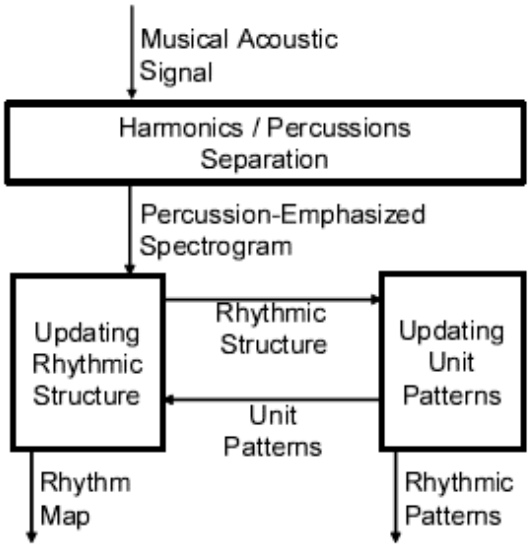



*Figure 3.8 A schema of the rhythm recognition system used for MER [312]*

Skowronek and his team along with other researchers investigated the role of percussiveness of sounds in terms of perception, including the influence of mood of music [221,287,288**].**

In MER, studies that refer to the ground truth play an important role. The aim of finding the ground truth is to create clear training sets that are easy to describe. This expression in MER tasks refers both to the set of music [287,288] as well as to mood classes within the categorical approach [113].

Recent years brought various tools and approaches that are involved in MER. Some of the most interesting ones are cited in the upcoming Section. Xu and collaborators [342] implemented the source separation to improve mood recognition. Results achieved for separated sources were better than for the regular analysis. Wang and collaborators [330] presented an interesting histogram density modeling approach with prediction of the

emotion distribution by a 2-D histogram over the quantized VA space. Chen and his team [56] proposed a model that adapts to the personal preferences of the listener in terms of mood of music. Park and collaborators [237] used Ranked Attributes Tree (RAT) which is able to recommend a music piece based on a combination of all ranked attributes, including mood. All of the works cited above show that MER is a very real and developing topic [224] and there is still a space for improvement and further studies.

### 3.2.4 Visualization Based on Mood of Music

Visualization of the results is especially important for these studies where dimensional models of mood are involved. In most cases music pieces are mapped onto the model plane or space. In these representations, pieces of music that are similar are placed close to each other, while a larger distance refers to smaller similarity. Pampalk introduced the idea of "Islands of Music" where whole music libraries are organized according to similarity between songs [232] (see Fig. 3.9). In this system the overall similarity is taken into consideration.

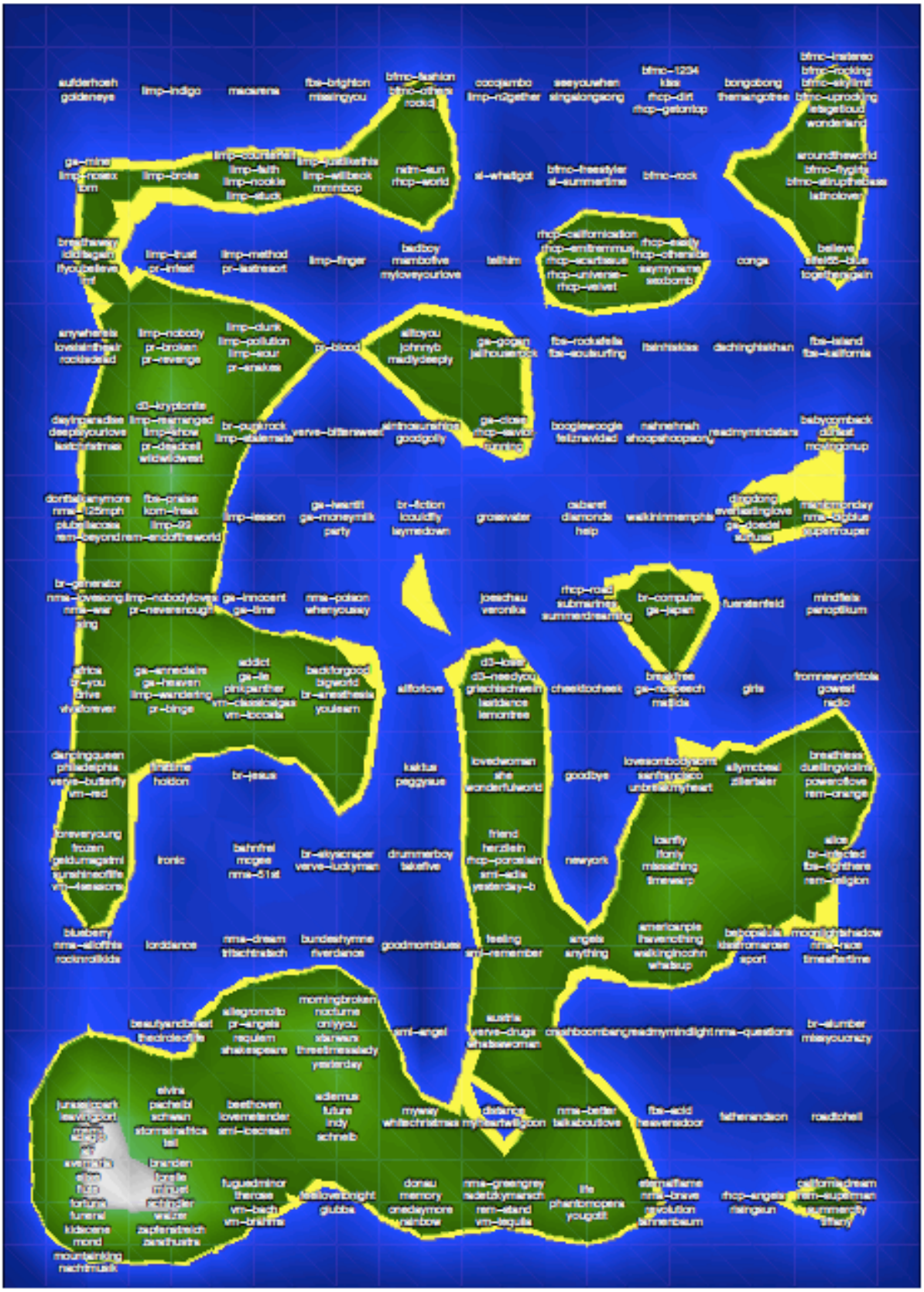


*Figure 3.9* *Example of music database organized according to the "Islands of Music" concept [232]*

For the subject of the presented work, visualizations based on mood of music are more interesting and relevant. Kim *et al*. mapped 50000 music pieces into a VA plane [133] (Fig. 3.10). In their approach songs are placed according to their mood, thus pieces with similar mood content are positioned close to each other.

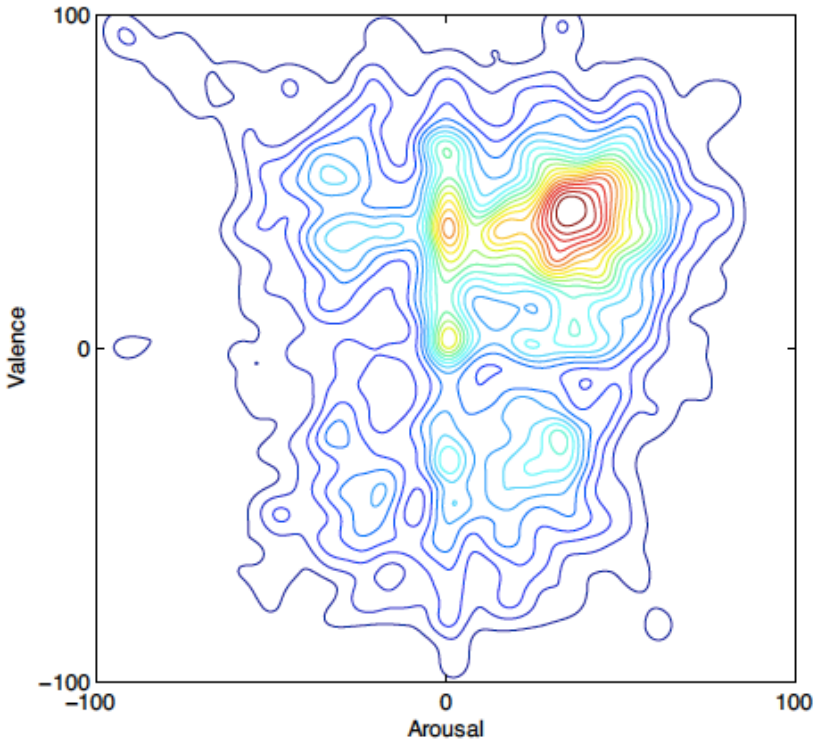



*Figure 3.10 Contour plot of the distribution of 50000 music pieces on Valence/Arousal plane [133]*

Yang *et al*. [345] proposed a different solution, based on a continuous model using Thayer's plane. The authors map each song to a point in the plane, using regression and audio features from various frameworks to predict the arousal and valence values. However, their approach reaches rather a low accuracy of up to 58.3% for arousal and 28.1% for valence.

A graphical representation of the songs in the mood plane is implemented in the Musicovery music platform [220] (Fig. 3.11). Musicovery and other music recommendation systems that include mood annotation are described in Section 3.3.

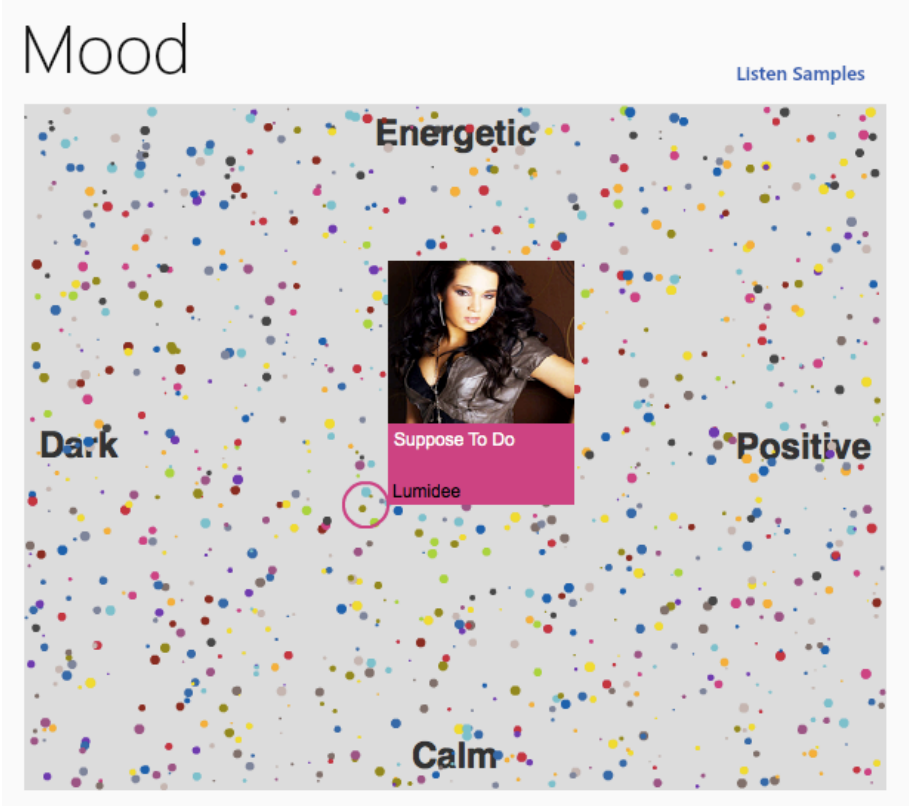



*Figure 3.11 Musicovery graphical representation of songs on the mood plane [220]*

### 3.2.5 Internet-based Systems of Mood of Music Data Collection

The internet is a common medium for the MER-based data collection. Researchers implemented several Internet-based games and applications. Moodswings is an interactive game, where players are placing a given song on the 2D VA plane [133,208]. The position can change during the song; moreover the participant can choose a game with a partner to see the difference and similarity in their mood judgment. The game incorporates each listener's subjective judgments of the emotional content (mood) of the music. The game is targeted at collecting dynamic (per-second) labels of the users' mood ratings in real-time using a two-dimensional space of emotional components [213]. That is a very interesting way of data collecting, although the interface is not exactly intuitive (the orientation of axes is different than in commonly used Thayer's model [308]), but includes very helpful emoticons representing mood (Fig. 3.12). Colors representing particular moods are almost coherent with the model proposed in this dissertation (see Section 6.2).

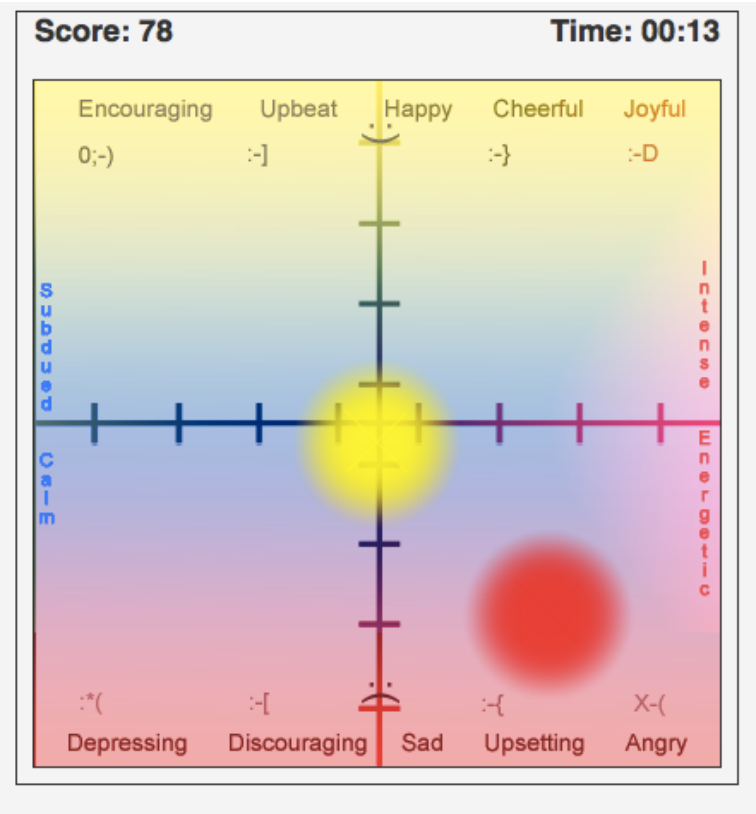



*Figure 3.12 Moodswings interactive interface [208]*

The ESP Game [3,80] is an idea in computer science for addressing the problem of creating difficult metadata. The same object is presented to both players and they are instructed to "think like each other" and type the same string (thus the name "ESP"). It turns out that the string on which the two players agree is typically a good label for the presented object. The idea behind the game is to use the computational power of humans to perform a task that computers cannot do (originally, image recognition) by packaging the task as a game. The ESP Game was shut down in 2011 as a part of Google Labs closure.

**Herd It** is a casual music game/music discovery tool developed by a team of students at the University of California [22]. Herd It connects music fans on Facebook and plays tunes from different musicals and asks players to share their opinions about what they hear in the music, including mood of music. MajorMiner is a music labeling game. The goal of the game, besides just listening to music, is to label songs with original, yet relevant words and phrases that other players agree with, including a description of music's emotional content. The authors use derived descriptions to train systems to recommend music [192]. All these applications were invented mostly for data collection purposes and are successful tools for subjective data acquisition.

## 3.3 SELECTED MUSIC RECOMMENDER SYSTEMS BASED ON MOOD OF MUSIC

Recently, due to a rapidly growing amount of online content, recommender systems have become extremely common and are applied in a variety of applications. The most popular ones are dedicated to movies, music, news, books, research articles, search queries, social tags, and products [7,260]. Music systems are often based on the collaborative filtering [162,236,297], that includes analysis of a large amount of users' behaviors and does not involve any information about content.

On the other hand, content-based filtering relies on the description or characteristics of an item. The description can be derived from tags assigned by experts, social tags or the content by itself. Recommendation can pertain to overall music preferences or similarity or a particular characteristic such as artist, genre, mood or others. WiMP [336] has options of searching by genre, artist, album and song. Recommendations and inspirations are created by the editors of the platform.

Some platforms combine both approaches. Last.fm [162] recommends music based on tags, similarity and similar preferences of users. Spotify [297] enables searching by artist, album, song, recommendation based on radio stations created by tag or artist or by other listener behaviors or preferences.

In the subsequent section, selected systems that can recommend music by mood are described.

Musicovery, an interactive Web-radio, includes different recommendation options, i.e. searching by artists or tags or a play your mood option. Songs are represented as colorful dots in the 2D space (Fig. 3.13).

Stereomood [298] will “Turn your mood into music”, but is based on tags, not always related to mood “I feel…”: ambient, action, dreamy, dirty, digital, drinking with friends, zombies, etc. (Fig. 3.14)

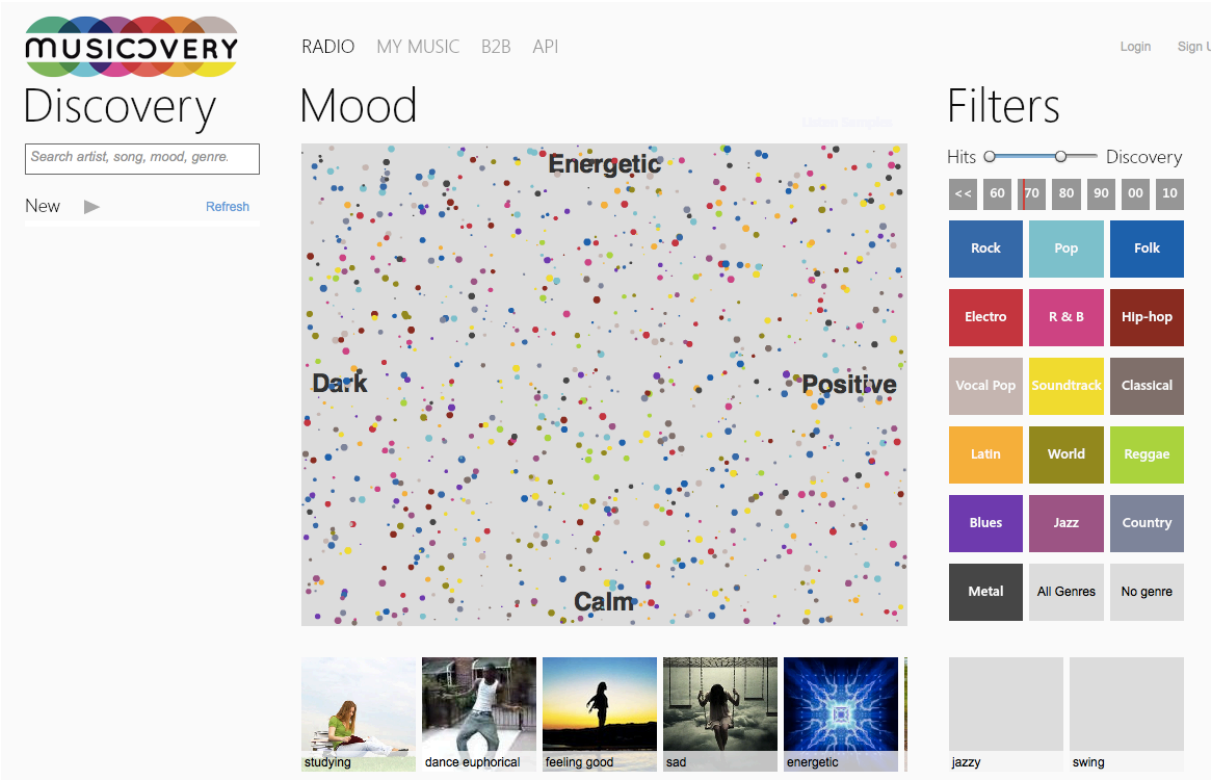



*Figure 3.13* *Graphical interface of Musicovery - music recommendation system based on music genre and mood of music [220]*

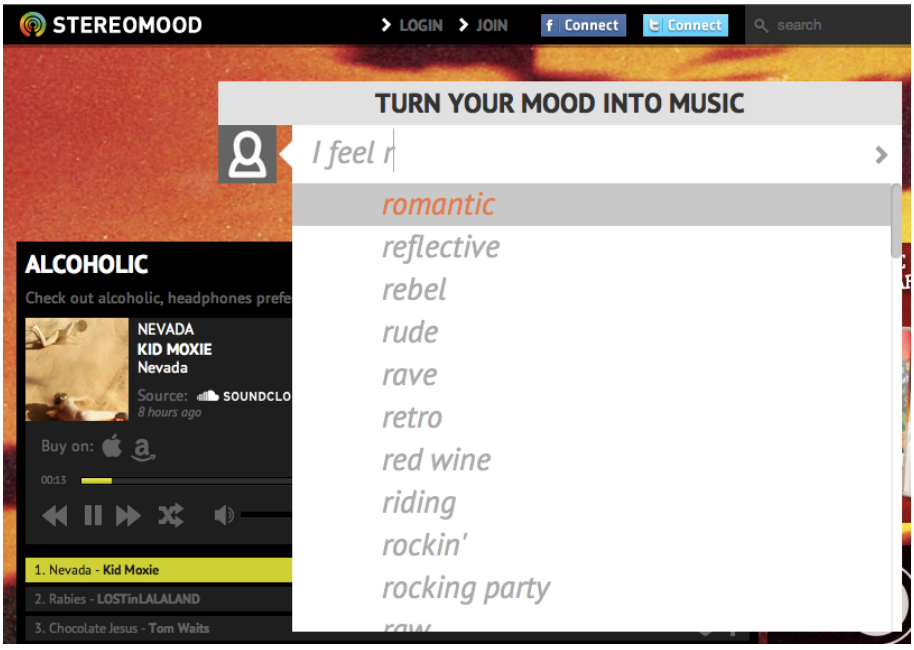



*Figure 3.14* *Graphical interface of Stereomood - music recommendation system based on tags related to music [298]*

In The Mood Music [119] is a New Age internet radio station that includes only music with calm and relaxing content. 8 Tracks [1] is based only on tags, very frequently not

correct ones. One can see many issues with this system; for example the tags "sad" and "sad/" are considered as separate categories.

Mood of music can be a very useful cue for music recommendation, not only for single users but especially for restaurants, bars, any type of background music use, music databases dedicated to film and advertisement, and many other applications. All presented systems are based on particular music sets. There is still space for development of the software/plugin that based on the content can organize any music library that is available online or is personally owned.

# 4 OVERVIEW OF AUDIO SIGNAL PARAMETRIZATION

In this Chapter first a brief summary of music parametrization in MIR is provided, then audio features applied to mood recognition are described and subsequently those features considered by the author of this thesis considered as having the potential to improve the results of the experiments are presented. In the final part of this Chapter, a case study is performed in which separate tracks and whole mixes of the songs are parameterized and subjected to thorough analysis. The proposed approach is based on the author's experience as a sound engineer working with musicians on the emotional content of their music.

As was already pointed out, within Music Information Retrieval there are two main approaches in music description: metadata-based and content-based [50]. While the metadata approach is based on textual information provided by experts or social exchange information platforms, the content-based approach involves music as a signal and its descriptors only. Discussion on the advantages and disadvantages of both approaches as well as more detailed information on particular solutions can be found in Section 3.1.

Content-based MIR is engaged in intelligent, automated processing of music such as music classification, sounds recognition, automatic music description and recognition, music creation and many other tasks [50,71,76,99,129]. The content-based approach constitutes the foundation of many studies and system implementations within the area of Music Information Retrieval [47,52,232,235]. Various types and segmentation and parametrization are used for this purpose. Parameters contain objective information about audio signals. The task of the researcher is to make the connection between descriptors and music perception and find the "hidden" sense behind them. This approach can pertain to a diversity of problems. In Music Emotion Recognition this applies to timbre, articulation, dynamics, harmony, key/mode, melody and time (rhythm/tempo) [20,50,174], thus signifies a necessity to create feature vectors containing many parameters related to the above music characteristics.

Depending on the aim, approach and computational method used, different features and parameters are involved. In the next few paragraphs, several examples of works are presented, where different sets of features are used to achieve various goals. Then groups of parameters commonly used for MER are described, including tools such as MPEG-7, MIR Toolbox and also those proposed by the author of the present thesis.

## 4.1 MUSIC MOOD RECOGNITION PARAMETRIZATION

Music as a form of art is perceived and interpreted in many different ways. It contains emotions, different meanings, references to other pieces and many other elements that are hard to interpret. On the other hand, audio signal can be treated as any other signal and parameterized according to characteristics of the temporal sequence. The relationship between music itself and parameters is very difficult to find. Therefore an additional layer of music features should be considered, which describes the characteristics of music. That said, music parametrization should consist of three layers (Fig. 4.1). Music is described by music features (characteristics) and then researchers are trying to identify parameters that are related to those features. Music features describe music elements and structure in special music language; some of them are rhythm, tempo, meter, key, harmony, dynamics articulation and many others. Terms related to music features are described in details in Section 2.2.

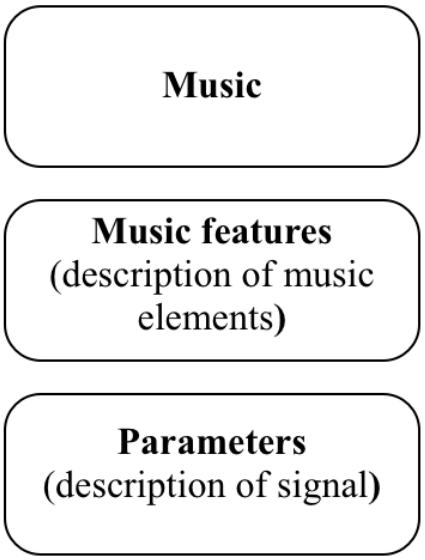



*Figure 4.1* *Three layers of music interpretation and description*

### 4.1.1 Music Features and Parameters Related to Mood of Music

In MIR and especially in MER, studies are performed to determine the relationship between music features and the impact on the listener. Music Emotion Recognition is the area where these relationships are crucial and underlie the whole concept of mood recognition. In the subsequent section relationships investigated by different researchers are cited and compared. On the other hand, it could be observed that composers commonly

use these rules; a skilled and conscious one uses particular elements of music to achieve desired impact on the listener.

Several works related to Music Mood Recognition refer to different groups and combinations of music features and specific parameters. However, it should be remembered that the basis of these descriptors had roots in research performed earlier within general Music Information Recognition [50,232].

Eronen [79] analyzed several features with regard to recognition performance in a musical instrument recognition system. He took into consideration a wide set of features covering both spectral and temporal properties of sound.

A number of studies considered interpretation between music features and valence and arousal [91,92,177,327]. Valence and arousal are expressions taken from Thayer's model of emotion, which is described in details in Section 2.5. A summarized set of music features important in the prediction of valence and arousal is listed in Tab. 4.1 [43]. Definitions and description of music features are included in Section 2.2.

*Table 4.1 Features in the prediction of valence and arousal [43]*

| **Valence** | **Arousal** |
|---|---|
| Chroma | Slow tempo |
| Percussiveness variability across bands | Loudness |
| Measure related to the ratio of fast and slow tempos | Chroma eccentricity |
| Modulation spectrum | Fast tempo |
| Harmonic strangeness | Spectral tilt |

Hevner summarized her findings related to the music features that create emotional content of music [108]. They are schematically shown in Tab. 4.2, according to eight clusters of adjectives included in Hevner's model of emotions (Fig. 2.20) [108].

The next step of the analysis is to determine the relationship between music features and particular parameters. Brinker [43] tested 79 features and proposed a schematic alignment, which is presented in Tab. 4.3.

*Table 4.2 Musical characteristics related to emotion groups with weights proposed by Hevner [108]*

| No. | Adjectives | Music Characteristic | Weight |
|---|---|---|---|
| 1 | spiritual, lofty, awe-inspiring, dignified, sacred, solemn, sober, serious | Firm rhythm | 18 |
| | | Slow tempo | 14 |
| | | Low pitch | 10 |
| | | Ascending melody | 4 |
| | | Major Mode | 4 |
| | | Simple harmony | 3 |
| 2 | pathetic, doleful, sad, mournful, tragic, melancholy, frustrated depressing, gloomy, heavy, dark | Minor mode | 20 |
| | | Low pitch | 19 |
| | | Slow tempo | 12 |
| | | Complex harmony | 7 |
| | | Firm rhythm | 3 |
| 3 | dreamy, yielding, tender, sentimental, longing, yearning, pleading, plaintive | Slow tempo | 16 |
| | | Minor mode | 12 |
| | | Flowing rhythm | 9 |
| | | High pitch | 6 |
| | | Simple harmony | 4 |
| 4 | lyrical, leisurely, satisfying, serene, tranquil, quiet, soothing | Slow tempo | 12 |
| | | Simple harmony | 10 |
| | | High pitch | 8 |
| | | Major mode | 3 |
| | | Ascending melody | 3 |
| | | Flowing rhythm | 2 |
| 5 | humorous, playful, whimsical, fanciful, quaint, sprightly, delicate, light, graceful | Major mode | 21 |
| | | High pitch | 16 |
| | | Simple harmony | 12 |
| | | Flowing rhythm | 8 |
| | | Fast tempo | 6 |
| | | Descending melody | 3 |
| 6 | merry, joyous, gay, happy cheerful, bright | Major mode | 24 |
| | | High pitch | 20 |
| | | Simple harmony | 16 |
| | | Flowing rhythm | 10 |
| | | Fast tempo | 6 |
| 7 | exhilarated, soaring, triumphant, dramatic, passionate, sensational, agitated, exciting, impetuous, restless | Fast tempo | 21 |
| | | Complex harmony | 14 |
| | | Low pitch | 9 |
| | | Descending melody | 7 |
| | | Firm Rhythm | 2 |
| 8 | vigorous, robust, emphatic, martial, ponderous, majestic, exalting | Low pitch | 13 |
| | | Firm rhythm | 10 |
| | | Descending melody | 8 |
| | | Complex harmony | 8 |
| | | Fast tempo | 6 |

*Table 4.3 Parameters related to musical features proposed by Brinker [43]*

| Feature class | Descriptor |
|---|---|
| Spectral | MFCC and modulations |
| Tonality | Chroma, key consonants, dissonants, harmonic strangeness, chroma eccentricity |
| Rhythm | Tempos (fast-slow), onsets, inter-onsets intervals |
| Percussiveness | Characterization and classification of onset per band |

This set allowed Brinker to achieve Valence and Arousal prediction with variance of 0.68 for Arousal and 0.50 for Valence [43].

As mentioned before, within the area of Music Emotion Recognition authors use different sets of parameters and algorithms. Panda for audio features extraction employed Marsyas and MIR Toolbox [234]. He fed parameters into SVMs classification and regression system, reducing the number of features using forward feature selection (FFS).

Rauber and his collaborators [256] executed 2-stage features extraction based on Psycho-Acoustic Models and used them in the SOM model (Fig. 4.2). They based their parameters on the basic of auditory perception that is loudness sensation and rhythm patterns per frequency band.

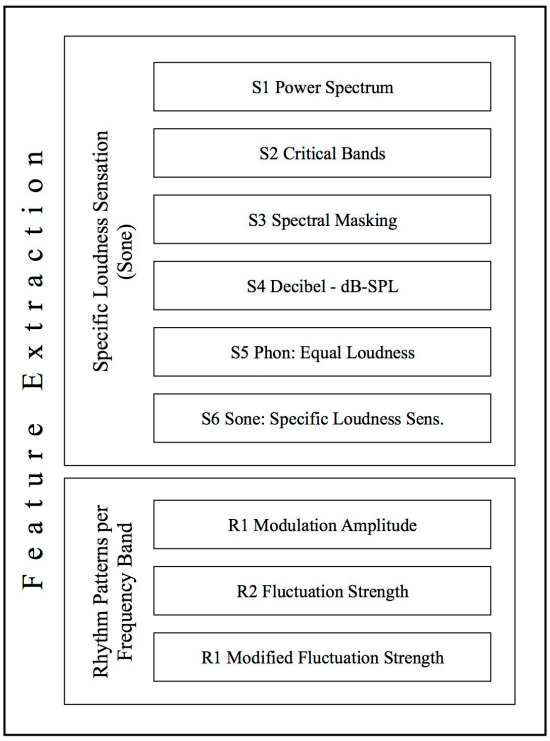



*Figure 4.2 2-stage feature extraction proposed by Rauber [256]*

Baume [25] described 47 different types of audio features and evaluated them for the purpose of music mood recognition. He tested these sets with different types of regressors (Fig. 4.3) as well as different subsets for each SVR regressor (Tab.4.4). Baume used different subset evaluation techniques that can be divided into three categories. He followed Liu's [184]: categorization of feature selection techniques the filter model, the wrapper model and the hybrid model [25]. The filter model relies on general characteristics of the data to evaluate feature subsets, whereas the wrapper model uses the performance of a predetermined algorithm (such as a support vector machine) as the evaluation criterion. The wrapper model gives superior performance as it finds features best suited to the chosen algorithm, but it is more computationally expensive and specific to that algorithm. The hybrid model attempts to combine the advantages of both. Results of his works are presented accordingly in Fig. 4.3 and Tab. 4.4.

For the purpose of MIR, including genre classification and MER, Li [175] used MFCC, STFT, DWCH and lyrics-based feature sets. At the same time Skowronek and her collaborators [287,288] employed mostly rhythm based, key and chroma features in their experiments. Schmidt [273,275] tested several subgroups of features (i. e. MFCC, Chroma, and Statistical Spectrum Descriptors) for emotion recognition and time-varying emotion regression. Schmidt analyzed individual sets of features and determined accuracy of 4-category classification for each of them (Tab. 4.5). Schmidt [273,274] tested several subgroups of features (i. e. MFCC, Chroma, and Statistical Spectrum Descriptors) for emotion recognition and time-varying emotion regression. Schmidt analyzed individual sets of features and determined accuracy of 4-categories classification for each of them (Tab. 4.5).

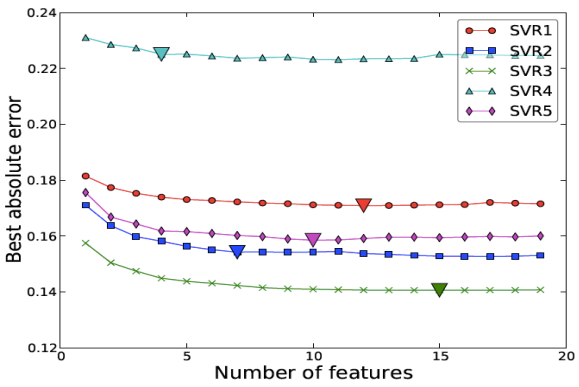



*Figure 4.3* *The absolute error of the best performing combinations for each of the five regressors. The first local minima are marked with triangles [25]*

*Table 4.4 Best feature combinations for each regressor [25]*

| Feature | **SVR1** | **SVR2** | **SVR3** | **SVR4** | **SVR5** |
|---|---|---|---|---|---|
| MFCC (20 coeffs) | ✔ | | | | |
| MFCC (12 coeffs) | | | ✔ | | |
| MFCC (7 coeffs) | | ✔ | | | |
| Spectral flatness | ✔ | | ✔ | | ✔ |
| Spectral spread | ✔ | | | | |
| Spectral inharmonicity | | | ✔ | ✔ | |
| Spectral smoothness | | | | ✔ | |
| Spectral sharpness | ✔ | | | | ✔ |
| Spectral irregularity | | | ✔ | | |
| Spectral centroid | | | | | ✔ |
| Bark coefficients | | | ✔ | | |
| Spectral valley (5 bands) | | | ✔ | | |
| Spectral valley (7 bands) | ✔ | | | | |
| Spectral valley (9 bands) | | ✔ | | | |
| Intensity ratio (9 bands) | | | ✔ | | |
| Tonal content space | ✔ | | | | |
| Tonal content function | | ✔ | | | |
| Key strength | ✔ | | ✔ | | |
| NNLS harmonic change | | | | | ✔ |
| Consonance | | ✔ | | | |
| Smooth power slope | ✔ | | ✔ | | ✔ |
| Scaled smooth power slope | ✔ | ✔ | ✔ | ✔ | |
| Peak-valley ratio | | ✔ | | | |
| Rhythm strength | | | ✔ | | ✔ |
| Mean correlation peak | ✔ | | | | |
| Mean onset frequency | | ✔ | ✔ | | ✔ |
| Beat counts | ✔ | | | | ✔ |
| RMS energy | | | ✔ | | |
| Zero-crossing rate | ✔ | | ✔ | | ✔ |
| Non-zero count | | | | ✔ | |
| Lowest value | | | | | ✔ |
| Highest value | | | ✔ | | |

*Table 4.5 Results of 4-way mood classification for several groups of parameters [275]*

| Feature Type | Accuracy |
|---|---|
| MFCC | 47.74±5.31% |
| Chromagram | 38.97±5.60% |
| Spectral Shape | 36.99±4.79% |
| Spectral Contrast | 48.67±6.10% |
| All Features | 38.24±4.60% |
| MFCC & Spectral Contrast | 50.18±4.18% |

As a result, Schmidt found different features appropriate for particular analysis [273-275]. Even within his research, length and the content of the feature vector varied.

[235] Panda and collaborators recently proposed a unique feature set consisting of standard and melodic features extracted directly from audio. Their results show that melodic features perform better than standard audio. They achieved the result of 64% F-measure, with only 11 features (9 melodic and 2 standard).

In many studies different layers of the music parametrization are mixed. Parameters are commonly included in music features set and vice versa. For example Brinker with collaborators [43] put on the same stage of analysis chroma and modulation spectrum (Tab. 4.2). On the other hand, systematics proposed by Thayer [308] are related only to the music features and are very hard to describe without the expert involved in the process. Methodology proposed by the author of the presented dissertation is based on 3-stage music analysis described before (Fig. 4.1). It involves attempt to create time-based parameters that describe particular musical content with mathematical tools. Proposed parameters and motivation behind them is presented in Section 4.6.

Each of the works presented in this Section refers to different sets of features and parameters, even though all of them aim at music mood recognition. Moreover, even within one computational method, different settings may require other parameters. Therefore it is difficult to determine one and only valid set of features which would be suitable for any approach to mood description and recognition of music.

### 4.1.2 Preprocessing

Preprocessing is a very important step that occurs before almost any analysis. The purpose of the process is preparing or adjusting data for the particular method or goal, extracting desired information and removing redundant content.

Usually, data values within a dataset may differ widely, which is one of the reasons of preprocessing. Normalization is often applied to bring various data to the same range of values. The procedure of and different types of normalization are described in Section 5.1.

Regardless of the type of the extracted feature, segmentation of the analyzed signal is first applied to set appropriate time resolution for the particular analysis and recognition tasks. Segmentation of audio piece is used to split it into its structural components such as vowels, phrases, notes, bars, etc. [159]. It is also commonly used for the purpose of analysis of varying signals to achieve more detailed information within time domain. During parametrization process, segmentation is implemented, i.e. to avoid bias caused by fragments of silence, observe differences between fragments, perform proper averaging process. Lengths of segments as well as their overlap, etc. are adjusted to match the requirements of the specific feature. In MER, segmentation is used not only for smoothing and determining whether some values are constant, but also for mood tracking [188].

## 4.2 MPEG-7-BASED AUDIO PARAMETERS

MPEG-7 audio parameters [215] are commonly used in MIR including MER [29,132,232], therefore they are listed and described in the subsequent section.

MPEG-7 standard is a set of standardized tools to describe multimedia content. MPEG-7 standard provides tools for audio, images and video data and are used both by humans as well as automatic systems. MPEG-7 Audio refers to audio content in any multimedia subject. Even though MPEG-7 Audio features are widely described and commented in the literature [132,215,216,267], therefore they will only be reviewed in the following Section shortly.

MPEG-7 Audio contains low-level descriptors that can be implemented in many applications as well as high-level descriptors, which are more specific to a set of applications described in standard [215]. Low-level descriptors are grouped and listed in Tab. 4.6. High-level tools include more complex schemes and procedures, which are: the audio signature Description Scheme, musical instrument timbre Description Schemes, the melody Description Tools to aid query-by-humming, general sound recognition and indexing Description Tools, and spoken content Description Tools. Since high-level descriptors are dedicated to specific tasks, which do not apply to the topic of presented dissertation, they are only mentioned briefly.

The MPEG-7 low-level audio descriptors are constructed to describe general attributes of audio signal. There are 17 temporal and spectral descriptors that can be extracted from audio automatically and may be used in a variety of applications. MPEG-7 descriptors are often used to determine similarity between different audio signals. Thus it is possible to identify identical, similar or dissimilar audio content. This also provides the basis for classification of audio content.

*Table 4.6 MPEG-7 Audio Low-level descriptors*

| Group | Low-level descriptor | Abbreviation |
|---|---|---|
| Basic | *Audio Waveform*<br>*AudioPower* | AW<br>AP |
| Basic Spectral | *Audio Spectrum Envelope*<br>*Audio Spectrum Centroid*<br>*Audio Spectrum Spread*<br>*Audio Spectrum Flatness* | ASE<br>ASC<br>ASS<br>ASF |
| Spectral Basis | *Audio Spectrum Basis*<br>*AudioSpectrumProjection* | ASB<br>ASP |
| Signal Parameters | *Audio Harmonicity Audio*<br>*Fundamental Frequency* | AH<br>AFF |
| Timbral Temporal | *Log Attack Time*<br>*Temporal Centroid* | LAT<br>TC |
| Timbral Spectral | *Harmonic Spectral Centroid*<br>*Spectral Centroid*<br>*Harmonic Spectral*<br>*Deviation Harmonic*<br>*Spectral Spread Harmonic*<br>*Spectral Variation* | HSC<br>SC<br>HSD<br>HSS<br>HSV |

### 4.2.1 Basic Descriptors

*Basic* Descriptors provide simple description of temporal structure of an audio signal. They are listed below including essential information.

**Audio Waveform**

*Audio Waveform (AW)* is defined to get a compact description of the shape of an audio signal. Whole signal is divided into non-overlapping frames (*hopSize*) and the lower (*minRange*) and upper (*maxRange*) limit of audio amplitude in the frame are stored. *AW* consist of *minRange* and *maxRange* time series, numbered accordingly to the frame index (*hopSize*). Comparison of the regular waveform and *AW* representation are shown in Figs. 4.4a and 4.4b.

**Audio Power**

Audio Power (AP) describes the temporally smoothed instantaneous power of the audio. The AP coefficient of the *m*-th frame of the signal is calculated according to the following formula:

$$AP(m) = \frac{1}{N}\sum_{n=0}^{N-1} |S(n+mN)|^2 \qquad (4.1)$$

An example of the AP description of a music signal is given in Figure 4.4c.

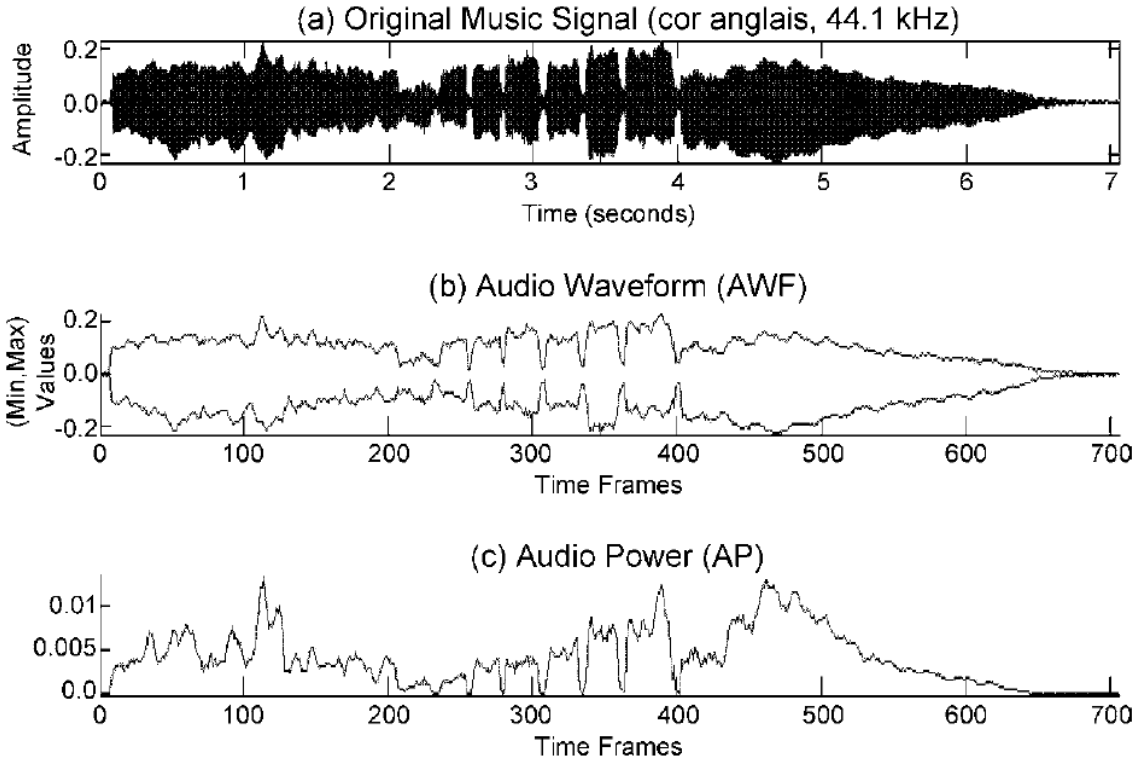



*Figure 4.4* *Comparison of representations of audio signal: a) original signal, b) Audio Waveform, c) Audio Power*

### 4.2.2 Basic Spectral Descriptors

*Basic Spectral Descriptors* provide time series of descriptions in the frequency domain. Frequencies are scaled logarithmically.

**Audio Spectrum Envelope**

*Audio Spectrum Envelope* (ASE) is a log-frequency power spectrum, which is obtained by summing the energy of the original power spectrum within a series of frequency bands. The

bands are distributed within the range [*loEdge, hiEdge*], according to the chosen resolution *r*, ranging from 1/16 of an octave to 8 octaves. The *ASE* within a band *b* is calculated as follows:

$$ASE(b) = \frac{1}{N} \sum_{k=loEdge(b)}^{hiEdge(b)} P(k) \qquad (4.2)$$

where $P(k)$ is the power spectrum (see Eq. 4.1).

**Audio Spectrum Centroid**

*Audio Spectrum Centroid* (ASC) stands for the center of gravity of a log-frequency power spectrum and is calculated as following:

$$ASC = \frac{\sum_{k=0}^{(N_{FT}/2)-low} \log_2(\frac{f(k)}{1000}) P(k)}{\sum_{k=0}^{(N_{FT}/2)-low} P(k)} \qquad (4.3)$$

Each frequency $f(k)$ of the power spectrum is weighted by the corresponding power coefficient $P(k)$. It is scaled with 1000 Hz as a central frequency. For ASC calculation low frequencies below 62.5 Hz are treated as a single band to avoid disproportionate weight of low-frequency components. Detailed information about particular is included in Kim's work [132].

**Audio Spectrum Spread**

*AudioSpectrumSpread* (ASS) is a measure of the spectral shape. It is defined as the second central moment of the log-frequency spectrum.

$$ASS = \sqrt{\frac{\sum_{k=0}^{(N_{FT}/2)-low} [\log_2(\frac{f(k)}{1000}) - ASC] P(k)}{\sum_{k=0}^{(N_{FT}/2)-low} P(k)}} \qquad (4.4)$$

**Audio Spectrum Flatness**

*Audio Spectrum Flatness* (ASF) characterizes an audio spectrum and provides a way to quantify how noise-like or how tone-like a given sound is [100,189]. It describes the amount of peaks or resonant structure in a power spectrum, as opposed to flat spectrum of white noise. A high spectral flatness (value 1.0 for white noise) indicates that the spectrum has a similar amount of power in all spectral bands. A low spectral flatness (approaching 0.0

for a pure tone) indicates that the spectral power is concentrated in a relatively small number of bands (mixture of sine waves) [29]. ASF is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum [189]. Spectral Flatness Measure is calculated as follows:

$$SFM_b(X) = sd\frac{[\sum_{k=0}^{K-1} X(k)]^{1/K}}{\frac{1}{K}\sum_{k=0}^{K-1} X(k)} \qquad (4.5)$$

where, $\boldsymbol{X(k)}$ is magnitude spectrum of signal $\boldsymbol{x(t)}$. The ASF is calculated within separate sub-bands *b*.

### 4.2.3 Spectral Basis

*Audio Spectrum Basis* (ASB) and *Audio Spectrum Projection* (ASP) descriptors were initially defined to be used in the MPEG-7 sound recognition high-level tool [132]. Their main concept includes the projection of an audio signal spectrum (high-dimensional representation) into a low-dimensional representation. This processing is aimed for classification systems. The extraction of ASB and ASP is based on normalized techniques which are part of the standard: the singular value decomposition (SVD) and the Independent Component Analysis (ICA).

### 4.2.4 Signal Parameters

*Signal Parameters* group of parameters describes the degree of harmonicity of audio signals.

**Audio Harmonicity**

*Audio Harmonicity (AH)* consists of two measures of the harmonic properties of a spectrum: *Harmonic Ratio* HR (the ratio of harmonic power to total power) and *Upper Limit* of *Harmonicity* ULH (the frequency beyond which the spectrum cannot be considered harmonic).

$$HR = \max_{l \le m \le N}\{\Gamma_l(m)\} \qquad (4.6)$$

where $\Gamma_l$ is defined as a normalized autocorrelation function of the signal within the frame *l*.

*Upper Limit of Harmonicity* is an estimation of the frequency beyond which the spectrum no longer has any harmonic structure.

**Audio Fundamental Frequency**

*Audio Fundamental Frequency* (AFF) provides estimations of the fundamental frequency $f_0$ in segments where the signal is assumed to be periodic. It can be interpreted as an approximation of the pitch of any music or speech signals.

Detailed calculation procedures *Signal Parameters* are included in [132].

### 4.2.5 Timbral Temporal

*Timbral Temporal* descriptors are extracted from the signal envelope in the time domain. They aim at describing perceptual features of instrument sounds based on ADSR envelope. It is schematically shown in Fig. 4.5.

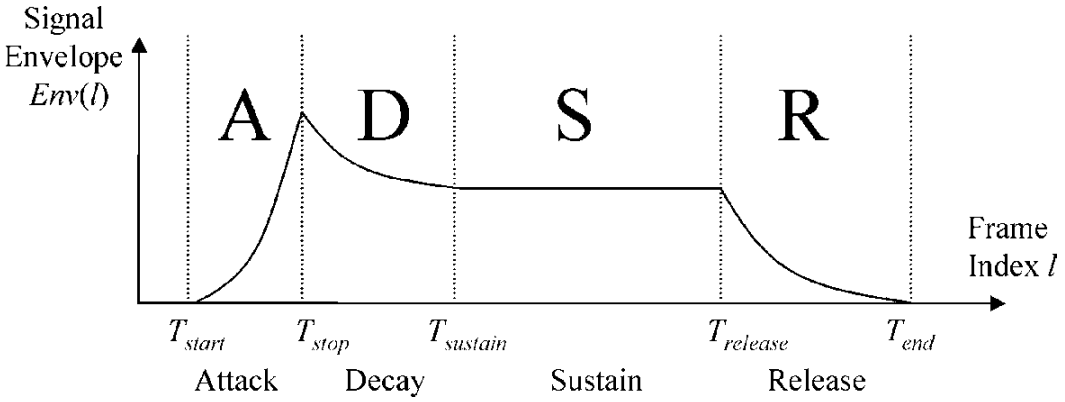



*Figure 4.5 Schema of ADSR envelope of a single sound*

Typical phases of ADSR are: **Attack** (the sound reaches its maximum volume), **Decay** (time when volume reaches the second volume level known as the sustain level), **Sustain** (is the volume level at which the sound sustains after the decay phase) and **Release** (volume reduces to zero).

**Log Attack Time**

*Log Attack Time* (LAT) is defined as the time it takes to reach the maximum amplitude of a signal from the minimum threshold time.

$$LAT = \log_{10}(T_{stop} - T_{start}) \tag{4.7}$$

**Temporal Centroid**

*Temporal Centroid (TC)* is defined as the time average over the energy envelope of the signal and is calculated as follows:

$$TC = \frac{N}{F_s} \frac{\sum_{l=0}^{L-1} (lEnv(l))}{\sum_{l=0}^{L-1} Env(l)} \tag{4.8}$$

where $Env(l)$ is the signal envelope.

### 4.2.6 Timbral Spectral Descriptors

*Timbral Spectral* describe the structure of harmonic spectra and are extracted in a linear frequency space.

**Harmonic Spectral Centroid**

*Harmonic Spectral Centroid* (HSC) is defined as the average, over the duration of the signal, of the amplitude-weighted mean (on a linear scale) of the harmonic peaks of the spectrum. For a given frame *l* it is defined:

$$LHSC_l = \frac{\sum_{h=1}^{N} f_{h.l} A_{h.l}}{\sum_{h=1}^{N} A_{h.l}} \tag{4.9}$$

where $f_{h,l}$ is frequency and $A_{h,l}$ is amplitude of *h*-th harmonic peak.

Thus, HSC value is obtained by averaging the local centroids over the total number of frames:

$$HSC = \frac{1}{L} \sum_{l=0}^{L-1} LHSC_l \tag{4.10}$$

**Spectral Centroid**

*Spectral Centroid* (SC) is not related to the harmonic structure of the signal. It gives the power-weighted average of the discrete frequencies of the estimated spectrum over the sound segment. SC is highly correlated with the perceptual feature of the brightness of sound [132] and is calculated as following:

$$SC = \frac{\sum_{k=0}^{N_{FT}/2} f(k) P_s(k)}{\sum_{k=0}^{N_{FT}/2} P_s(k)} \tag{4.11}$$

**Harmonic Spectral Deviation**

*Harmonic Spectral Deviation* (HSD) measures the deviation of the harmonic peaks from the envelopes of the local spectra. To achieve HSD, local measures are averaged over the total duration of the signal:

$$HSD = \frac{1}{L}\sum_{l=0}^{L-1} LHSD_l \qquad (4.12)$$

The calculation procedure of LHSD is described in details by Kim and collaborators [132].

**Harmonic Spectral Spread**

*Harmonic Spectral Spread* (HSS) is a measure of the average spectrum spread in relation to the HSC. At the frame level, it is defined as the power-weighted RMS deviation from the local HSC LHSC (Eq. 4.9).

**Harmonic Spectral Variation**

*Harmonic Spectral Variation* (HSV) reflects the spectral variation between adjacent frames. At the frame level, it is defined as the complement to 1 of the normalized correlation between the amplitudes of harmonic peaks taken from two adjacent frames.

## 4.3 OTHER PARAMETERS

In Music Emotion Recognition music features that need to be covered by parameters are timbre, articulation, dynamics, harmony, key/mode, melody and time (rhythm/tempo) [20,50,174]. Thus few groups of parameters relevant for MER can be identified. Some of the most important characteristics are presented and described below.

### 4.3.1 Timbre-Related Parameters

***Mel-Frequency Cepstral Coefficients (MFCC)***

Mel-frequency cepstral coefficients (MFCCs) are among the most widely used acoustic features in speech and audio processing. They were introduced by Mermelstein [199] as a tool for speech recognition. MFCCs are also commonly used in music information retrieval

applications such as genre classification, audio similarity measures and many others [50,221]. They are described as a low-dimensional representation of the spectrum warped according to the mel-scale, which reflects the nonlinear frequency sensitivity of the human auditory system [210,274]. Mathematically MFCC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of the pitch (Fig. 4.6) [185,207].

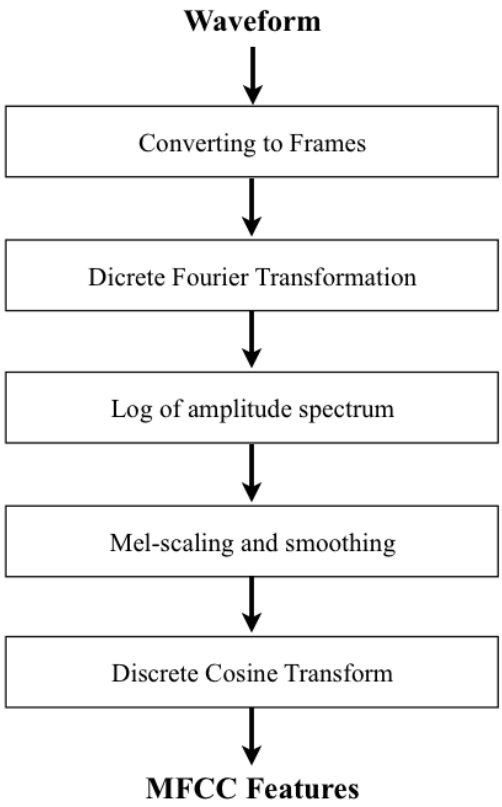



*Figure 4.6 A schema of MFCC calculation procedure*

A simple interpretation of MFCC within MIR is timbre of music.

***Statistical Spectrum Descriptors (SSD)***

In music and audio processing, *statistical spectrum descriptors* (SSD) are often related to timbral texture [50]. SSD are derived from a psycho-acoustically transformed Bark-scale spectrogram and comprise several statistical moments, which are intended to describe fluctuations on a number of critical frequency bands. The Bark scale spectrogram is then transformed into the decibel scale. Subsequently, the values are transformed into Sone values, in order to approximate the loudness sensation of the human auditory system. From this representation of a segment spectrogram the following statistical moments are computed in order to describe fluctuations within the critical bands: mean, median,

variance, skewness, kurtosis, min- and max-value are computed for each critical band, forming the SSD feature subset (Fig. 4.7) [178].

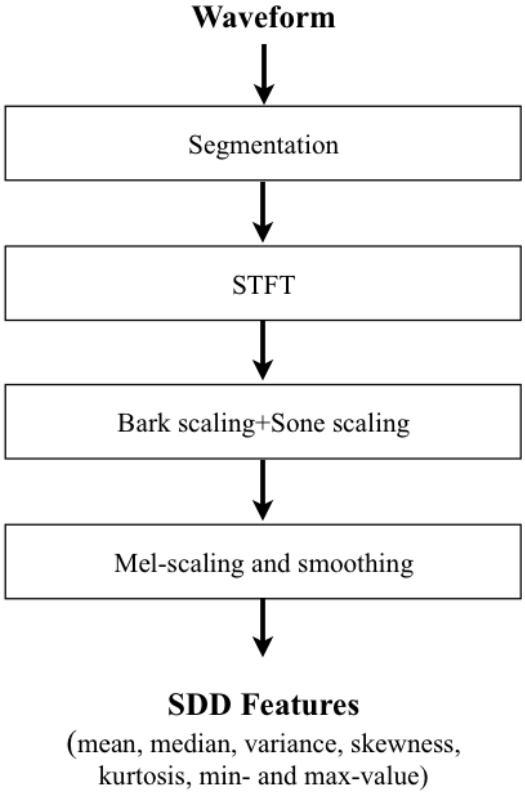



*Figure 4.7 SDD calculation process*

### 4.3.2 Time-Based Parameters

Tempo and rhythm are strongly connected with mood of music. This relationship was widely examined in many studies [64,108,127]. Main results of Hevner [108] studies are presented in Section 4.1.1 and listed in Tab. 4.2. Correlation between tempo and rhythmic patterns and mood of music was also studied by the author of presented dissertation [242]. Experiment and results are invoked in Section 6.3.

Time-based parameters are created to describe the music features such as tempo, rhythm, meter, accents and others. Description of time structures in music and related terminology is referred in Section 2.2.1.

Various researchers attempt to correlate time-based music features with parameters derived from audio signal [50,126,272]. Automatic estimation of the temporal structure of music, such as musical beat, tempo, rhythm, and meter, is not only essential for the computational modeling of music understanding but also useful for MIR. Rhythmic similarity has been used extensively in the audio domain for classification tasks [50].

Temporal properties estimated from a musical piece can also support Music Emotion Recognition.

There are numerous features such as tempo, Rhythm Patterns and Rhythm Histograms, beat detection in sub-bands, etc. Various content-based rhythm retrieval models exist and enable to extract information either from audio signal [90,99,163] or symbolic representation [6,258].

Rhythm Patterns describe modulation amplitudes for a range of modulation in the human auditory range, i.e. fluctuations (or rhythm) on a number of frequency bands. The feature extraction process for the Rhythm Patterns consists of two stages (see Figs. 4.8a and b). First, power spectrum that reflects human loudness sensation (Sonogram) is calculated. In the second step, the spectrum is transformed into a time-invariant representation based on the modulation frequency, which is achieved by applying discrete Fourier transform, resulting in amplitude modulations of the loudness in individual critical bands. From that data, reoccurring patterns in the individual critical bands, resembling rhythm, are extracted, which result in a time-invariant, comparable representation of the rhythmic patterns in the individual critical bands.

The Rhythm Histogram features are used to determine general rhythmic content [10]. The magnitudes of each modulation frequency bin of all critical bands are summed up, to form a histogram of "rhythmic energy" per modulation frequency. The histogram consists of 60 bins, which reflect modulation frequency between 0 and 10 Hz. For a given piece of audio, the Rhythm Histogram feature set is calculated by taking the median of the histograms of every 6-seconds segment processed (see Fig. 4.9).

An approach based on the recognition of rhythmic patterns is used also in various fields of MIR, i.e. automatic genre classification, dance music analysis, assisted annotation and many others [4, 83,158,172].

Going deeper into the structure of sound, zero-crossing (ZCR) or particular level crossing (i.e. RMS, 2xRMS etc.) are being calculated. These time-base parameters provide information about overall loudness or changes of loudness of the piece, and indirectly about dynamics.

Even though automatic recognition of rhythmic structures, tempo etc. is developing, created methods and descriptors are not completely covering the topic and do not enable sufficient recognition straight from audio signal.

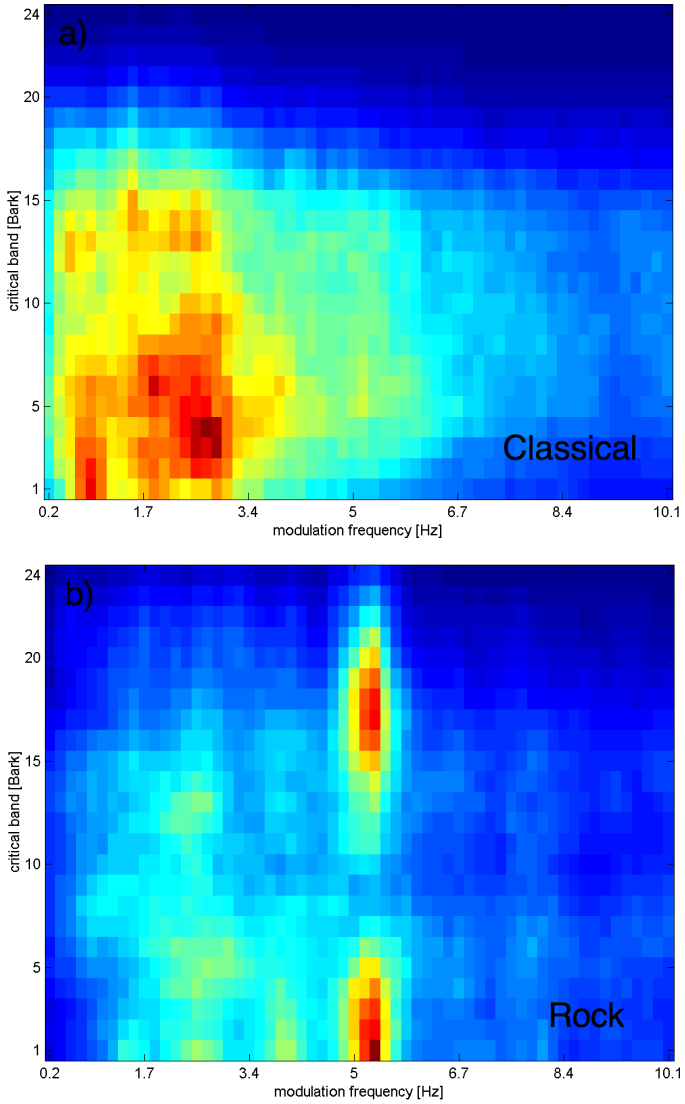



*Figure 4.8* *Rhythm Patterns of a) Classical and b) rock musical excerpts [10]*

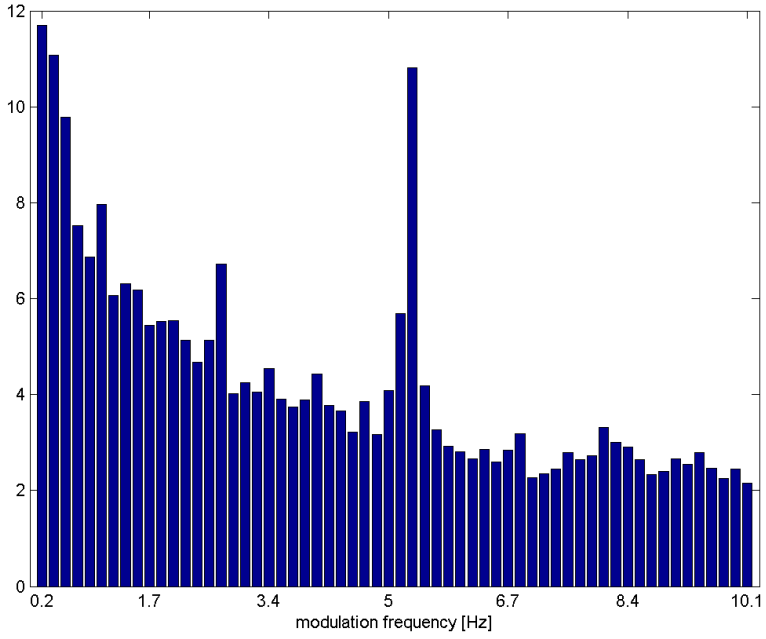


*Figure 4.9 Rhythm histogram for rock music piece [10]*

### 4.3.3 Chroma and Key Descriptors

Musical chords and key information are an important part of Western music and this information can be used to understand the structure of music. Detailed description of music features related to key, harmony, scales and chords is presented in Section 2.2.2.

The chromagram is a well-established method for estimating the Western pitch class components within a short time-interval [50]. It is essentially a circular version of the logarithmically warped spectrogram, where the frequencies corresponding to chroma in different octaves are grouped together and summed to estimate the energy at each of the 12 pitch classes. Using this feature, it is sometimes possible to obtain an indication of the overall musical key and modality.

The best performing chord- and key-recognition systems use Hidden Markov models (HMMs) to unify recognition and smoothing into a single probabilistic framework [30,171,286]. Most of systems consist of a transition matrix - a probabilistic model of the state sequence - and an output model - a probabilistic distribution that encodes the probability that one of the states produces the signal that we measure. Recognition systems consist of 24 or 36 chords (including major, minor and diminished triads for each of pitch classes), unless other limitations such as key or specific scale are implemented. The acoustic signal is represented as a set of chromagram frames so the output model represents the probability that each state (chord) produces any given chromagram signal. An example of automatic 12-bin pitch chord recognition performed by Lee and Slaney [171] is shown in Fig. 4.10.

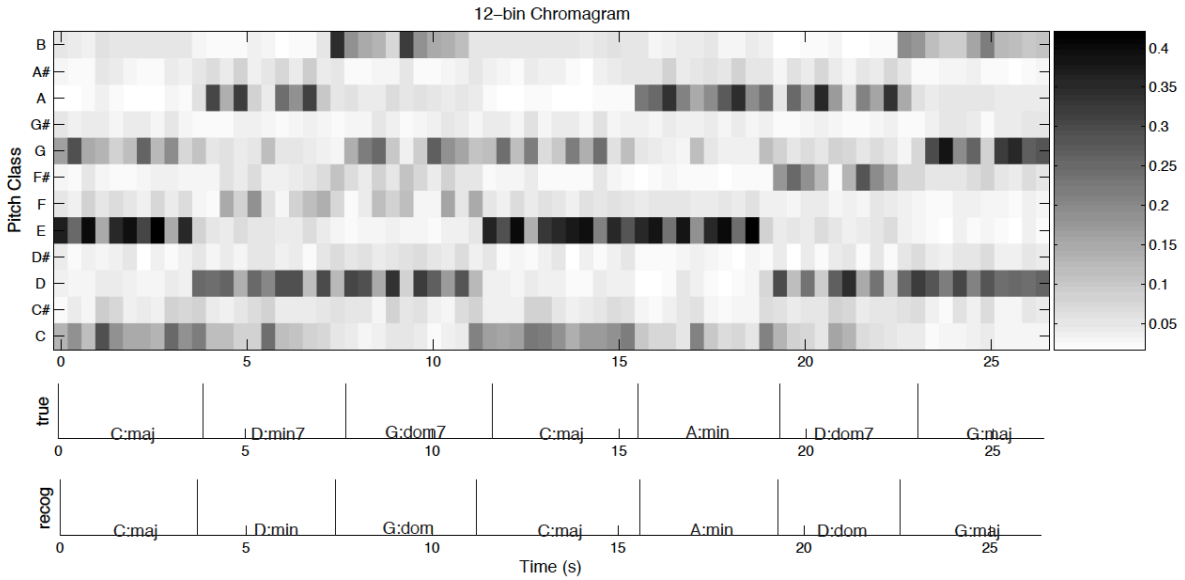



*Figure 4.10 12-bin chromagram of an excerpt from Bach's Prelude in C Major (BWV 846) performed by Glenn Gould. At the bottom chord labels with boundaries can be observed: "true" corresponds to the ground-truth annotation, and "recog" corresponds to the system output [171]*

## 4.4 PARAMETRIZATION TOOLS USED IN MIR

Numerous tools can help to organize, understand and search large collections of music in digital form. Some researchers created apparatus in different environments (i.e. Matlab, C++, C#) and shared them with MIR community [164,314]. Some of them are commonly used in MER and are described in the subsequent section.

### 4.4.1 MIR Toolbox

Lartillot with his team constructed a set of functions written in Matlab, dedicated to the extraction from audio files of musical features related to timbre, tonality, rhythm or form. MIRtoolbox contains good visualization capabilities as well as user-friendly implementation and manual. The tool is free and open-source [205]. MIRtoolbox consists of wide range of features including timbre, dynamics, tempo and others. The main structure of the MIRtoolbox parametrization is presented in Fig. 4.11. Detailed information on the MIR Toolbox can be found in papers by Lartilot and his collaborators [164-155].

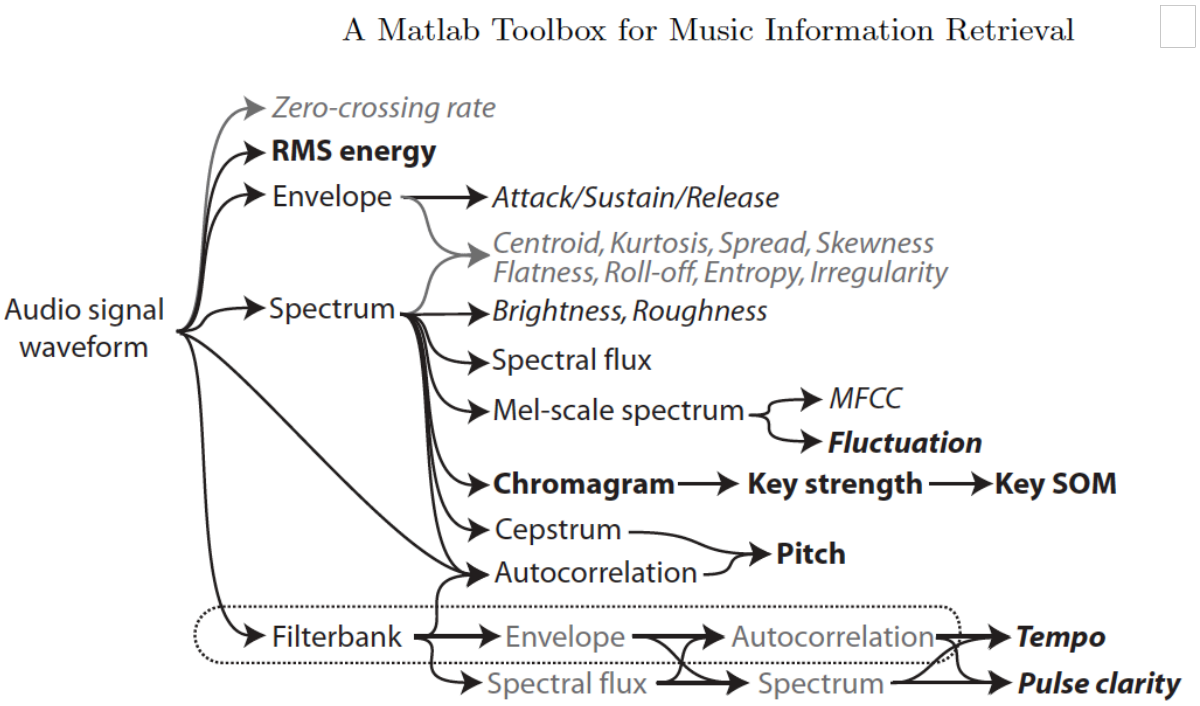



*Figure 4.11 Overview of the musical features that can be extracted with MIRToolbox [166]*

MIR Toolbox includes also "*miremotion*" script, which attempts to predict such description of emotion based on the analysis of the audio and musical contents of the recordings [75]. The output of "*miremotion*" corresponds to this underlying localization of emotional content within the 5 basic classes (happy, sad, tender, anger, fear) and within the 3 dimensions. Model of emotions, which combines classes and dimensional approach, used in MIRtoolbox is shown in Fig. 4.12.

Some of MIR Toolbox calculations are based on the Auditory Modeling Toolbox for Matlab (AMT) [9], which provides a model chain for the auditory hearing system. It includes models of all stages of the auditory system - from the outer ear up to the cortex. Detailed information about modules and components of AMT can be found in the source material developed by several auditory research groups [9].

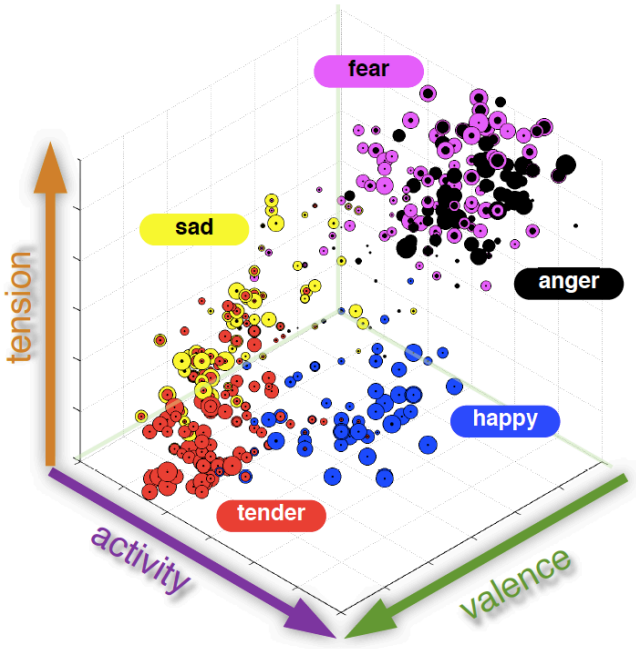



*Figure 4.12 Model of emotions used in MIRtoolbox [164]*

### 4.4.2 MARSYAS Parametrization

MARSYAS (Music Analysis Retrieval and Synthesis for Audio Signals) is a software framework for audio analysis, synthesis and retrieval introduced by Tzanetakis and his collaborators [41,314,315]. They have implemented a number of the features that have been proposed in the literature in the C++ environment. The list of features employed in MARSYAS is presented in Tab. 4.7.

*Table 4.7 List of features supported by MARSYAS*

| Feature name |
|---|
| Spectral Centroid |
| Spectral Moments |
| Spectral Flux |
| Pitch |
| Harmonicity |
| MFCC |
| Linear prediction (LPC) reflection coefficients |

### 4.4.3 MIDI as "Quasi Parametrization"

Another commonly used music description is MIDI (Musical Instrument Digital Interface), which is a technical standard protocol that contains symbolic representation. Despite the fact that MIDI is not a direct parametrization tool but rather notation of music

features, it enables specification of many aspects of music signal. MIDI notation contains information about tempo, instruments particular notes, their length, attack, decay etc. It allows extracting musical features such as melody, rhythm, key, tempo, etc. Thus MIDI is used in Music Information Retrieval, especially for high-level descriptors [50,74] Nevertheless, this approach requires operating on the databases that contain MIDI notation, which is not often the case. Moreover, this technique does not include information related to timbre and overall sound, which seem to be very important features for Music Emotion Recognition. Erolaand and Toiviainen created an open-source computational instrument in Matlab, dedicated to the analysis of symbolic representations of music called MIDItoolbox [74].

## 4.5 SYNAT PARAMETRIZATION

The SYNAT database is a collection of 52532 pieces of music described with a set of parameters obtained through the analysis of MP3-quality recordings. The database stores 173-parameter vectors, which in majority are the MPEG-7 standard items (109). However, the vector was additionally supplemented with 20 Mel-Frequency Cepstral Coefficients (MFCC), 20 MFCC variances and 24 time-related 'dedicated' parameters. The SYNAT project was realized by the Gdansk University of Technology (GUT) [155] and the music was collected from the Internet by means of a music robot. As this feature vector (FV) was examined in several MIR studies [110,153,156,159], also those performed at the earlier stage of this doctoral study [242,243], thus its content may be treated as a very thoroughly analyzed. The same FV, but extended to higher frequency bands, which resulted in 191 parameters, was used in the ISMIS'2011 conference in music recognition contest [155], in which more than 100 teams participated, thus in addition it may be treated as a kind of benchmarking. That's why the whole parameter set has been taken into consideration in the presented dissertation, especially as the author of this thesis participated in this study. The data analysis performed by the author of the present thesis leads to the conclusion that only some of them might be useful for mood recognition (see Section 6.2). In the next Section, sets of parameters chosen on various stages of the presented research are listed and described.

As MPEG-7 features and *Mel-Frequency Cepstral-Coefficients* are presented in Sections 4.2 and 4.3.1, thus they are only listed in Tab. 4.8 with consecutive numbers assigned. They

are used in the feature vector applied in the experiments performed by the Author. For the SYNAT database, the analysis band is limited to 8kHz. More details are included in the summary below:

**parameter 1:** *Temporal Centroid* (defined as the time averaged over the energy envelope),

**parameter 2:** Spectral Centroid (computed as the power weighted average of the frequency of the bins in the power spectrum) average value calculated for all frames,

**parameter 3:** *Spectral Centroid* variance calculated for all frames,

**parameters 4-32:** *Audio Spectrum Envelope* (ASE) - describes the spectrum of the audio according to a logarithmic frequency scale; average values in 29 frequency bands (calculated at one-fourth-octave intervals over the range of 62.5 Hz to 8 kHz),

**parameter 33:** ASE average value calculated for 29 frequency bands,

**parameters 34-62:** ASE variance values in 29 frequency bands (the same frequency bands as cited above),

**parameter 63:** averaged ASE variance parameter,

**parameters 64, 65:** *Audio Spectrum Centroid* (defined as the power weighted log-frequency centroid) - average and variance values,

**parameters 66, 67:** *Audio Spectrum Spread* (describes the second moment of the log-frequency power spectrum) - average and variance values,

**parameters 68-87:** *Spectral Flatness Measure* (SFM) average values for 24 frequency bands, calculated at one-fourth-octave intervals over the range of 250 Hz to 8 kHz; (SFM describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands); calculated according to ASF descriptor specified in MPEG-7 standard (described in Section 4.3.2).

**parameter 88:** SFM average value (averaged for 24 frequency bands);

**parameters 89-108:** Spectral Flatness Measure (SFM) variance values for 24 frequency bands,

**parameter 109:** averaged SFM variance parameters (averaged for 24 frequency bands),

**parameters 110-129:** 20 first MFCC (mean values),

**parameters 130-149:** 20 first MFCC (variance values),

**parameters 150-173:** dedicated parameters of the time domain obtained through the analysis of the envelope distribution in relation to the RMS (root mean square) value.

Parameters 150-173 refer to the time domain. They are based on the analysis of the distribution of sound sample values in relation to the root mean square values of the signal (RMS). Three reference levels were defined: $r_1$, $r_2$, $r_3$ – equal to 1, 2, 3 RMS values of the samples in the analyzed signal frame. Parameters 150, 151 and 152 correspond to the number of samples exceeding levels: $r_1$, $r_2$ and $r_3$.

$$p_n = \frac{count(samples\_exceeding\_r_n)}{length(x(k))} \tag{4.13}$$

where $n$=1,2,3 and $x(k)$ is the analyzed signal fragment.

In order to solve the issue of RMS varying in the analyzed frame, Kostek and her collaborators [155] have devised and introduced another approach. In this approach, each 5-second frame is divided into 10 smaller segments. In each of these segments parameters $p_n$ (Eq. 4.13) are calculated. As a result a sequence $P_n$ is obtained:

$$P_n = \{p_n^1, p_n^2, p_n^3, ...., p_n^{10}\} \tag{4.14}$$

where $k$ = 1 . . . 10 and $n$ = 1, 2, 3 as defined in Eq. 4.13.

In this way, six new features (parameters 153-158) were defined on the basis of $P_n$ sequences. Features are denoted as the mean ($q_n$) and variance ($v_n$) values of $P_n$, $n$ = 1, 2, 3. Index $n$ is related to different reference values of $r_1$, $r_2$ and $r_3$

$$q_n = \frac{\sum_{k=1}^{10} p_n^k}{10} \tag{4.15}$$

$$v_n = \mathrm{var}(P_n) \tag{4.16}$$

Three additional parameters (159-161) calculated as the ‘peak to rms’ ratio were specified. They are achieved in three different ways described below:

- parameter $k_1$ calculated for the 5-second frame,
- parameter $k_2$ calculated as the mean value of the ratio calculated in 10 subframes,
- parameter $k_3$ calculated as the variance value of the ratio calculated in 10 sub-frames.

Parameters 162-173 are based on the observation of the zero crossing rate and threshold crossing rate (TCR), which are calculated by counting the number of the signal crossings in relation to the 0, $r_1$, $r_2$ and $r_3$ values. These values (similarly as other previously presented parameters) are defined in three different ways: for the entire 5-second frame and as the mean and variance of the TCR calculated for 10 sub-frames.

*Table 4.8 The list of parameters within the SYNAT music database*

| No. | Parameter |
|---|---|
| 1 | *Temporal Centroid* |
| 2 | *Spectral Centroid* |
| 3 | *Spectral Centroid* variance |
| 4-32 | *Audio Spectrum Envelope* for particular bands |
| 33 | ASE average for all bands |
| 34-62 | ASE variance values for particular bands |
| 63 | averaged ASE variance |
| 64 | average *Audio Spectrum Centroid* |
| 65 | *Audio Spectrum Centroid* variance |
| 66 | average *Audio Spectrum Spread* |
| 67 | *Audio Spectrum Spread* variance |
| 68-87 | *Spectral Flatness Measure* for particular bands |
| 88 | SFM average value |
| 89-108 | *Spectral Flatness Measure* variance for particular bands |
| 109 | averaged SFM variance |
| 110-129 | *Mel-Frequency Cepstral Coefficients* (MFCC) for particular bands |
| 130-149 | MFCC variance for particular bands |
| 150 | number of samples exceeding RMS |
| 151 | number of samples exceeding 2×RMS |
| 152 | number of samples exceeding 3×RMS |
| 153 | mean value of samples exceeding RMS, averaged for 10 frames |
| 154 | variance value of samples exceeding RMS, averaged for 10 frames |
| 155 | mean value of samples exceeding 2×RMS, averaged for 10 frames |
| 156 | variance value of samples exceeding 2×RMS, averaged for 10 frames |
| 157 | mean value of samples exceeding 3×RMS, averaged for 10 frames |
| 158 | variance value of samples exceeding 3×RMS, averaged for 10 frames |
| 159 | peak to RMS ratio |
| 160 | mean value of the peak to RMS ratio calculated in 10 subframes |
| 161 | variance of the peak to RMS ratio calculated in 10 subframes |
| 162 | Zero Crossing Rate |
| 163 | RMS Threshold Crossing Rate |
| 164 | 2×RMS Threshold Crossing Rate |
| 165 | 3×RMS Threshold Crossing Rate |
| 166 | *Zero Crossing Rate* averaged for 10 frames |
| 167 | *Zero Crossing Rate* variance for 10 frames |
| 168 | RMS *Threshold Crossing Rate* averaged for 10 frames |
| 169 | RMS *Threshold Crossing Rate* variance for 10 frames |
| 170 | 2×RMS *Threshold Crossing Rate* averaged for 10 frames |
| 171 | 2×RMS *Threshold Crossing Rate* variance for 10 frames |
| 172 | 3×RMS *Threshold Crossing Rate* averaged for 10 frames |
| 173 | 3×RMS *Threshold Crossing Rate* variance for 10 frames |

## 4.6 ANALYSIS BY SYNTHESIS

Composers intend to communicate an idea, vision, feelings, etc. through music and to induce emotions in listeners. They use melodic, harmonic, rhythmic, dynamic, etc. elements of music structure to achieved a desired effect. Some of these techniques are referred to in Section 2.2. A relationship between these elements and emotions included in music was also analyzed by Hevner [108] (her findings are cited in Section 4.1.1). For example minor mode is strongly related to pathetic and tragic emotions, while slow tempo indicates dreamy or lyrical mood (Tab. 4.2). Depending on music genre, arrangement, artistic vision, time period, esthetics, etc., different instruments play different lines and have a different role. In Western music bass is very often a foundation and is responsible for overall groove of a piece, i.e. linking the harmony (chords) of a song with a distinctive rhythm (groove). At the same time, crash cymbal of the drum set is used to emphasize accents or important moments. Harmonic instruments such as guitar, piano or organs create harmony and melodic parts are performed by solo instruments, i.e. vocal, saxophone, etc. Sometimes absence of a specific instrument in a particular fragment can also be a mean of artistic expression. For example lack of rhythmic instruments can make music less energetic and smoother. All these assumptions are listed only for demonstrative purposes and are rather examples than rules strictly followed by composers.

An experience of the author, gained as a sound engineer and music producer, as well as all dependences described above lead to an approach, which is based on the role and impact of single instruments. In addition, also some other researchers, i.e. Xu and collaborators [342] implemented source separation to improve mood recognition. Results achieved for separate sources were better than for the regular analysis, therefore the approach based on the separate tracks seems to be well justified.

### 4.6.1 Separate Tracks vs. Mix

Following the idea of an analysis of single instruments, multi-track recordings and mixes were collected to enable a complete source separation. This was gathered as a kind of ground-truth in the experiments. Recordings from four different music genres: jazz, metal, pop and rock were selected to provide diverse examples. Since some of instruments are recorded using multi-microphones technique (described in Section 6.3), one track for each instrument section was sum-mixed. The list of musical excerpts along with tracks is

presented in Tab. 4.9. Since they are also included in the listening tests, indexes are assigned according to numbering in the main listening test (Section 6.2) and are listed in Appendix I with indexes A1-A4. Details regarding song titles, artists and albums are included in Appendix I.

*Table 4.9 List of music pieces selected for multi-track analysis. Details regarding song titles, artists and albums are included in Appendix I*

| Song No. | Music Genre | Tracks |
|---|---|---|
| A1 | jazz | drums |
| | | bass |
| | | piano |
| | | guitar |
| A2 | metal | drums |
| | | bass |
| | | guitars |
| | | vocal |
| A3 | pop | drums |
| | | bass |
| | | acoustic guitar |
| | | electric guitar |
| | | piano |
| | | vocal |
| A4 | rock | drums |
| | | bass |
| | | guitars |
| | | piano |

Most of parameters that describe music features are easy to implement on the separate tracks. In the final songs all sources overlap and particular aspects of sound are blurred. It depends on music genre, mixing concept, esthetics, and characteristics of every single song. For general overview, spectrograms of mixes of 4 selected songs are presented in Figs. 4.13-4.16. These representations show clearly that songs are different in terms of overall tempo, duration of single notes, spectral content, sections of the piece. All spectrograms are scaled from 0 to 30 seconds and from 0 to 3000 Hz. It is easy to conclude that jazz piece is much slower than metal (115 BPM vs. 240 BPM). But it is also worth noting that the onset time between notes is much longer for jazz and the difference is bigger than it should result from a simple tempo change. That is one of the examples where a simple description of music

feature does not directly translate to perception. Since the instrumentation for the pop song is less dense, it is possible to follow the solo line of the guitar, while for pop song (A3) spectra of guitars, piano and vocal overlap and they are not possible to separate. Mixing techniques called "Selective mixing" is based on the phenomenon that the removal of large parts of musical tracks in the time-frequency domain may not be perceived in the mix at all [136,137]. When sounds are mixed, in a small area of the time–frequency plane all respective segments of sounds can be removed except sound with the highest energy in that area. Despite that interferring process, quality remains satisfactory and in some cases can improve the accuracy of details [139,140].

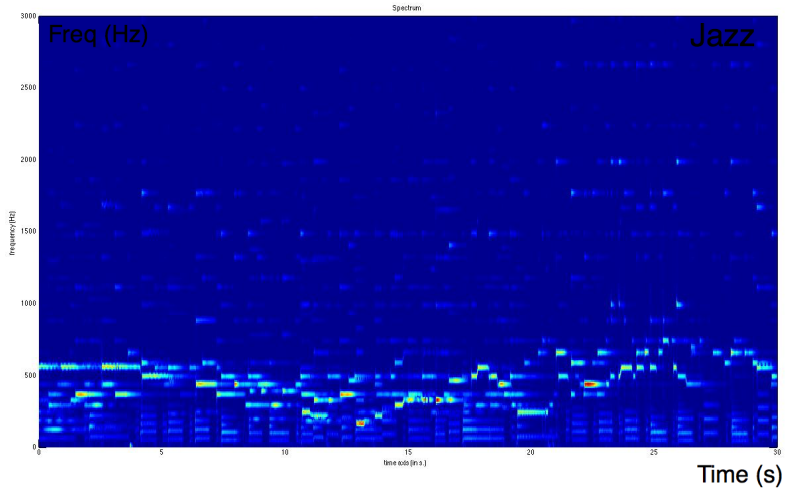



*Figure 4.13* *A spectrogram of the 30-sec. excerpt of jazz music (A1 according to 4.Tab. 9 and Appendix I). Axes denote time range of 30 seconds and frequency from 0 to 3000 Hz*

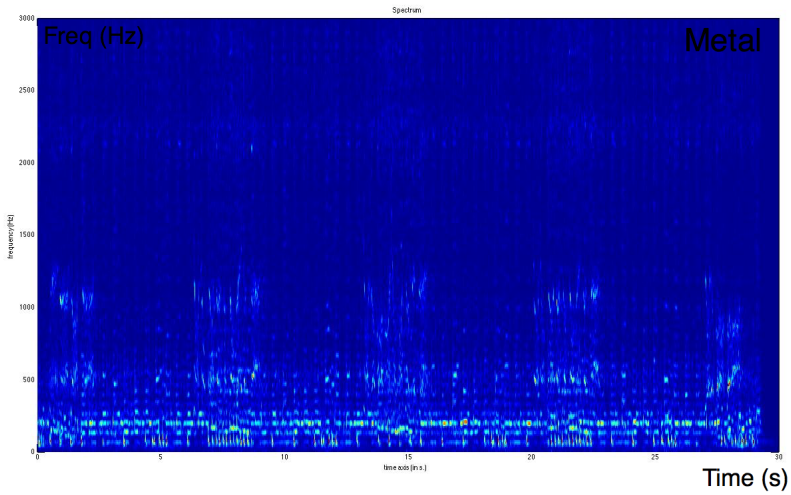



*Figure 4.14* *A spectrogram of the 30-sec. excerpt of metal music (A2 according to Tab. 4.9 and Appendix I). Axes denote time range of 30 seconds and frequency from 0 to 3000 Hz*

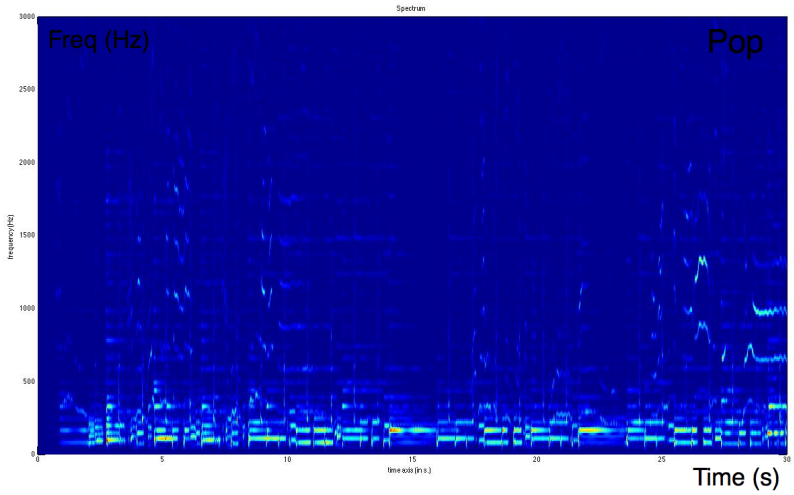



*Figure 4.15* *A spectrogram of the 30-sec. excerpt of pop music (A3 according to Tab. 4.9 and Appendix I). Axes denote time range of 30 seconds and frequency from 0 to 3000 Hz*

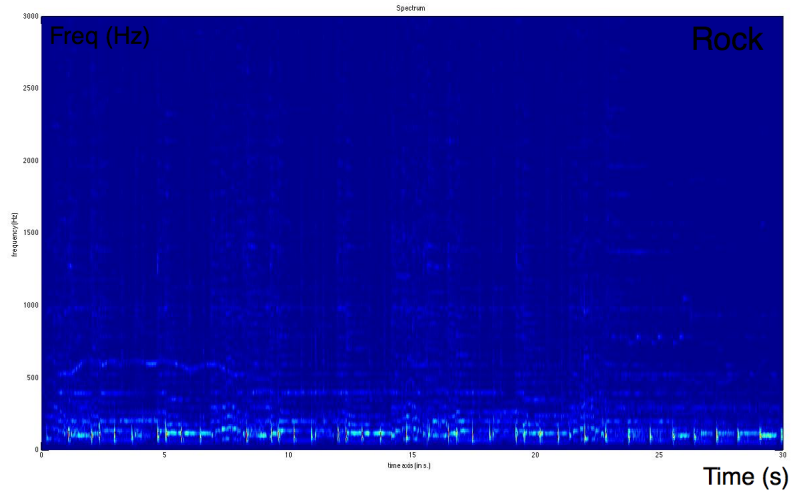



*Figure 4.16* *A spectrogram of the 30-sec. excerpt of rock music (A4 according to Tab. 4.9 and Appendix I). Axes denote time range of 30 seconds and frequency from 0 to 3000 Hz*

Some aspects of music are easy to observe on separate tracks but not within the complete mix. As mentioned before, during mixing process particular instruments are included in the final song in appropriate proportions (often changing during the song), with various processing and effects. All these steps make the analysis more complex and difficult. Sometimes the character of particular components causes problems with data interpretation. The snare drum played using brushes in the jazz tune (A1) is very hard to be discerned within the whole mix (especially in higher frequencies where it overlaps with harmonics of other instruments), even though it is very present while listening to. For demonstration purposes a spectrogram of snare drum track separately and the whole mix are shown respectively in Figs. 4.17 and 4.18.

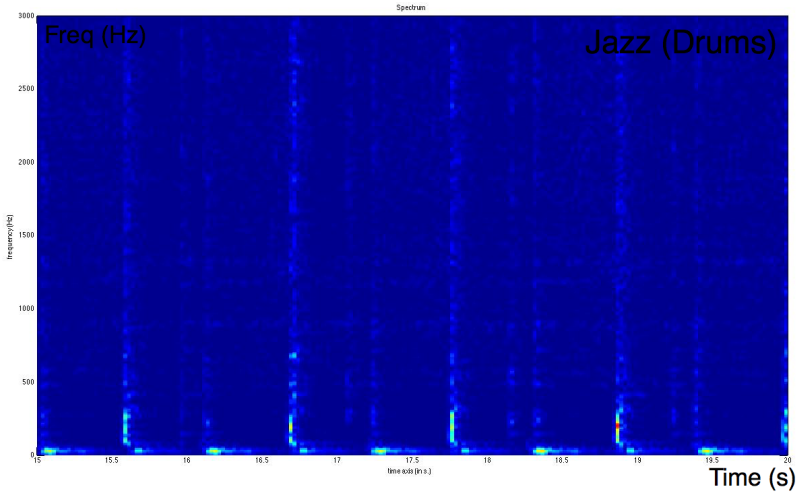



*Figure 4.17 A spectrogram of the 5-sec. fragment of solo drums track that is a part of jazz piece of music (A1 according to Tab. 4.9 and Appendix I). Axes denote time range of 5 seconds and frequency from 0 to 3000 Hz*

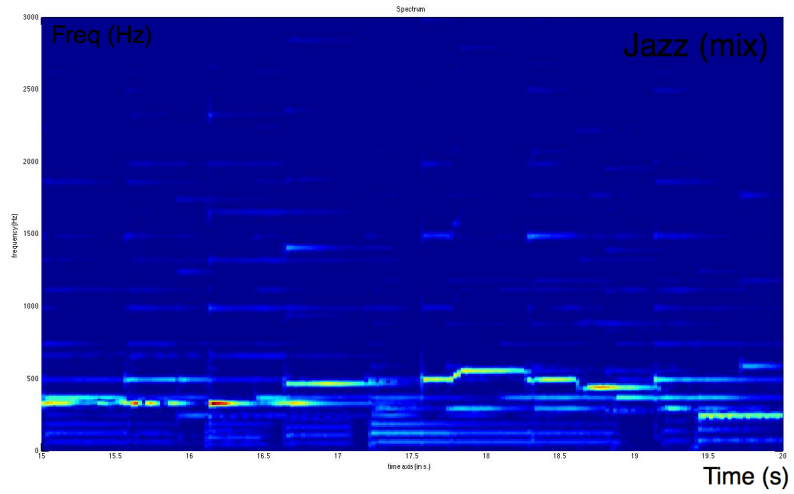



*Figure 4.18 A spectrogram of the 5-sec. fragment of jazz music (A1 according to Tab. 4.9 and Appendix I). Axes denote time range of 5 seconds and frequency from 0 to 3000 Hz*

For pop song (A3) due to dense instrumentation, spectra of guitars, piano and vocal overlap and are not possible to be separated. As an example a solo piano track (Fig. 4.19) and the whole mix (Fig. 4.20) are presented. What is also interesting, the vocal line and especially aliquots are easy to be recognized in the complete mix (Fig. 4.20).

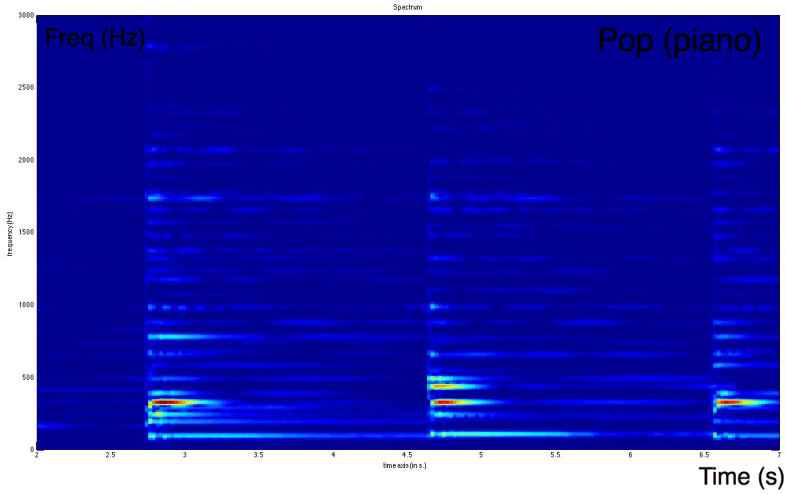



*Figure 4.19 A spectrogram of the 5-sec. fragment of solo piano track that is a part of jazz pop of music (A3 according to Tab. 4.9 and Appendix I). Axes denote time range of 5 seconds and frequency from 0 to 3000 Hz*

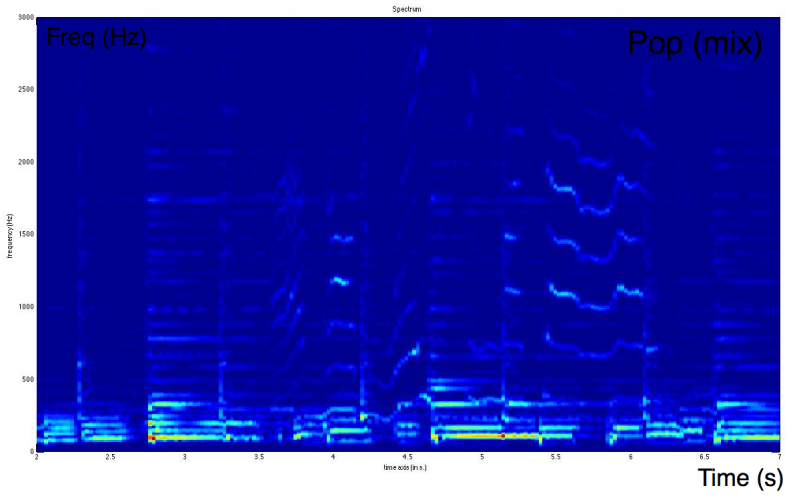



*Figure 4.20 A spectrogram of the 5-sec. fragment of pop music (A3 according to Tab.4. 9 and Appendix I). Axes denote time range of 5 seconds and frequency from 0 to 3000 Hz*

Also parameters related to specific music characteristics are more difficult to extract from the prepared mix than from separate tracks, even though some harmonic information can still be retrieved. Chromagrams calculated for tracks of a single harmonic instrument and for the whole mix are compared for pop and metal and presented in Figs. 4.21 and 4.22. Instruments for the analysis were selected according to the harmonic role they play in the arrangement (they are the main element that determines the harmonic content). It is important to notice that all analyzed songs have quite simple and traditional harmonic structure.

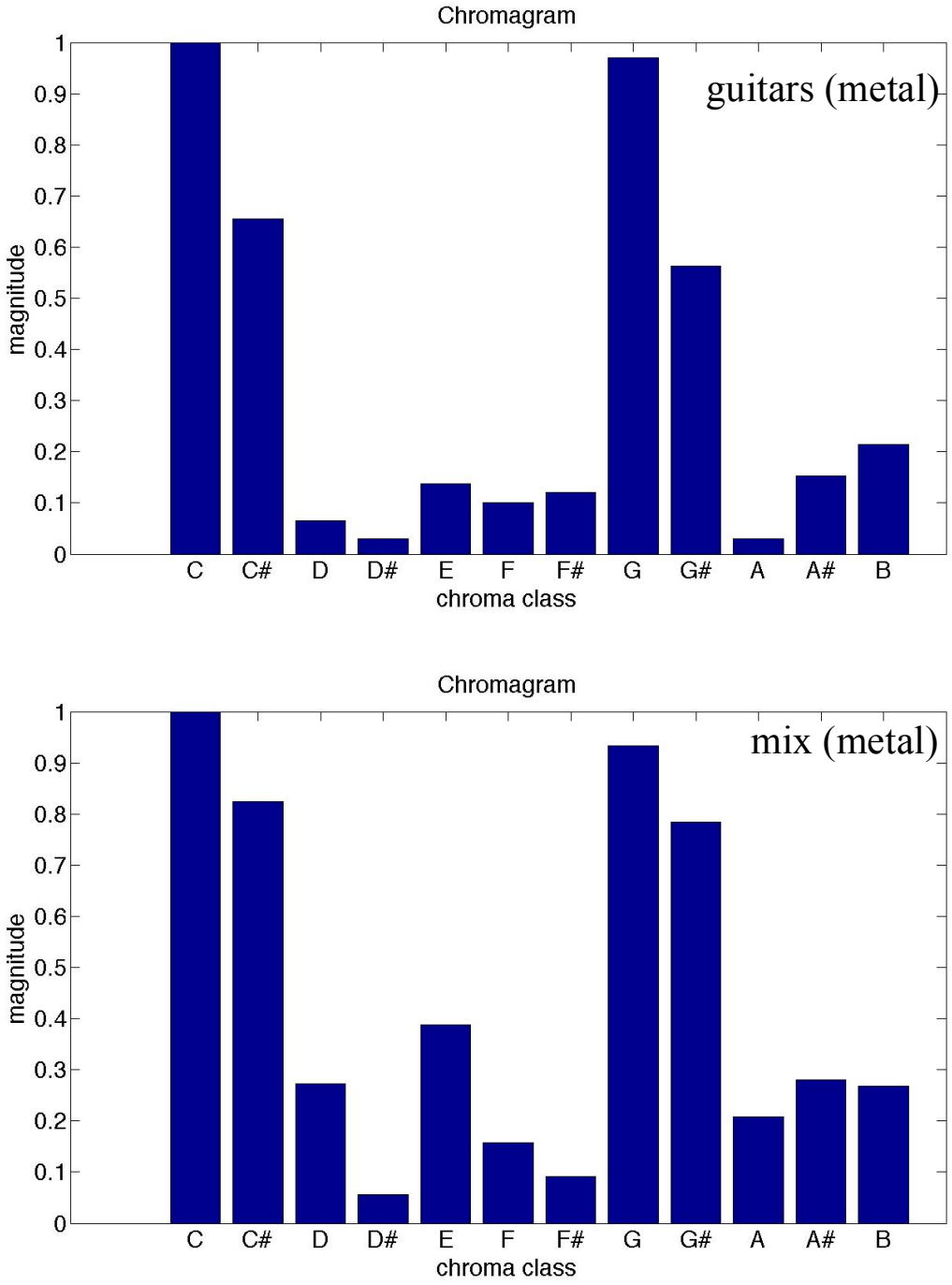



*Figure 4.21 Chromagram calculated for a single harmonic instrument track (guitars) and the whole mix of metal music*

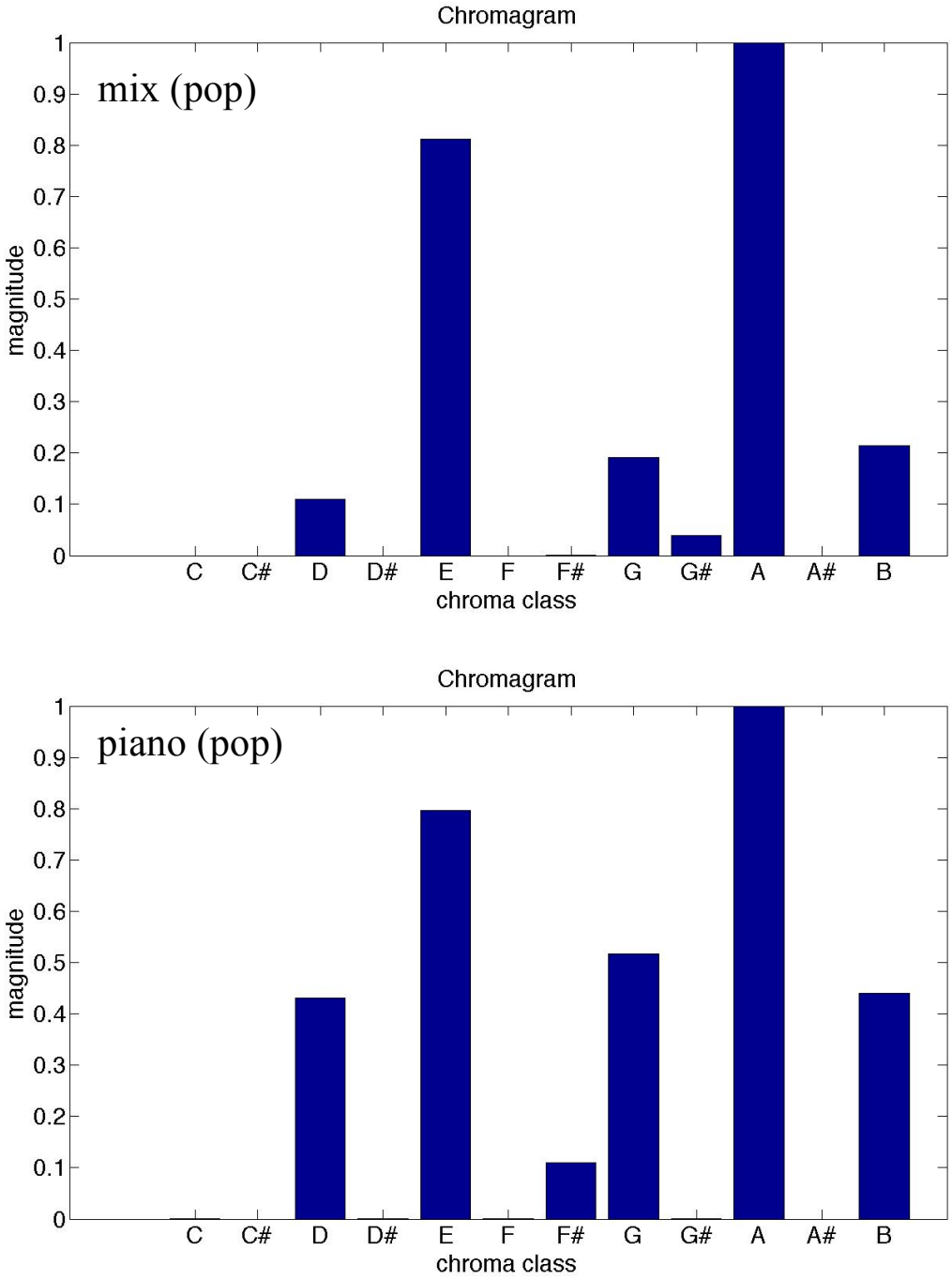



*Figure 4.22 Chromagram calculated for a single harmonic instrument track (piano) and the whole mix of pop music*

While the harmonic content is still possible to extract for the whole mix, rhythmic characteristics are much more difficult. Specific values related to temporal information are hard to retrieve from the whole mix. Despite these difficulties, using dedicated tools it is possible to achieve some information, but unfortunately not for all music arrangements and characteristics. An example of the rhythmic content representation is presented in Fig. 4.23,

where rhythm histogram is calculated for separate tracks from drums and the whole mix of pop music. It is interesting that the description of the harmonic content is possible, while rhythmic it is too difficult. In this case it might be related to rhythmic parts played by guitar and piano, which overlap with drums onsets.

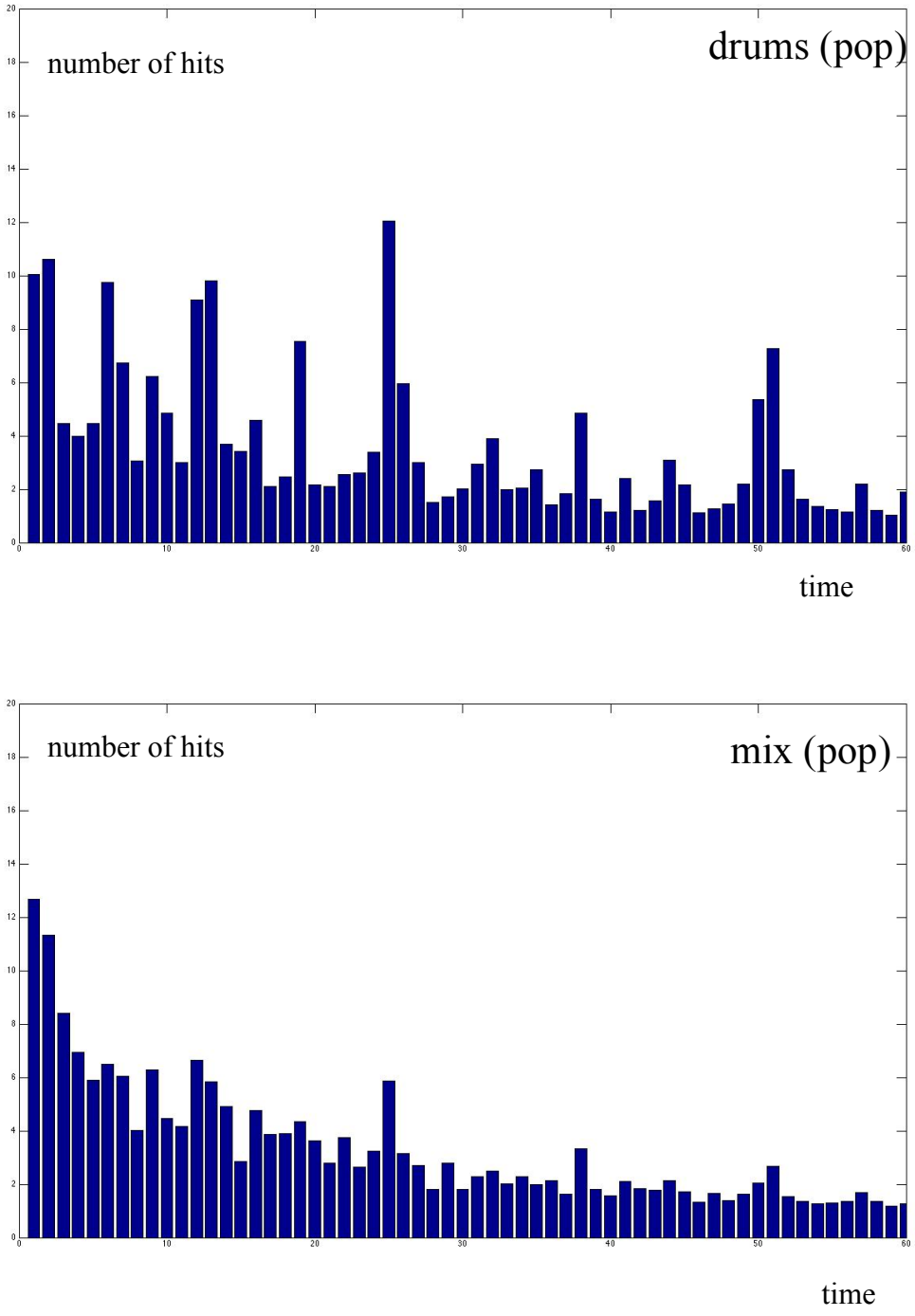



*Figure 4.23 Chromagram calculated for a single rhythmic track (drums) and the whole mix of pop music*

All examples presented above show how tracks of single instruments or sections enable deeper analysis of music features and are easier to be describe by signal representations and parameters. An approach based on the analysis of single tracks might be very reasonable and efficient for mood recognition because it enables more thorough analysis of particular music elements. Sound separation is an area of studies of many works related to various tasks within MIR. Ule *et al*. [316] created a system for drum beat separation based on Independent Component Analysis. Saragdis and Brown [292] applied Non-Negative Matrix Factorization (NMF) to create a system for transcription of polyphonic music that showed remarkable results on piano music. Helen and Virtanen [105] combined NMF with feature extraction and classification process and achieved promising results in drum beat separation for popular music. Similar techniques were used by Paulus and Virtanen [239] and Moreau and Felxer [212] for drum separation.

Separation of music in multi-pitch material is especially difficult and challenging [101,225]. Separation algorithms usually operate on a spectral analysis basis in order to determine the fundamental tones of individual voices and their harmonics. However, there are a number of technical difficulties to overcome, which are the result of the compromise between time and frequency resolution of the analyzed signal.

Applying drum-beat separation for tempo and key detection shows that the separation into single signals parts (only drums or only harmonic parts) does not necessarily improve the results in comparison with the original signal [282]. Due to that fact, it was considered to use different mixtures of at least two signal representation types (original, drum and/or harmonic). The paper by Rosner *et al*. [159] confirms the assumptions that such a mixture of signals is a promising approach to music classification.

The main principle of the drum separation algorithm is employing a semi-supervised approach based on non-negative matrix factorization (NMF). The general idea of NMF is to separate input audio track into several isolated audio tracks, representing specified components such as rhythmic or melodic part. NMF is an efficient method in the blind separation of drums and melodic parts of music recordings.

It is worth noting that there are challenges such as musical articulation, e.g. tremolo or glissando and/or transients with non-harmonic spectra. Thus, separation of music sources is never perfect [26,136,102,187,294]. Another type of problem involves the overlapping harmonics of individual sounds. This phenomenon makes it difficult to obtain not only the

original timbre of the separate sound sources but even to perceive main characteristics of signals. The solution to these problems is to have all sound sources recorded separately and then mixed, but this requires additional resources, which in most of cases are not available. Even though source separation is subject of recent studies [159,292], the tools dedicated to source separation in music are still under development and are very difficult to use without knowing instrumentation of the piece.

Source separation can be enhanced by choosing the frequency band where particular instrument has the fundamental harmonic or majority of the energy. The success of this approach depends strongly not only on the instrumentation but also on spectral properties of sound, which is characteristic for a given music genre or esthetics. For example in metal music bass plays in parallel with guitars, which also have a lot of content in lower frequencies to achieve intense impact. Therefore single bass line is not easy to extract from this kind of arrangement.

### 4.6.2 Proposed Time-Based Parameters

The idea behind the proposed time-based parameters (TBP) was to describe rhythmic content in separate sub-bands. As a sound engineer, the author had an opportunity to consider the role of particular elements of rhythmic section: i.e. kick drum, snare drum and cymbals. Instruments listed above (kick drum, snare drum and cymbals) often exist in specific ranges, respectively in low frequencies (70-200 Hz), mid-range (200-2500 Hz) and high frequencies (above 2500 kHz). At the same time these bands are too capacious to enable an analysis of details. Therefore parameters were calculated in 25 narrow sub-bands and their center frequencies are listed in Tab. 4.10. TBF are named accordingly to the number of the band; that is from TBF1 to TBF25.

The calculation of parameters started with spectrograms with desired frequency and time resolution related to most dense rhythmic arrangements in Western music: i.e. 0.2 sec, to be able to achieve final resolution of 0.4 sec. Then logarithm of the achieved matrix and STFT were calculated. As a result, a matrix where rows represent band and columns represent time steps was achieved. For each band the maximum peak is found and related time value is returned. The final outcome for one audio track is presented in a form of 25 values of maximum energy and corresponding time value. The schema of the calculation process is presented in Fig. 4.24.

*Table 4.10 Frequency bands used in analysis*

| Band No. | Center Frequency [Hz] |
|---|---|
| 1 | 14 |
| 2 | 28 |
| 3 | 42 |
| 4 | 56 |
| 5 | 74 |
| 6 | 96 |
| 7 | 124 |
| 8 | 159 |
| 9 | 204 |
| 10 | 260 |
| 11 | 330 |
| 12 | 419 |
| 13 | 531 |
| 14 | 672 |
| 15 | 850 |
| 16 | 1075 |
| 17 | 1357 |
| 18 | 1713 |
| 19 | 2161 |
| 20 | 2726 |
| 21 | 3438 |
| 22 | 4335 |
| 23 | 5465 |
| 24 | 6890 |
| 25 | 8684 |

Parameters were tested and designed for prepared tracks consisting of separate hits of kick drum, snare drum and cymbals with a different spectral content. The analysis shows that the rhythmic content is visible in different bands for different timbres of a particular instrument. For example, rhythm characteristics of one kick drum were visible in band No. 5, while for another kick they clearly appeared in band 8. The same situation occurred for all instruments depending on how instrument sounded, how was recorded and mixed. Therefore the idea of three wider sub-bands was introduced and Modified Time-Based Features (MTBF) were created. In each range the maximum value from all bands was selected and it was returned as a final result along with the corresponding time value. This concept is presented in schema in Fig. 4.25. Various ranges of these bands were tested by

the author and the best results were achieved for the setup presented in Tab. 4.11. Modified Time-Based Features used in further analyses are named MTBF1, MTBF2 and MTBF3 according to the frequency ranges specified in Tab. 4.11.

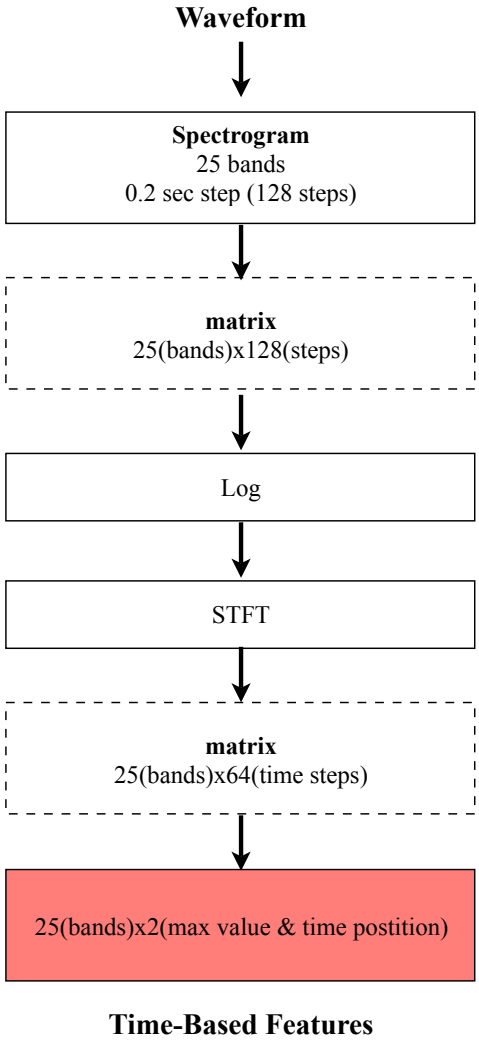



*Figure 4.24 Calculation process of the Time-Based Features (TBF)*

As mentioned before, the proposed parameters were designed and tested using separate tracks, and their performance was satisfying in this configuration. Implementation on the whole mixes returned much more errors but still parameters can describe rhythmic content of different ranges. Thus parameters were included in the feature vector used as an input to artificial intelligence methods (Section 7.5).

*Table 4.11 Frequency ranges used for MTBF analysis*

| Frequency range | Bands |
|---|---|
| I | 5-10 |
| II | 11-19 |
| III | 20-24 |

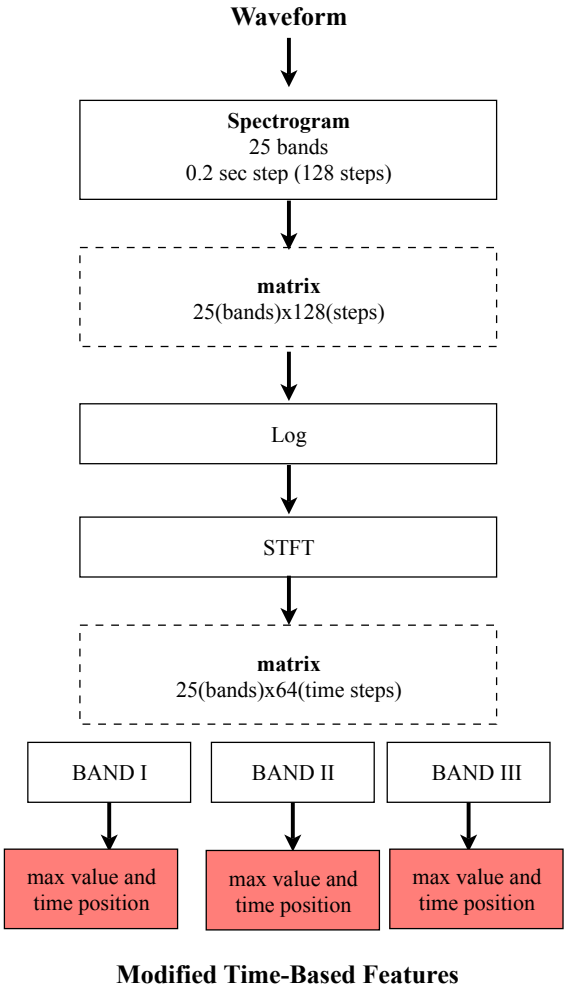



*Figure 4.25 Calculation process of the proposed Modified Time-Based features (MTBF)*

### 4.6.3 MIR Toolbox- Based Additional Parameters Based on Music Characteristics

The idea of parameters describing particular music characteristics was introduced in Section 4.1. Also an analysis of music features of single instrument tracks and mixes for different music genres was performed (Section. 4.6.1). Following the idea of finding descriptors of music signal that contain information related to rhythm, tempo, meter, key, harmony, dynamics articulation, parameters proposed in MIR Toolbox (Section 4.4.1) were analyzed for this purpose. Selection of the parameters was based on the findings of Hevner

[108], who studied relationship between music characteristics and emotions in music. Her studies are described in details in Section 4.1.1 and summarized in Tab. 4.2. A list of features chosen from MIR Toolbox is presented in Tab. 4.12. Detailed information about Matlab scripts and settings can be found in study by Lartillot [165]. All selected parameters were calculated for the set of 150 songs used in the key experiment (Chapter 7). Also the correlation analysis was performed to extract significant descriptors in terms of their relationship with mood of music. Parameters correlated with subjective mood evaluation were included in the final vector used as an input to artificial intelligence methods (Section 7.4).

*Table 4.12 List of additional parameters based on music features*

| No. | Parameter | Description |
|---|---|---|
| 1 | mirtempo | Estimates the tempo by detecting periodicities from the onset detection curve. |
| 2 | mireventdensity | Estimates the average frequency of events, i.e., the number of note onsets per second. |
| 3 | mirpulseclarity [165] | Estimates the rhythmic clarity, indicating the strength of the beats estimated by the mirtempo function. |
| 4 | mirattacktime | Temporal duration of attack time. |
| 5 | mirattackslope | Description of the attack phase is related to its average slope. |
| 6 | mirattackleap | Estimation of the amplitude difference between the beginning and the end of the attack phase. |
| 7 | mirrolloff | The frequency such that a certain fraction (here 0.85) of the total energy is contained below that frequency. |
| 8 | mirbrightness | The amount of energy above cut-off frequency. |
| 9 | mirroughness | An estimation of the sensory dissonance (roughness), depending on the frequency ratio of each pair of sinusoids. |
| 10 | mirregularity | The irregularity of a spectrum is the degree of variation of the successive peaks of the spectrum. |
| 11 | mirkeystrength | Strength of music key based on a cross-correlation of the chromagram. |
| 13 | mirinharmonicity | The amount of partials that are not multiples of the fundamental frequency. |
| 14 | mirchromagram | Shows the distribution of energy along the pitches or pitch classes. Characteristics such as centroid, averaged pitch, etc. are calculated. |

# 5 ANALYSIS METHODS

During the course of presented study, several statistical and computational methods were employed. Moreover, due to interdisciplinary character of the research on emotions in music, traditional statistical methods are accompanied by psychology-based approaches. This Chapter is organized as follows: analysis methods are presented along with references to practical implementation of particular method within the music technology. Common issues (i.e. normalization and correlation) are briefly introduced, while more sophisticated methods such as Multidimensional Scaling, Artificial Neural Networks, Principal Component Analysis or Self-Organizing Maps, are discussed in more detail.

## 5.1 NORMALIZATION

Usually, data values within a dataset may differ widely. That is why normalization is often applied. In essence, normalization is performed to have the same range of values. This process is especially important when data processing by computational methods that involve various parameters. Normalization aligns the importance of each parameter within the data set by assigning the same range. Normalization brings all of the variables into proportion with one another. This process is recommended and sometimes required for methods such as Artificial Neural Networks, Principal Components Analysis, Fuzzy Logic, various classification methods and many others [303,320]. Features normalization is also known as data scaling and is generally performed during the data pre-processing step.

Several types of the normalization exist. They differ in function and structure depending on the normalization range as well as the center point and are chosen accordingly to the nature of the data [82].

### 5.1.1 Normalization I

The most basic and commonly used normalization method is linear unity-based normalization that transform data values into range [0,1]:

$$x_{normI} = \frac{x - x_{\max}}{x_{\max} - x_{\min}} \qquad (5.1)$$

where $x_{normI}$ is the value after normalization I, *min* is the minimum value and *max* is the maximum value of a particular feature. It rescales the data from original values into a desired range. The range can be set arbitrarily to another range, but in most of cases it covers unity range.

### 5.1.2 Normalization II

In a two-steps normalization, data after Normalization I are rescaled, so the average value becomes 0.5 (linear normalization in two ranges):

$$x_{normII} = \begin{cases} 0.5\dfrac{x_{normI}}{avg_{normI}}, x_{normI} \in [0,0.5) \\ 0.5\dfrac{x_{norm1} + 1 - 2avg_{normI}}{1 - avg_{normI}}, x_{normI} \in (0.5,1] \end{cases} \quad (5.2)$$

where $x_{normII}$ is the value after normalization II, and $avg_{normI}$ is the average value of a particular parameter after normalization I.

### 5.1.3 Normalization with Centralized Data

For the cases, where data are distributed symmetrically with the central point in 0, normalization into [-1,1] range is recommended. This procedure is given by the following Equation:

$$x_{norm[-1,1]} = \frac{x - \dfrac{x_{\max} + x_{\min}}{2}}{\dfrac{x_{\max} - x_{\min}}{2}} \quad (5.3)$$

### 5.1.4 Standardization (Z-score Normalization)

The special case of normalization is standardization, where mean and standard deviation for each feature are used for the normalization process. Z-Score is calculated to standardize the data, and it reflects how many standard deviations data fall from the mean (i.e. the variation of across the standard normal distribution. It is computed by substracting from each single point the mean value and dividing the result by the standard deviaiton of the feature values.

$$x_{normσ} = \frac{x - \mu}{\sigma_x} \quad (5.4)$$

where $\boldsymbol{\sigma_x}$ is a standard deviation and $\boldsymbol{\mu_x}$ is a mean value of particular feature.

Standardization is suitable for the data where all of the features are normally distributed.

### 5.1.5 Normalization of the SYNAT Database

SYNAT database consists of various types of parameters (described in detail in Section 4.5), not all of the parameters are from the same domain. Therefore the whole set of the parameters was normalized to range [0,1]. Each of the 173 parameters was normalized separately. Normalization I and II were tested in the carried out research work (Section 6.1.2). Only these two types of normalization were applied because there were the most appropriate for the nature of the data gathered.

The important point is that parameters were normalized according to maximum and minimum parameter values, which occurred in the SYNAT database. Therefore, it is possible that parameters of a song not included in this database may exceed range [0,1]. The SYNAT database contains more than 50,000 of music pieces from different time and genres, thus this kind of normalization seems to be reasonable.

Results of the correlation analysis (presented in Chapter 6, Tab. 6.6 and 6.7) showed that the type of normalization does not have significant influence on the results. Therefore in the next study stages only Normalization I was used.

Once the normalization process was completed, only normalized parameters were used at all stages of the research. New parameters added to the vector either were defined not to exceed the desired range, or were normalized based on the values achieved for the whole SYNAT database. This will be presented in Chapter 6.

## 5.2 CORRELATION ANALYSIS

Correlation coefficient $\boldsymbol{\rho}$ is used to determine whether two qualitative variables are dependent [84]. It is defined as a covariance of two variables *X* and *Y* divided by the product of their standard deviations [320]:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{5.5}$$

where ***cov* is covariance and $\boldsymbol{\sigma}$ is** standard deviation.

Covariance is an expected value $\varepsilon$ of the product of the deviations of two random variables $X$ and $Y$, from their respective means $\mu_X$ and $\mu_Y$ and is given by the following formula [84]:

$$\mathrm{cov}(X,Y) = \varepsilon(x-\mu_X)(y-\mu_Y) \tag{5.6}$$

Correlation is commonly employed to determine dependent factors (i.e. subjective evaluation results and parameters) and is widely used in the area of music technology [110,156,191].

## 5.3 T-STUDENT TEST

Student's t-distribution is commonly used to assess statistical significance. In the case of the presented study, t-distribution shows whether correlation-based results are significantly different. It is defined as a significance test for assessing hypotheses about population means [84]. t-Student for correlation $\rho_{X,Y}$ between parameters $X$ and $Y$, with $n$ -2 degree of freedom, can be calculated using Equation:

$$t = \frac{\rho_{XY}}{\sqrt{1-\rho_{XY}^2}}\sqrt{n-2} \tag{5.7}$$

t-Student is dedicated to small trials, arbitrarily employed when the number of samples $n$ is ≤30 [320] and has been successfully used numerous times as a statistical tool for analysis of audio parameters [110,156,175,232].

## 5.4 MULTIDIMENSIONAL SCALING ANALYSIS

Multidimensional Scaling (MDS) was firstly developed in the area of psychology. The key problem encountered in this field is the recognition of "underlying dimensions" that would explain similarities or dissimilarities observed by subjects. In the study by Borg *et al.* [38] 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 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. The Euclidean distance of points *i* and *j* in *m*-dimensional space X is calculated from:

$$d_{ij} = \sqrt{\sum_{a=1}^{m} (x_{ia} - x_{ja})^2} \tag{5.8}$$

MDS maps the proximities $p_{ij}$ (similarities obtained from subjects) into corresponding distances $d_{ij}$ in MDS space X:

$$f(p_{ij}) = d_{ij}(X) \tag{5.9}$$

Since (depending on the data) the exact representation does not always exist, there is a need to define a value, which reflects error of the map:

$$Stress-1 = \sigma_1 = \frac{\sum_{(i,j)} [f(p_{ij}) - d_{ij}(X)]}{\sum d_{ij}^2(X)} \tag{5.10}$$

While the rules presented above apply to the main concept of Multidimensional Scaling, several different types of MDS models are defined. Only within the metric approach, transformations such as absolute, ratio, interval and others are distinguished. The basic idea of a few of them is presented in Fig. 5.1 [38]. They represent various properties of the data related to algebraic operations (addition, subtraction, multiplication, division).

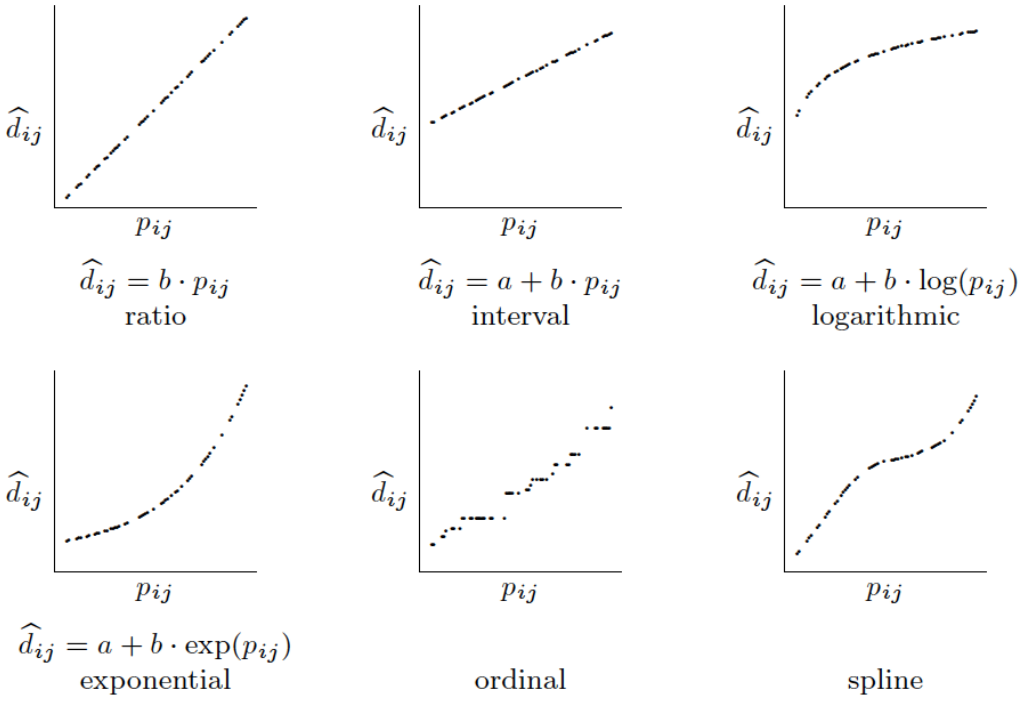



*Figure 5.1 Transformation plot of several transformations. [38]*

In contrast, nonmetric models represent only the ordinal properties of the data. Ordinal models typically require the following condition:

$$\text{if } p_{ij} \le p_{kl}, \text{ then } d_{ij} \le d_{kl} \qquad (5.11)$$

Metric models also lead to distances ordered in the same way as the corresponding proximities. But they are all special cases of non-metric model, where no particular function $\boldsymbol{f}$ is required for the monotone relation. Chosen MDS model should correspond to the nature of analyzed problem and especially to the methodology of the subjective.

Multidimensional Scaling was used in various applications related to sound and music. Trochidis et al. [310] applied MDS to analyze similarity within the wide area of Western music. They determined two main dimensions and created graphical representation of 27 musical excerpts on 2D plane. Wagenaars used multidimensional scaling test to determine optimum values of compression in music [325]. Małecki found three dimensions that determine perception of similarity between acoustics of different rooms. MDS was also applied to emotional responses to music [32]. Bigand et. al. [32] stated that 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 research and the presented Ph.D. work, therefore results may be different. While the aim of presented work is to analyze the audio content to determine mood included in audio signal, Bigand asked listeners specifically " to focus their attention on their private emotional experience". Additional differences occurred during data collection. Bigand [32] asked participants to look for excerpts that induced similar emotional experience and to drag the corresponding icons in order to group these excerpts. The author of the presented thesis collected direct similarity data and introduced application of MDS to Music Mood Recognition [243]. Differences in the collection data approach might cause significant inconsistency. The MDS experiment reported in [243] is described in details in Section 6.4.

## 5.5 ARTIFICIAL NEURAL NETWORKS (ANN)

Artificial Neural Networks are a family of statistical learning models inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Neural networks are very important tool providing a wide range of solutions, especially for tasks such as classification

and cluster analysis of data [12,145,304,353]. They are commonly used especially within the area of speech and image recognition or optical character recognition, but became also popular in musical sounds classification and mapping [44,154,211,214,227].

ANN are very useful tools for tasks with high complexity and low knowledge about the rules [303, Rauben]. Perception is one of the common problems with unknown principles or with low knowledge of principles. ANN can only solve issues where the declarative memory is involved. Based on the previous experiences, they can help predicting the reaction or the repercussion of the action with unknown rules. They are commonly used in various areas of industry and research such as: identification of military objects based on their noise, image recognition, sound separation and many others [61,150,304].

Artificial Neural Networks can be divided into three main subgroups:

- Feedforward Networks
- Recurrent Networks
- Self-Organizing Maps (SOM).

In the presented dissertation thesis, only feedforward and SOM networks were implemented; therefore Recurrent Networks are adduced roughly to maintain the methodology and structure of the subject.

### 5.5.1 Feedforward networks

Feedforward network are structures where the information moves in only one direction, forward, from the input nodes, where the input data (tasks) are delivered, through the hidden nodes (if any) and to the output nodes, where the solution in a form of output data is received. These networks are the most commonly used and the most helpful of the artificial intelligence tools. The structure and operation of feedforward networks is based on a set of artificial neurons. A single neuron consists of inputs, processing element and a single output (Fig. 5.2).

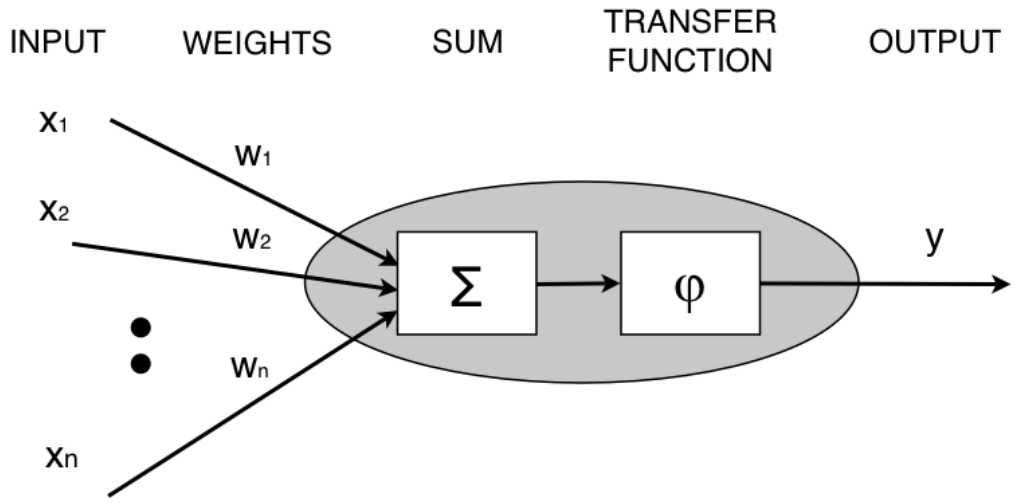



*Figure 5.2 Schema of a single neuron.*

The simplest case is the linear neuron where, the output is calculated as following:

$$y = \sum_{i=1}^{n} w_i x_i \tag{5.12}$$

where $x_i (i=1,2,...,n) \in [-1,1]$ is vector of inputs, $w_i (i=1,2,...,n)$ are synaptic weights and $y \in [-1,1]$ is value on the output. All weighted signals $w_i x_i$ are summed in the summing component and directed to the output. In this case transfer function is linear and is included in weights values.

More sophisticated form of a single neuron is perceptron, where non-linear transition function **φ**, also called activation function, is involved. The output of the perceptron is given by the following formula:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i\right) \tag{5.13}$$

The most commonly used transfer functions are sigmoid, its variations and occasionally other functions i.e. unipolar, bipolar, hyperbolic tangent, Gaussian etc. [150,257], with the assumption that the activation function has to be differentiable. Examples of different activation functions are presented in Fig. 5.3.

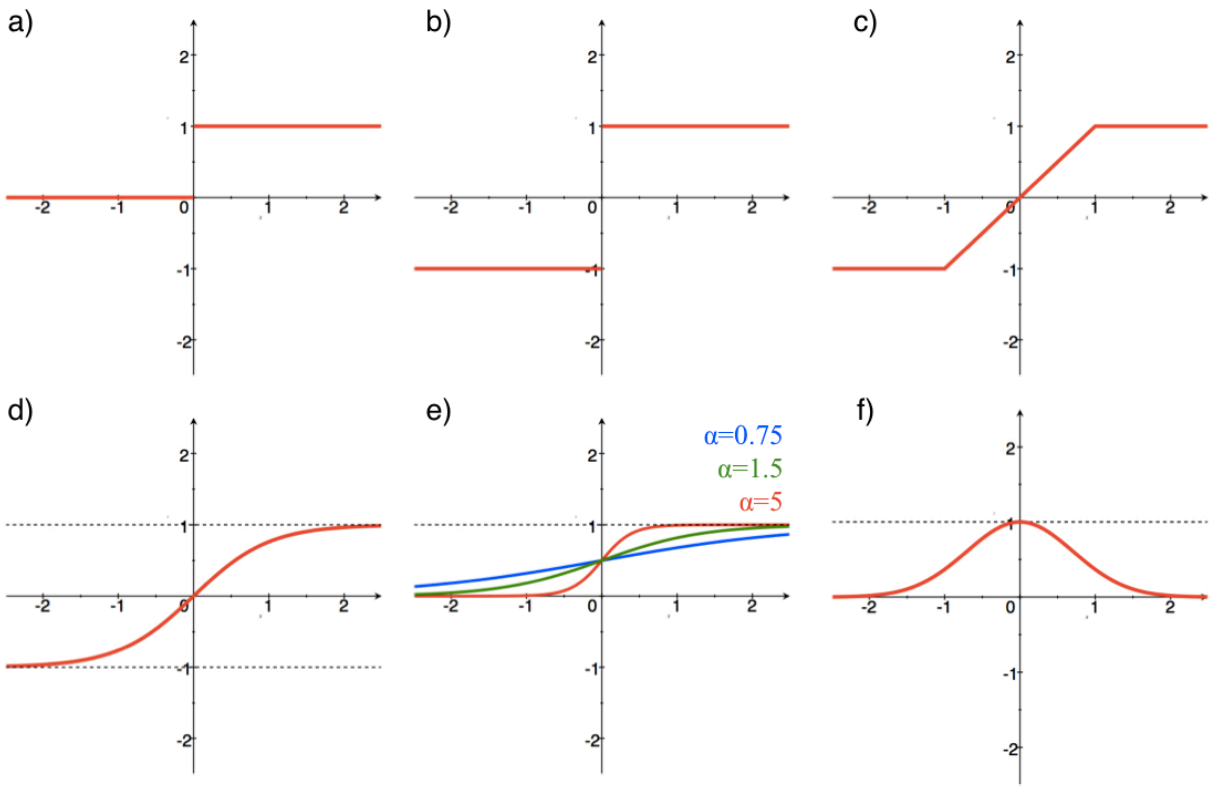



*Figure 5.3 Examples of transfer functions used in neural networks: a) unipolar binary, b) bipolar binary, c) bipolar threshold linear, d) hyperbolic tangent, e) sigmoid with different values of α, f) Gaussian.*

The sigmoidal transfer function is described by the following formula:

$$f(x) = \frac{1}{1+e^{-\alpha x}} \tag{5.14}$$

where $\boldsymbol{\alpha}$ is slope coefficient. Sigmoidal functions with different $\boldsymbol{\alpha}$ are presented in Fig. 5.3e.

Neurons with linear or non-linear activation function are connected into networks. Although it is possible to create a network with very complex structure, most of implemented networks have layer arrangement. Example of the 2-layer neural network (with one hidden layer) is shown in Fig. 5.4. The input layer includes only weight values and distributes the weighted input values into the first layer of neurons, where the signals are summed and re-calculated trough the activation function $\boldsymbol{\varphi}$. Then signals are weighted with another set of weights and distributed into output layer, where they are summed, re-calculated and directed to outputs.

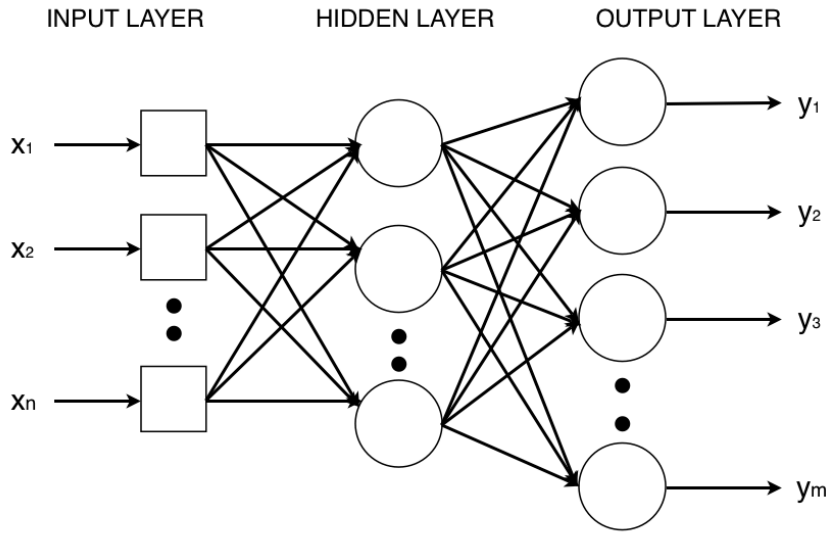



*Figure 5.4 Schema of feedforward neural network with one hidden layer.*

Neural networks can form two different types of models: regression and classification. As an output of a regression model, solution in a form of an objective numerical value is expected. Tasks defined as classification problems, seem to be more successfully solved with ANN [303]. Therefore even problems with expected numerical value should be rather defined as a set of sub-ranges, with one neuron assigned to each. At this point another decision has to be done, whether one network with multiple outputs should be defined or multiple networks with a single output each. Neuron with the highest value is interpreted as a winning one and range assigned to this unit is treated as a solution of particular question. There is no unequivocal answer to the structure issue and the choice has to be made accordingly to the characteristics of analyzed phenomenon [304].

The non-winning output values for the classification problem can obtain values different than 0, what is one of the most interesting and useful characteristics of neural networks. This fact can be a good starting point for joint action with fuzzy logic.

In the classification tasks ANN can find the features that allow the solution, even though the sense and meaning of a particular characteristic is unknown or concealed in the hidden layer.

The network structure impact on the ability to solve problems is not crucial. Nevertheless unconsidered and random choices can cause less efficient and more difficult learning process. [304]. However, the size of the network should carefully be considered. Too big network will not be able to generalize, while too small might not have enough "intellectual potential" to solve particular problem.

In most of cases, networks are trained under supervision. The learning process can strongly depends on the default weights, because their primary values can (but not necessary have to) cause failure.

To enable learning model has to be supplemented with two additional elements: mechanism of changing weights and error detector. Neuron in this form is called ADAptive LiNear Element (ADALINE). In supervised learning, algorithm is given a set of example pairs (x, y), and the aim is to find a function *f:*

$$f(x) = y \tag{5.15}$$

that matches the examples. The cost function is related to the mismatch between mapping and the data and it contains prior knowledge about the problem domain.

A commonly used cost function is the mean-squared error, which tries to minimize the average squared error between the network's output *f(x)* and the target value *y* for all elements of training set.

There are several methods of training feedforward networks. The most significant methods are listed in Tab 5.1 [85,107,145,150,254,303,353]. Further details on each of them can be found in adequate literature sources. Recognizing the errors in the hidden layer is the crucial part of the method used in the learning process.

*Table 5.1 Selected methods of supervised training feedforward networks with corresponding references*

| **Selected Methods of Supervised Training Feedforward Networks** |
|---|
| Backpropagation [303] |
| Conjugate gradient method [303] |
| Heuristic Algorithms [150] |
| Quickprop Algorithm [85] |
| Rprop Algorithm [261] |
| Levenberg-Marquardt Method [303] |

Regardless of numerous different methods of learning, one has to choose the appropriate learning rate $\boldsymbol{\alpha}$ as well as the value of momentum $\eta$. Both of these factors have large impact on the training process and their properly adjusted value can determine either successful or unsuccessful result. To avoid enormous amount of input data, pre-processing and adequate selection of the signals should occur.

One-layer networks can also take a part in unsupervised training. Most of methods dedicated to this application are based on the competition and using Hebb rule

[150,304,353]. The network dedicated to unsupervised learning has to contain at least three times more neurons in output layer than expected answers [304]. The training set has to include groups of similar objects that can be classified. Unsupervised training occurs only if behind the training set stays the regularity, on which the network may base. The competition makes the learning process more effective, the best results are achieved for medium-size networks [304].

For classification tasks, typically the data set is split into three parts: training set, validation set and testing set [196,303]. Elements from training part are presented to the network during training, and the network is adjusted according to its error. Validation set is used to measure network generalization, and to finish training when generalization stops improving. Finally testing set has no effect on training but provides an independent measure of network performance during and after training. The division process is random and commonly shares are distributed as following: training (70%), validation (15%) and testing (15%). These values can vary depending on the data set.

**Feedforward networks applied to music technology**

There are numerous works related to research that involves ANN into musical sound classification (i.e. [52,79,129,157] and many others). Kostek [149] showed results of the experiment where groups of instruments (strings, woodwinds and brass) and 10 single instruments were classified and high effectiveness was achieved. The effectiveness was 97.6% for single instruments classification and 96% for groups of instruments (both for two-stage neural network algorithm). Different configuration of descriptors and configurations of neural networks were tested by Kostek and her team [148,302] and accuracy was always clearly above 90%.

Kaminskyj [129] developed systems based on Artificial Neural Networks (ANN) as well as K-Nearest Neighbor Classifier and compared them in terms of efficiency. In the classification task that involved 4 instruments (guitar, piano, marimba and accordion within one octave), KNN and ANN achieved very good results (max. 98.1% for KNN and max. 97.7% for ANN). The surprising conclusion of his work was that only temporal, but not frequency, data were utilized. Kaminskyj broadened his work and achieved accuracy of 93% in instrument recognition, 97% in instrument family recognition, and 100% for sustain/impulsive instruments within a group of 19 instruments and 3 octaves (C3–C6).

Cemgil and Gürgen [52] used STFT to classification of 40 sounds within one octave (C3-C4) from 10 instrument groups based on ANN, recurrent networks and SOM. The best results were achieved for ANN and recurrent network with efficiency rate above 97%.

Eronen [79] performed *k*NN classification of 30 orchestral instruments (within the full pitch ranges) from the string, brass and woodwind families, played with different techniques. The correct instrument family was recognized with 94% accuracy and individual instruments in 80% of cases. His vector of parameters consisted mostly of **cepstral coefficients and temporal features.** He continued his work with larger set of features and published in the research studies by Eronen [79].

Żwan [354] conducted experiments aimed for automated classification of singing voices. Neural network was trained to determine voice quality as well as voice type (bass, baritone, tenor, alto, mezzo-soprano, and soprano). Żwan achieved approximate accuracy of 90%.

Another type of neural network application is analysis of sound spaciousness. Palomaki [231] trained multilayer perceptron with localization cues computed using a binaural model.

Artificial Neural Networks can also support recognition of musical structures such as phrase, rhythm and harmony, as well as prediction of musical elements (melody, rhythm and harmony) [301]. In the area of psychoacoustics ANN was used to determine the auditory noise-masking threshold created by input data [214].

### 5.5.2 Recurrent Networks

Feedback from the outputs at the several stages of the processing is a characteristic attribute of recurrent networks. Connections create numerous and complex loops, where signal is floating until it achieves (if possible) the steady state [304]. In consequence, the output signal depends not only on current outputs but also the whole history of stimulation. Example of recurrent network is presented in Fig. 5.5.

Various topologies are widely described in literature dedicated to neural networks [61,77,257,304]. To name few interesting structures: Hopfield network (with peer to peer connections) [304], Elman's structure [78] (with signals from first hidden layer are delayed and fed into input), Jordan's network (context units are fed from the output layer) [61] and many others. Recurrent networks are successful with time sequences processing and are employed for time-consuming and complex optimization processes [150,257].

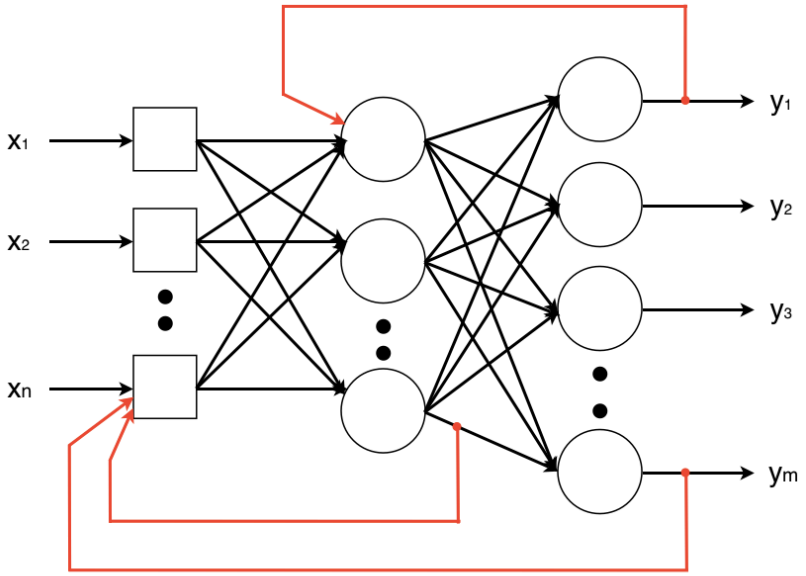



*Figure 5.5 Example of feedback network.*

**Recurrent networks applied to music technology**

Recurrent networks were employed to create models of string instruments [176]. They were also used to support struck strings instrument synthesis. [54] Chang and Su's work attempts to automatically extract the synthesis parameters by using a neural-network training algorithm without the knowledge of physical properties of the instruments. Mion [204] involved Bayesian networks into automatic recognition of expressive content of piano improvisations. Cemgin and Gürgen [52] successfully used recurrent networks (Time Delay Neural Network) for classification of instrument sound with accuracy of 97%.

Another type of problems where recurrent network that was involved is automatic identification of a sound source position [63,109,152].

### 5.5.3 Self-Organizing Maps (SOM)

The SOM (Self-Organizing Map) is an unsupervised neural network providing a mapping from a high-dimensional space to few-dimensional (in most of cases two-dimensional $K \times L$) representation [257,317]. The topological relations between objects are preserved as detailed as possible. This self-organized process is called Voronoi mosaic and neurons specialize in detecting and signalization of different groups of input signals. Presented objects are grouped accordingly to the similarity. The SOM consists of 2-dimensional grid of neurons, with a weight vector related to each unit (Fig. 5.6). SOM is forced by vector $x_1, x_2, \ldots,$

and activation $y_1, y_2,..., y_m$ for each neuron unit for the presented object is calculated. This type of networks was introduced by Kohonen [144].

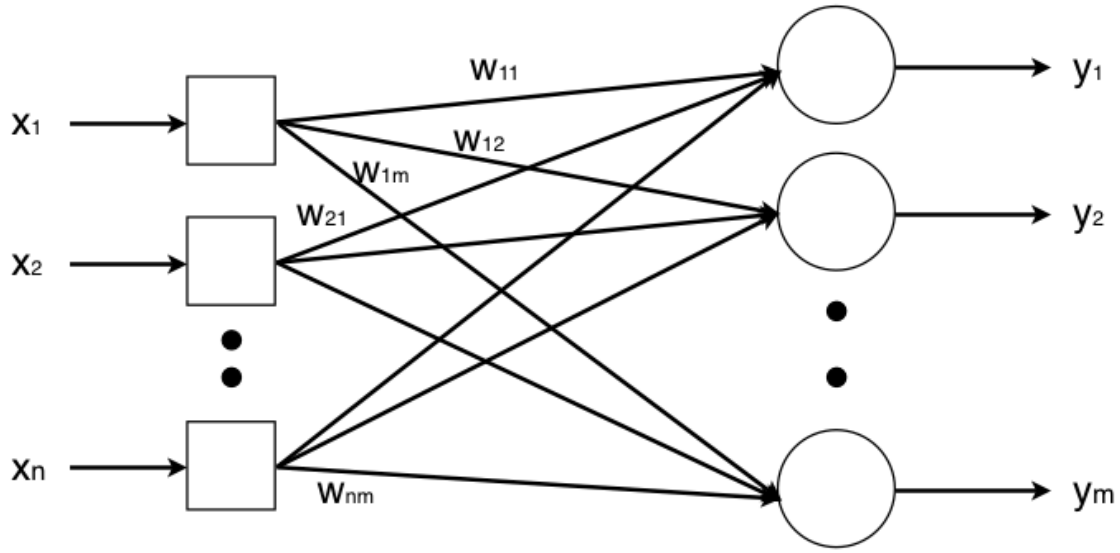



*Figure 5.6 Schema of the SOM network.*

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" [196,254]. The winning neuron *a* is selected from the *K×L* network consisting of *i* elements, according to the following relation:

$$d(x, w^{(a)}) = \min_{1 \le m \le K \times L} d(x, w^{(m)}) \tag{5.16}$$

where *d* is a measure of distance between *n*-dimensional vector *x* and the weights vector *w* of the output vector in *K×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:

$$w_i^{(a)(j+1)} = w_i^{(a)(j)} + \eta^{(j)} \left[ \left( x_i^{(j)} - w_i^{(a)(j)} \right) \right] \tag{5.17}$$

*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" [196,254,255, 303].

$$\forall_{m \in N_a} w_i^{(m)(j+1)} = w_i^{(m)(j)} + \eta^{(j)} h[x^{(j)} - w_i^{(m)(j)}] \tag{5.18}$$

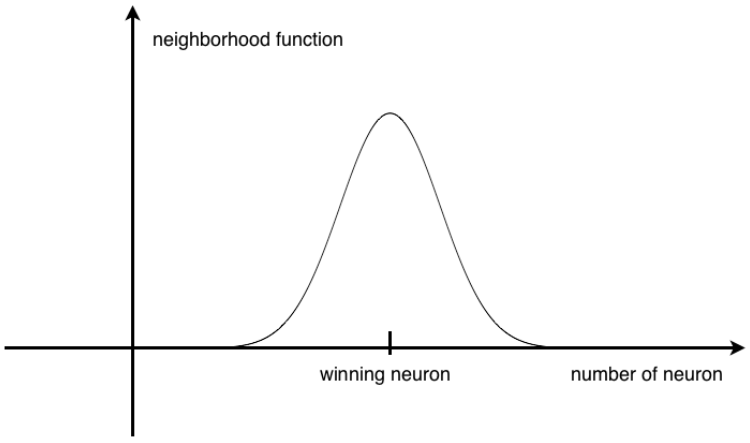
where $N_a$ is a set of units adjacent of the winning unit *a*, and *h* is the neighborhood function of neuron *a*. Neighborhood function can vary from simple to complex functions. Therefore *N* can include i.e. only connected units or the influence on the weights of the neurons can shaped accordingly to Gaussian function (Fig. 5.7).

Both speed of learning $\eta$ and neighborhood function *h* are changing during learning process and monotonically decrease to avoid compensation at the final stage of the learning process.

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

- Competitive learning: the prototype vector most similar to a 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.

*Figure 5.7 Example of Gaussian neighborhood function* ***h***

The learning process can vary, depending on the architecture of the network, default weight values and a training set. There are numerous conditions that have to be completed so the network can function properly. The size of the network should be appropriate for the nature of the problem. Too small network might not represent well the details of the problem, while too big one, might not be able to generalize. The learning set should contain objects from the whole range of the features. Absence of objects in particular are causes that this region is not covered by neurons, therefore in case of occurrence, network will not be able to assign the correct unit. Also default weights settings and the topology of the output layer can affect the course of the training process. Weight values are rather random and only their range should be adjusted, so the process is faster and stable. The topology of the output layer and can be arranged on i.e. a rectangular, hexagonal or random lattice (Fig. 5.8). That determines the number of connections of a single neuron. Useful extensions include using toroid grids where opposite edges are connected.

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.

While supervised neural networks need the external judgment, SOMs are strongly related to human perception. SOM are dedicated to complicated tasks, where rules are not much known or unknown i.e. cluster analysis, creation of models and mapping features. They are successfully used in the areas such as medicine, economy and image recognition.

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 i.e [232,233,254], also taking into account the music genre [255].

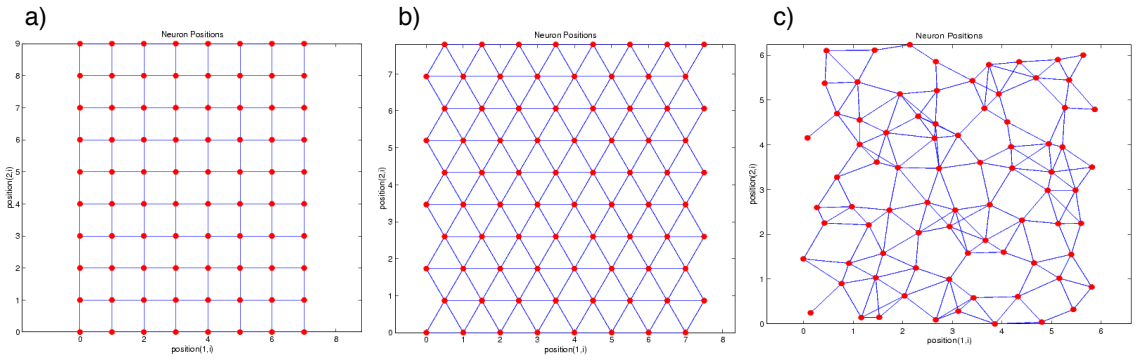



*Figure 5.8* *Examples of Self-organizing Map topologies: a) rectangular, b) hexagonal, c) random. Red circles represent neurons and blue lines represent connections between units*

There are several works (i.e. [2,52,350] and others) where SOMs are used for musical instrument sound classification. Zhang [350] achieved effectiveness of 80%, within a

polyphonic music with a dominant instrument. Using discrete STFT, Cemgin and Gürgen [52] combined SOM with RBF network, what enabled them to achieve classification accuracy of 91%, but also gave additional information how sounds are organized according to timbre.

The aim of Palomaki's work [231] 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 [313] created system for reduction of impulsive noises based on SOM.

Barbedo [19] proposed a Cognitive Model for Objective Assessment of Audio Quality. His 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. Feiten [87] introduced classification of musical instrument sounds based on STFT, involving hybrid self-organizing maps (SOM-RBF) as well as another types of neural networks (MLP and TDNN). Rauber *et al*. [254] from Vienna University of Technology proposed a SOM-enhanced JukeBox (SOMeJB) system [89] to organize their music database analogically to the text library. The classification is mostly content-based and genre-based. The idea developed and system that automatically organizes any music collection according to music similarity was presented by their inventors [255]. Introduced system consisted of 2-dimensional SOM representation that could be generated for any music set. More complex variation involved GHSOM (Growing Hierarchical Self-Organizing Maps) with 3-layer architecture [256]. GHSOM was fed with 1200 psychoacoustic loudness and rhythm descriptors. Simplified concept is presented in Fig. 5.9 and examples of maps are shown in Figs. 5.10 and 5.11. It is worth noting that the organization does not follow clean “conceptual” genre styles but rather reflects the overall sound similarity.

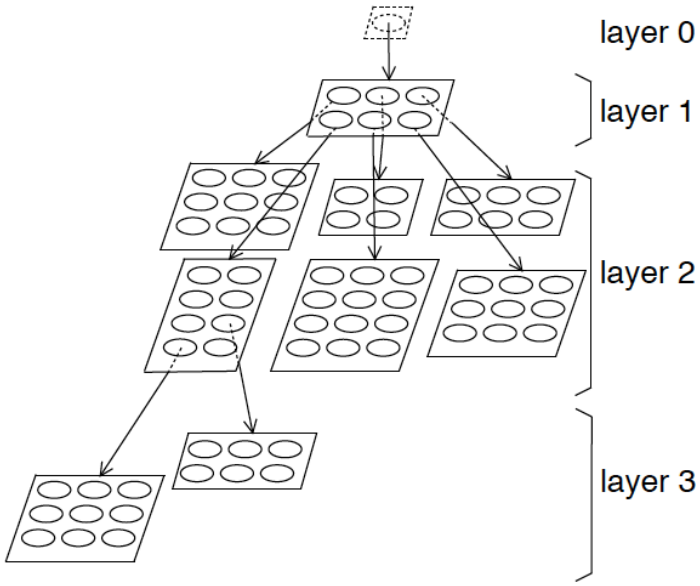



*Figure 5.9* *GHSOM architecture used for music database representation [256]*

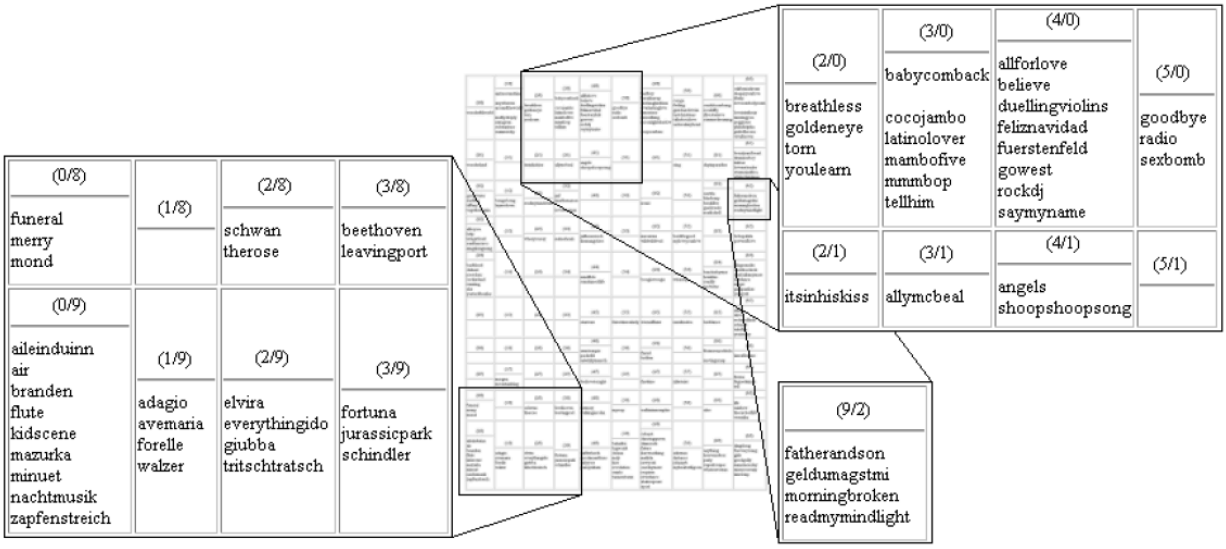



*Figure 5.10* *SOM representing 230 pieces of music [254]*

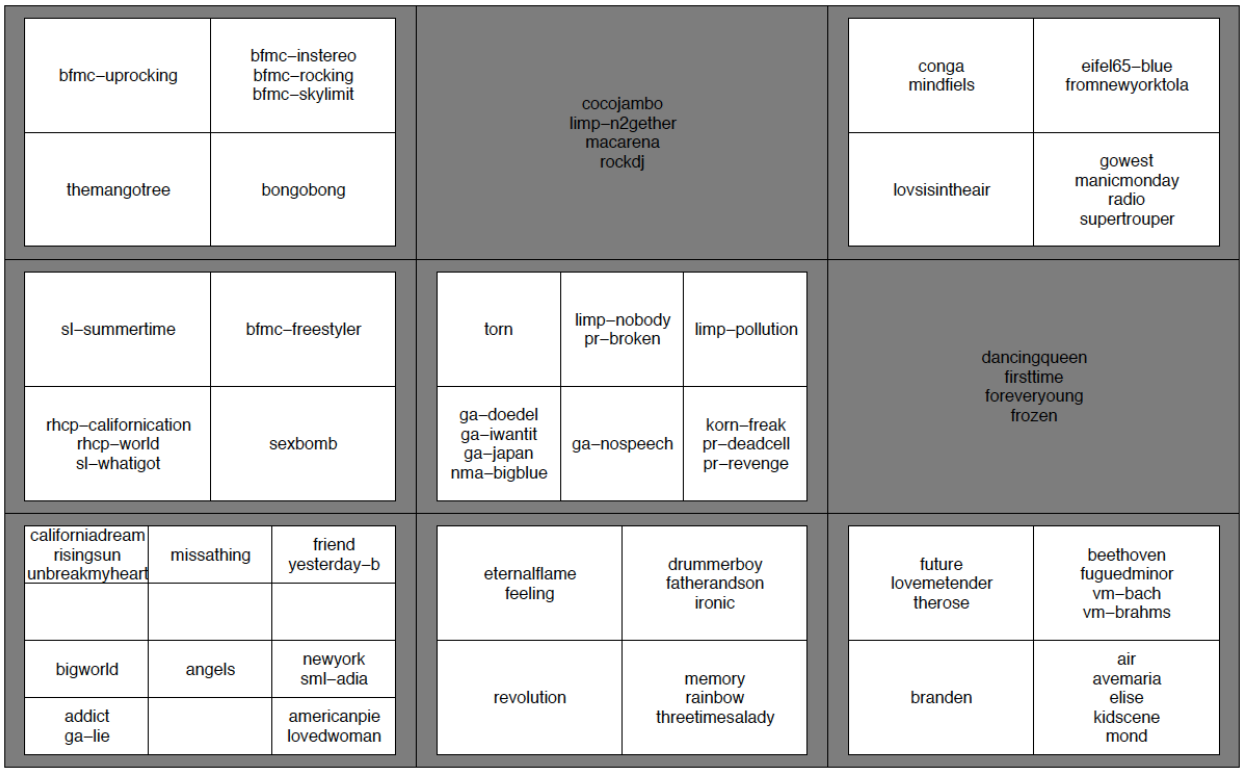



*Figure 5.11 A GHSOM of 77 pieces of music [256]*

Pampalk [233] introduced a method to visualize the clusters of a SOM based on smoothed data histograms. His research involved into *Islands of Music* system, where he used metaphor of geographic maps with islands resembling genres or styles of music. His solution is dedicated to exploration of unknown music collections [232]. At the same time Mayer [197] proposed mnemonic shaped SOMs, according to concept that conventional rectangular maps are not satisfactory for memorizing and associations purposes.

Most of the mapping examples presented in this Section base on the general and genre-based similarity between music pieces. Therefore a reasonable step was to apply and perform SOM mapping founded on the features related to mood of music. Results of this concept are presented in this thesis in Section 6.5.

## 5.6 PRINCIPAL COMPONENTS ANALYSIS

Principal Component Analysis method is defined as a procedure for analyzing multivariate data, which transforms the original variables into new ones that are uncorrelated to reduce the dimensionality of the data [84]. PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system [123]. The new variables, called the principal components, are defined as linear functions of the original

variables and are meant to be new, orthogonal dimensions [293]. 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. [123]. Principal Components Analysis is included in various software such as XLStat [340], Simka-P [318], Matlab [195] and is commonly used as a tool for data reduction, also in the area of music technology, where large sets of data are frequently encountered.

Małecki [191] applied Principal Component Analysis to reduce features describing acoustics of the sacral objects. Kaminsky [129], using PCA, reduced 80 elements vector to 3 components (covering 88.8% of total variation) for the purpose of KNN and ANN classification.

## 5.7 FUZZY LOGIC

Fuzzy logic is a logic system that recognizes more than simple true and false values, where the truth values of variables may be any real number between 0 and 1. Fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0). With fuzzy logic, propositions can be represented with degrees of truthfulness and falsehood.

Fuzzy logic has proved to be particularly useful in expert system and other artificial intelligence applications. It is also used in some spell checkers to suggest a list of probable words to replace a misspelled one.

The concept of Fuzzy Logic (FL) was conceived by Lotfi Zadeh [347], a professor at the University of California at Berkley, who was working on the problem of computer understanding of natural language.

Fuzzy logic, the extension of fuzzy set theory, utilizes degrees of truth to determine the nature of a system. In particular, while mathematical variables often are represented by numerical values, fuzzy logic permits the use of linguistic variables [347,348]. These variables may be associated with qualifying terms such as short, fast or poor.

Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths, which we aggregate further into higher truths, which in turn, when certain thresholds are exceeded, cause certain consequence results such as motor

reaction. A similar kind of processing is used in artificial computer neural networks and expert systems.

In fuzzy set theory, a set is described by a pair *(A, μ).* Membership of *k* in the set *(A, μ)* is defined as *μ(x)* an is called the membership grade of x. x is fully included in the set *(A, μ)* if μ(x)=1, and fully excluded from set *(A, μ) if μ(x)=0*. Moreover x is a fuzzy member of the set *(A, μ)* if *μ(x)* is within range (0,1), where fuzzy set allows a member to belong to a set to some partial degree [18]. A comparison between classical and fuzzy sets is presented in Fig. 5.12. In this case traditional sets are marked with red dashed line and represent crisp membership, where each sound pressure level (SPL) belongs to one set (either quiet, medium or loud). On the other hand, fuzzy sets (bold black line) determine fuzzy membership, where i.e. SPL of 50dB is QUIET to high extent and MEDIUM to low extent. This reflects the rules of human's perception, where no crisp threshold is defined, where SPL changes impression of loudness from quiet to medium, but the transition is blurred.

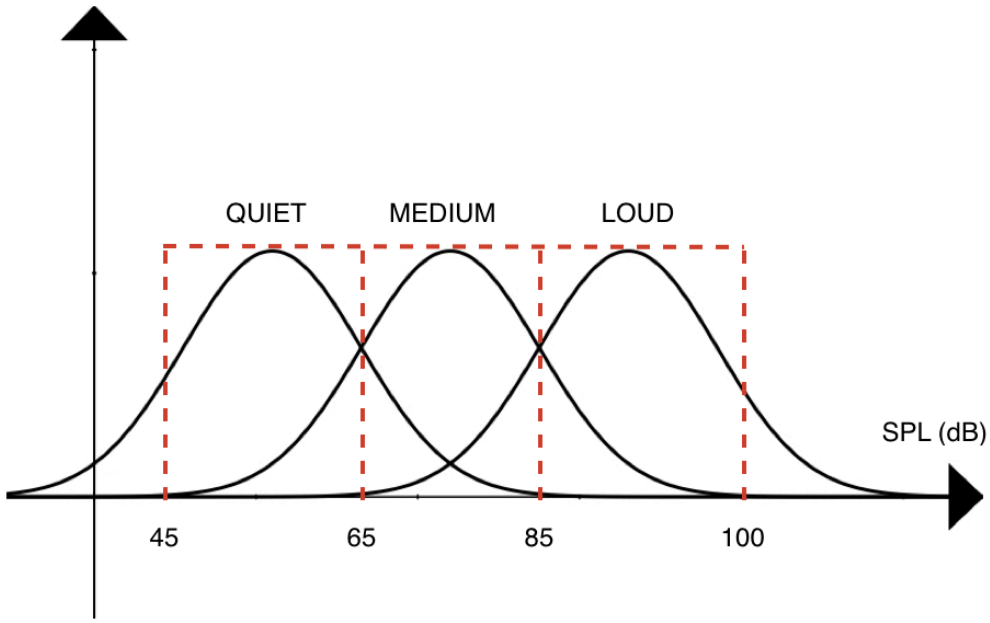



*Figure 5.12 A comparison between classic sets (black bold line) and fuzzy sets (red dashed line)*

Fuzzy set contains elements, which have varying degrees of membership in the set, and this is contrasted with the classical or crisp sets because members of a classical set cannot be members unless their membership is full or complete in that set. A fuzzy set allows a member to have a partial degree of membership and this partial degree membership can be mapped into a function or a universe of membership values. Assume that we have a fuzzy set *A*, and if an element *x* is a member of this fuzzy set *A*, this mapping can be denoted as:

$$\mu_A(x) \in [0,1]$$
$$(A = (x, \mu_A(x) \mid x \in X) \qquad (5.19)$$

and can be read: a fuzzy subset $A$ with an element $x$ has a membership function of $\mu_A(x)$.

Fuzzy control rules can be considered as the knowledge of an expert in any related field of application. The fuzzy rule is represented by a sequence of the form IF-THEN, leading to algorithms describing what action or output should be taken in terms of the currently observed information. The law to design or build a set of fuzzy rules is based on a human's knowledge or experience, which is dependent on particular application. A fuzzy IF-THEN rule associates a condition described using linguistic variables and fuzzy sets to an output or a conclusion. The IF part is mainly used to capture knowledge by using the elastic conditions, and the THEN part can be utilized to give the conclusion or output in linguistic variable form. This IF-THEN rule is widely used by the fuzzy inference system to compute the degree to which the input data matches the condition of a rule. Let's imagine that we would like sound to be accompanied with appropriate light assigned. Here is an example of the rule based on the fuzzy sets presented in Fig. 5.12.

**IF loudness is LOW, THEN light should be DARK.**

For other input SPLs, different rules should be developed. An example of fuzzy rules for sound/lighting system are shown in Tab. 5.2. Rows and columns represent two inputs, respectively: SPL and tempo of music.

*Table 5.2 An example of fuzzy rules for sound/light system*

| | **SLOW TEMPO** | **MEDIUM TEMPO** | **FAST TEMPO** |
|---|---|---|---|
| **QUIET** | DARK | MEDIUM | MEDIUM |
| **MEDIUM** | DARK | MEDIUM | BRIGHT |
| **LOUD** | MEDIUM | BRIGHT | BRIGHT |

All these relations can be presented in a form of rule IF-THEN:

**IF loudness is QUIET, and tempo of music is SLOW, THEN the output (brightness of the lighting system) should be DARK**

Fuzzy logic was employed in several applications related to MIR [157]. Kostek [147] implemented fuzzy control in acoustic organ controlling. The flow of air was determined by

fuzzified inputs from MIDI controller. Rough-Fuzzy Based Classifiers were used for timbre classification as well as pitch assignment [63].

Fuzzy logic was also applied to studies dedicated to emotions in music. Blewitt executed exploration of psychologically grounded theories of emotion in music through fuzzy logic systems [36]. He also constructed two models of emotions based on fuzzy logic along with their computer implementation. Jun and collaborators [125] created music emotion recognition system, where music fragments are analyzed and mapped into VA plane by a fuzzy inference engine. They achieved promising result of average 12% of distance between subjective and automatic VA assignment. Although it is important to note that their system was based on the set of 50 songs evaluated by 5 listeners, which can suggest that formulated rules are rather specific.

It is also interesting that relations between music features and emotions in music described by Hevner are also based on the concept of rules, which can be considered as fuzzy [108]. She summarized her findings related to the music features that create emotional content of music by assigning to each of eight adjectives group as set of music characteristics with weights (Tab. 4.2).

In this Chapter, various computational methods used in the course of presented study were described. It is worth noting that most of them refer to human's cognition and perception (i. e. MDS, ANN, fuzzy logic). Choice of methods is supported by numerous observations related to performed experiments (Chapter 6) as well as nature of the study, which is closely related o psychology and perception. Therefore it is hard to imagine using tools that would not contain "human's" element.

# 6 PRELIMINARY EXPERIMENTS AND ANALYSES

The role of the experiments in Music Emotion Recognition (MER) is very important. Commonly MER is based on the subjective evaluation by an expert [208], group of listeners or even whole social network [162,]. Listening tests enable determination of the listeners' perception of mood of music and are crucial for the verification of automatic mood description. The main approach in the area of Music Emotion Recognition leads to retrieve information from the music signal (although there exist some systems based mostly on tags or social networking). At the same time the main interest is focused on the listener's opinion, since that is the value that is modeled and predicted. The final goal of many MER researches is a system that would allow automatic mood annotation. Listening tests are the only true possible way to verify the results. Most studies presented in Chapter 3 refer to listening tests on different stages of the presented research.

Listening tests conducted in the course of the presented dissertation were divided into two phases: preliminary and final. Experiments were designed for the following purposes (Fig. 6.1):

- Creating a dictionary.
- Determining the model of mood.
- Collecting data for further analysis.
- Selecting features essential for mood recognition and finally verification and discussion of the different classification algorithms.

In the literature related to the subject many different approaches exist [25,167,188,234,213]. Therefore preliminary tests enabled the determination of the reference point. The outcomes and conclusions of the critical review of the literature led to the assumptions and foundations of the analysis and solutions proposed by the author of this dissertation.

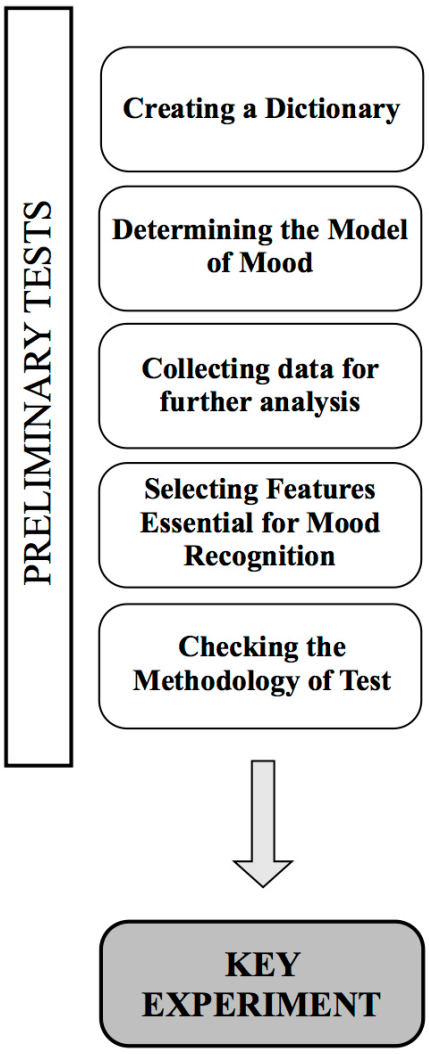



*Figure 6.1 Goals of subjective tests performed in the course of dissertation*

The aim of preliminary tests was: to create an appropriate dictionary related to mood of music, to determine and choose the model of mood, to collect data for further analysis and to select features essential for mood recognition. Preliminary experiments along with results and findings are presented in this chapter. Also some conclusions that are relevant for further work are included. Checking the form and methodology of the final test, which is described in chapter 7, was also one of the main roles of the preliminary tests.

This Section is organized as follows: the table with an overview of the experiment is placed at the beginning of each sub-section. Subsequently procedures and details are described in each sub-section.

## 6.1 DICTIONARY CREATION

There are many works that proposed various approaches and solutions to build mood-related dictionaries, but the vast majority of them are in English. It occurred that a simple translation of terms from English into Polish and using them in research is insufficient. Many such words are inadequate to describe music and the emotions that it creates. Although these words may easily be understandable to the audience, in Polish, they are not commonly used in the context of music. For example, in the translation of the Thayer's model (Valence/Arousal) there is a problem with the term "valence" (in Polish "walencja"). Social networking sites, and those recommending music as well, are available in various languages, although they are based on the same core resources. There is a question of whether the mood associated by a listener with a particular music track can be described in English or should it rather be expressed in a national language. Tab. 6.1 includes information with regard to an experiment that aimed to create a dictionary appropriate for music mood description in Polish.

*Table 6.1 Dictionary creation experiment*

| Title | Creating the dictionary |
|---|---|
| **Objectives** | The experiment presented below was conducted to create a dictionary of Polish words and/or terms that could be used to adequately describe music mood and to achieve subjective assessment ratings on the energy-arousal plane. |
| **Protocol** | - 36 listeners; 30 music fragments<br>- part A - description of mood of music using Polish adjectives<br>- part B - evaluation of mood on Energy/Arousal plane |
| **General Results and Conclusions** | - The set of words used for music mood description<br>- 30 musical excerpts mapped onto Energy/Arousal plane |

**Detailed description of the experiment**

The test consisted of two parts:

- Part A was aimed at verifying the appropriateness of mood descriptors in Polish (using a set of Polish adjectives)

- Part B was the assessment of mood at the energy-arousal plane.

Before the test, the subjects listened to a sequence of 3 examples, different from any of the test sets however, prepared in the same way. This was done to familiarize the listeners with the time pattern of the tracks' appearance and with the volume of audio examples at which the test was carried out.

Each part consisted of 30 examples. Parts A and B were separated by a 5-min. break. The duration of the whole test (including breaks) was approximately 35 minutes long. The stages of the test are presented in Fig. 6.2.

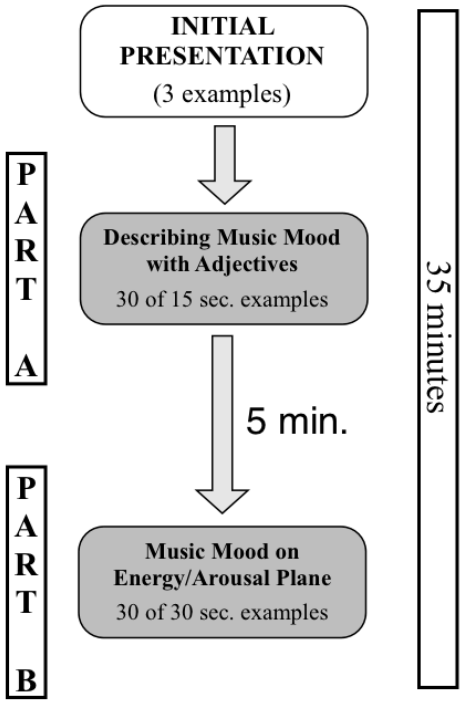



*Figure 6.2* *Subjective test arrangement related to music mood recognition and mood adjective searching (creating a mood dictionary in Polish)*

In both parts of the test, the same set of 30 songs was used. Tracks were chosen from five genres: Alternative Rock, Classical, Jazz, Opera & Vocal, and Rock. Songs used in the experiment came from the SYNAT music database [151,155]. The complete and detailed list of the music tracks is listed in Table 6.2. For the purpose of the test, 15-sec. and 30-sec. long excerpts were extracted. The 15-sec. fragments were also part of the corresponding 30-sec. excerpts.

The order of the tracks was random and different for both parts. However it was constant for each of the listeners. The subjects were informed about their tasks and given a printed instruction with questionnaires for their answers.

*Table 6.2 List of the music tracks used in the experiment*

| No | Gendre | Artist | Album | Title |
|---|---|---|---|---|
| 1 | Alternative Rock | Sigur Ros | Takk... | Hoppípolla |
| 2 | Classical | Maria Callas-Georges Prêtre-Orchestre | Maria Callas - La Divina | Tosca (1987 Digital Remaster)- Vissi d'arte |
| 3 | Classical | George Gershwin | 50 Greatest Hits Of Classical Music | Rhapsody in Blue |
| 4 | Alternative Rock | Super Furry Animals | Phantom Power | Hello Sunshine |
| 5 | Classical | Luciano Pavarotti | Nessun Dorma | Turandot - Act 3 - Nessun Dorma! |
| 6 | Opera & Vocal | Marlene Dietrich | Falling In Love Again | Lili Marlene |
| 7 | Jazz | Herbie Hancock | Future Shock | Mega Mix |
| 8 | Rock | Queen | Stone Cold Classics | Bohemian Rhapsody |
| 9 | Jazz | George Benson | The Best Of George Benson | On Broadway |
| 10 | Jazz | Diana Krall | When I Look In Your Eyes | Why Should I Care |
| 11 | Rock | U2 | The Joshua Tree | In God's Country |
| 12 | Opera & Vocal | Lucia Popp-Philharmonia Orchestra-Otto Klemperer | Mozart- Die Zauberflöte | Die Zauberflöte. |
| 13 | Rock | ZZ Top | Chrome. Smoke & BBQ-The ZZ Top Box | Concrete And Steel |
| 14 | Alternative Rock | Deftones | Saturday Night Wrist | Xerces (Album Version) |
| 15 | Opera & Vocal | Nina Sky | Move Ya Body | Move Ya Body |
| 16 | Jazz | The Dave Brubeck Quartet | Dave Brubeck Greatest Hits | Take Five |
| 17 | Alternative Rock | Kings Of Leon | Come Around Sundown | The End |
| 18 | Jazz | Bela Fleck & The Flecktones | Rocket Science | Life In Eleven |
| 19 | Classical | Ronan Hardiman | Michael Flatley's Lord Of The Dance | Gypsy |
| 20 | Opera & Vocal | Maroon 5 | Songs About Jane | She Will Be Loved |
| 21 | Rock | Billy Joel | The Essential Billy Joel | The Downeaster _Alexa_ |
| 22 | Opera & Vocal | Jacques Brel | Infiniment | Ne Me Quitte Pas |
| 23 | Alternative Rock | Mark Ronson & The Business Intl | Record Collection [Explicit] | Bang Bang Bang |
| 24 | Rock | Within Temptation | The Silent Force | Destroyed |
| 25 | Rock | Coldplay | Viva La Vida - Prospekt's March Edition | Lovers In Japan |
| 26 | Classical | Arthur Fiedler | Stars And Stripes | Stars and Stripes |
| 27 | Opera & Vocal | Linda Eder | Soundtrack | Falling Slowly |
| 28 | Classical | Pearl Jam | Big Fish - Music from the Motion Picture | Man Of The Hour |
| 29 | Jazz | Paco De Lucia | J Mclaughlin-P De Lucia-Al Di Meola | Manha De Carnaval |
| 30 | Alternative Rock | Imogen Heap | Lifeline | Lifeline |

**Part A - mood description in Polish**

The first part was the presentation of 30 fragments each 15 sec. long and separated by a break of 10 seconds. After every five examples a break of 15 seconds was made. As mentioned before, the listeners' task was to describe the mood of the music with any adjective (or adjectives) they considered most adequate. No predefined dictionary was offered to them. Both fast reaction and first impression were of key importance. This point was especially highlighted in the instruction.

**Part B - mood on the energy/arousal plane**

This part started with the presentation of 30 fragments each 30 sec. long and separated by a break of 10 seconds. Every six examples (6 graphs per page were presented) were separated by a 15 sec. period of silence. The listeners were asked to describe the mood of the music by marking a point on the energy (negative/positive) and arousal (low/high) plane.

Musical excerpts were presented in the MP3 stereo format, and the subjects were using closed Beyerdynamic DT 150 headphones. The reproduction system consisted of a PC computer and an audio interface ALESIS iO2.

The panel of listeners consisted of 36 subjects: 27 men and 9 women. The age of the subjects ranged from 20 to 26. Most of them were students of a sound and vision engineering course, and had at least elementary experience in sound engineering. Only a few were educated in music. As indicated in questionnaire forms filled in during the preparatory phase, listening to so-called background music is very common among students.

**Results**

In part A. the subjects were asked to describe the mood of the music by writing down at least one adjective. As the task was free-form, some editorial work had to be done. All adjectives were rewritten into masculine form and expressions other than adjectives were classified as a specific group of terms describing emotions or personal preferences. Some examples of the expressions assigned to this group are: sadness after breakup, autumn meditation, boring, mobilizing. etc.

Another separate group are words connected with music genre. A few of them are: chill out, swingy meditation, pop, opera etc.

The adjectives assigned by the listeners were examined and the occurrence of terms proposed by the audience was calculated. The overall quantities of the most popular expressions are presented in Tab. 6.3.

*Table 6.3 The overall quantity of the most frequent adjectives in part A*

| Adjective (English) | Adjective (Polish) | Number of occurrences |
|---|---|---|
| calm | *spokojny* | 122 |
| sad | *smutny* | 112 |
| happy | *wesoły* | 73 |
| joyful | *radosny* | 65 |
| brisk | *energiczny* | 58 |
| relaxing | *relaksujący* | 36 |
| melancholic | *melancholijny* | 32 |
| exalted | *podniosły* | 31 |
| positive | *pozytywny* | 31 |
| lively | *żywy* | 30 |
| serious | *poważny* | 27 |
| stimulating | *pobudzający* | 26 |
| energetic | *energetyczny* | 25 |
| pleasant | *przyjemny* | 25 |
| romantic | *romantyczny* | 24 |
| reflective | *refleksyjny* | 20 |

The results indicate four different trends:

- There is one favorite expression; other expressions have a similar meaning (Fig. 6.3);

- There is one favorite expression; other expressions have a different meaning (Fig. 6.4);

- There is no favorite expression; most of the expressions have a similar meaning (Fig. 6.5);

- There is no favorite expression; most of the expressions have a different meaning (Fig. 6.6).

Examples of the four tracks, which are representatives of the above four trends, are presented in Figs. 6.3-6.6.

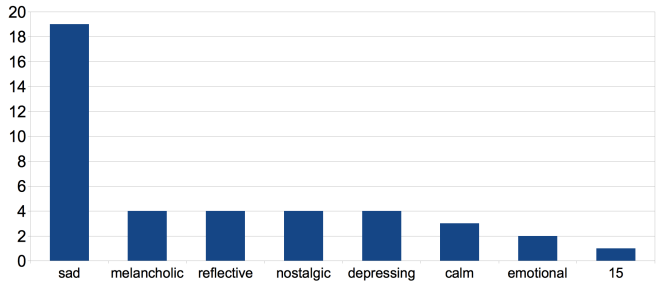



*Figure 6.3 Expressions given by listeners to describe mood of a music track. The last position in this graph represents the amount of other expressions, which occurred only once for a given song. Example No.* ***28****. Genre: Classical. Artist: Pearl Jam. Album: Big Fish - Music from the Motion Picture. Title: Man Of The Hour*

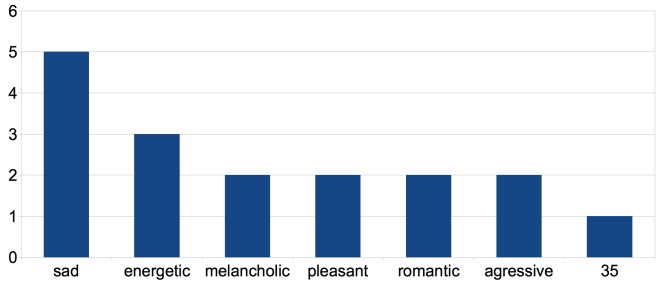



*Figure 6.4 Expressions given by listeners to describe mood of a music track. The last position in this graph represents the amount of expressions, which occurred only once for a given song. Example No.* ***24****. Genre: Rock. Artist: Within Temptation. Album: The Silent Force. Title: Destroyed*

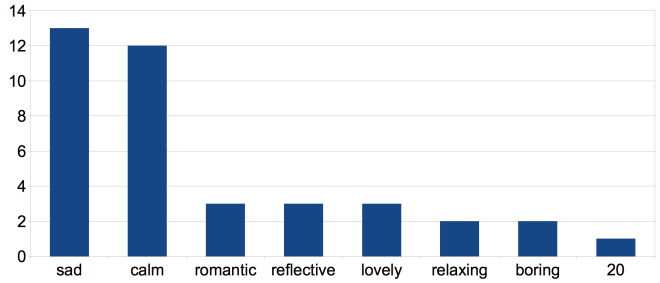



*Figure 6.5 Expressions given by listeners to describe mood of a music track. The last position in this graph represents the amount of expressions, which occurred only once for a given song. Example No.* ***27****. Genre: Opera & Vocal. Artist: Linda Eder. Album: Soundtrack. Title: Falling Slowly*

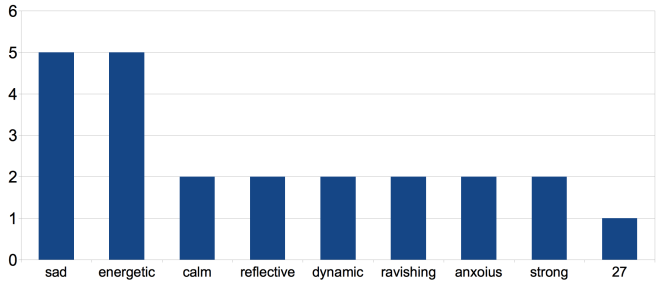



*Figure 6.6 Expressions given by listeners to describe mood of a music track. The last position in this graph represents the amount of expressions, which occurred only once for a given song. Example No.* ***17****. Genre: Alternative Rock. Artist: Kings Of Leon. Album: Come Around Sundown. Title: The End*

In part B the results from all of the listeners were averaged and for every song the average value along with the standard deviation for energy and arousal were calculated. The outcomes are presented in Fig. 6.7.

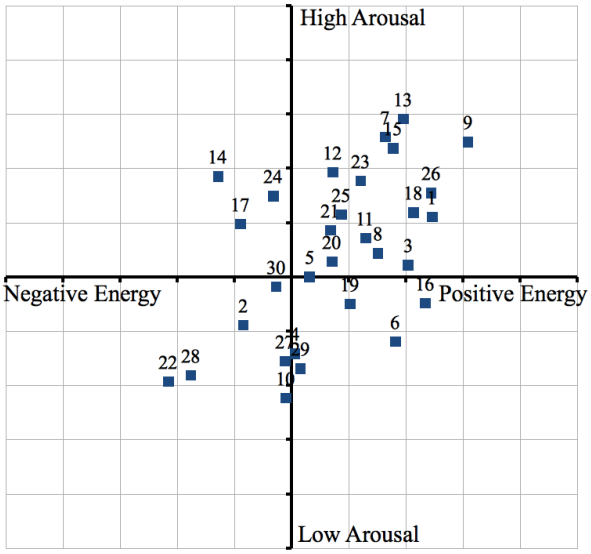



*Figure 6.7* *Results of Part B averaged for all subjects. Labels are marked in accordance with Table 6.4*

On this basis, all music samples fall into one of four general categories of the Thayer's model: exuberance (high arousal. negative energy), frantic (high arousal. positive energy), contentment (low arousal. positive energy) and depression (low arousal. negative energy). The results and classification are shown in Tab. 6.4. Some research has already been performed examining the relationship between mood and genre [168,160,182]. To follow the direction of these works, the presented results are divided into genre groups (Fig. 6.8).

Parallel analysis of the part A and B results is presented in Table 6.4.

Adjectives/mood descriptors from part A are grouped into four clusters according to the part B classification (similar to the Thayer's [308] model and Laurier's *et al.* clusters [167]). A particular adjective is mentioned in Table 6.5 if it occurred at least once in the particular mood cluster. Descriptors listed in Table 6.5 are in alphabetical order.

*Table 6.4 Results of the Part B averaged for all of the subjects. Mood is assigned in accordance to the Thayer's Energy/Arousal model*

| No. of Music Sample | Mood | Energy | | Arousal | |
|---|---|---|---|---|---|
| | | Average | St. Dev. | Average | St. Dev. |
| 1 | exuberance | 12.3 | 7.7 | 5.6 | 10.7 |
| 2 | depression | -4.3 | 9.6 | -4.4 | 12.0 |
| 3 | exuberance | 10.2 | 8.1 | 1.1 | 12.0 |
| 4 | contentment | 0.3 | 11.5 | -7.1 | 11.9 |
| 5 | exuberance | 1.5 | 11.0 | 0.0 | 12.6 |
| 6 | contentment | 9.1 | 9.5 | -5.9 | 10.5 |
| 7 | exuberance | 8.2 | 9.9 | 12.9 | 9.0 |
| 8 | exuberance | 7.5 | 10.5 | 2.2 | 14.0 |
| 9 | exuberance | 15.4 | 6.9 | 12.4 | 9.3 |
| 10 | frantic | -0.5 | 11.9 | -11.1 | 12.4 |
| 11 | exuberance | 6.4 | 9.7 | 3.6 | 11.4 |
| 12 | exuberance | 3.6 | 11.0 | 9.7 | 10.7 |
| 13 | exuberance | 9.8 | 9.8 | 14.6 | 9.1 |
| 14 | frantic | -6.5 | 11.6 | 9.3 | 11.9 |
| 15 | exuberance | 8.9 | 10.4 | 11.9 | 7.5 |
| 16 | contentment | 11.7 | 10.2 | -2.4 | 11.3 |
| 17 | frantic | -4.5 | 10.3 | 4.9 | 9.7 |
| 18 | exuberance | 10.6 | 11.1 | 5.9 | 12.1 |
| 19 | contentment | 5.1 | 11.2 | -2.5 | 12.7 |
| 20 | exuberance | 3.5 | 10.3 | 1.4 | 11.4 |
| 21 | exuberance | 3.4 | 9.2 | 4.3 | 11.7 |
| 22 | depression | -10.8 | 10.6 | -9.6 | 11.4 |
| 23 | exuberance | 6.0 | 14.2 | 8.9 | 13.6 |
| 24 | frantic | -1.6 | 12.1 | 7.5 | 11.4 |
| 25 | exuberance | 4.3 | 11.7 | 5.8 | 10.8 |
| 26 | exuberance | 12.2 | 11.4 | 7.8 | 9.1 |
| 27 | depression | -0.6 | 11.7 | -7.8 | 11.0 |
| 28 | depression | -8.9 | 9.4 | -9.1 | 9.8 |
| 29 | contentment | 0.8 | 13.2 | -8.4 | 10.2 |
| 30 | depression | -1.4 | 9.8 | -0.9 | 11.0 |

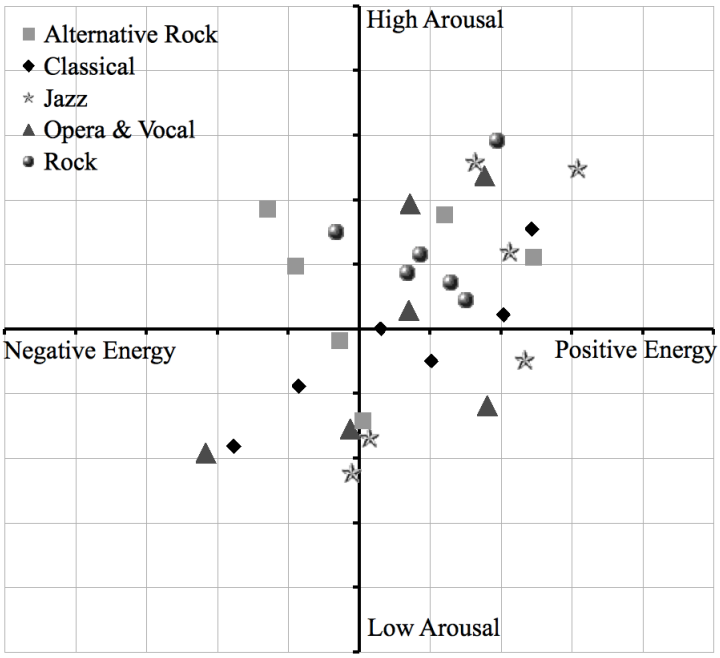



*Figure 6.8 Music samples presented on Energy/Arousal plane with the assigned genre*

In addition, the analysis with the aim of defining the so-called target/leader mood descriptors for a given music sample was performed. This means that the most frequent adjectives (in average it was 55 times per adjective) should be treated as target descriptors. It occurred that for most music samples there were a few target descriptors repeated within the same cluster. This strongly suggests that multi-label description may be a good way to perform music mood tagging. On the other hand, it may be sufficient to use two or three of the most significant descriptors to simplify the evaluation process.

**Discussion**

Considering the "target/leader" analysis for all of the songs in the experiment, it is reasonable to allow multi-label mood classification. Another rule should be to limit the number of expressions to maximum 3. It is important to observe that 6 out of 30 analyzed examples had three "leaders".

The classification derived from part B (Thayer's model mood description) is quite coherent with the lexical description from part A.

*Table 6.5 Adjectives obtained during part A. grouped by part B classification (Thayer's model)*

| contentment | depression | exuberance | frantic |
|---|---|---|---|
| calm<br>delicate<br>depressing<br>happy<br>joyful<br>lovely<br>melancholic<br>moody<br>mysterious<br>nostalgic<br>pleasant<br>positive<br>reflective<br>relaxing<br>romantic<br>sad | boring<br>calm<br>depressing<br>dramatic<br>emotional<br>exalted<br>lovely<br>melancholic<br>mysterious<br>neutral<br>nostalgic<br>opera<br>piercing<br>reflective<br>relaxing<br>romantic<br>sad<br>serious | brisk<br>calm<br>classic<br>dancing<br>depressive<br>dynamic<br>electronic<br>energetic<br>epic<br>exalted<br>happy<br>irritating<br>joyful<br>lively<br>loud<br>moody<br>neutral<br>noble<br>optimistic<br>party<br>pathetic<br>pleasant<br>positive<br>recreating<br>relaxing<br>rhythmic<br>sad<br>serious<br>stimulating<br>touching<br>vivacious | aggressive<br>anxious<br>calm<br>delicate<br>depressive<br>dynamic<br>energetic<br>exalted<br>lovely<br>melancholic<br>nostalgic<br>pleasant<br>ravishing<br>recreating<br>reflective<br>romantic<br>sad<br>strong |

Models, which consist of four types of mood (like Thayer's model) [308] or a folksonomy system (like Laurier's *et al.* [167]), are not sufficient. Despite the fact that some expressions belong to all four groups, most of them may be treated as properly assigned to the given cluster. Most of the adjectives belonging to a particular cluster are coherent with the mood it represents. It was also noted that Polish expressions are less diverse than English.

Thus, the general conclusion regarding this part of the experiments is that it is reasonable to use classification with more than four main categories of mood and permit multi-label classification or at least classification using more labels.

Some listeners reported problems with understanding the 2-dimensional Energy/Arousal representation. Their comments were related not only to the problematic deciphering of the nomenclature but they also said that the whole concept of 2D

representation was not intuitive. This is an important cue for carefully choosing an appropriate model of emotions for further research.

## 6.2 PRELIMINARY TESTS - CORRELATION ANALYSIS

All of the musical excerpts tested in the experiment described in Section 6.1 are elements of the SYNAT database. Each excerpt is described by a set of 173 parameters obtained through the analysis of mp3-quality recordings. The majority of the parameters are MPEG-7 standard items (109), along with 20 Mel-Frequency Cepstral Coefficients (MFCC), 20 MFCC variances and 24 time-related 'dedicated' parameters. Detailed information on the parameters and set of features contained in SYNAT is included in Section 4.5. The whole parameter set has been taken into consideration but the experiment was conducted to determine which of the parameters are relevant to mood recognition. Tab. 6.6 includes information with regard to correlation analysis in preliminary tests.

*Table 6.6 Correlation analysis applied to results of preliminary tests*

| Title | Correlation analysis applied to results of preliminary tests |
|---|---|
| **Objectives** | The data obtained from the subjective preliminary tests (described in Section 6.1) were checked in terms of correlation with a feature vector containing 173 parameters. The aim was to shorten the vector by choosing only the parameters correlated with mood labels. Different types of normalization were applied at this stage. |
| **Protocol** | Described in Section 6.1 |
| **General Results and Conclusions** | - A list of parameters ordered by the correlation coefficient value with regard to the music mood description was created<br>- Long (91) and short (8) vectors describing mood of music of parameters were determined<br>- While correlation between parameters and the Arousal label is moderate, it is not sufficient to derive any decisive conclusions for Energy<br>- Arguments for multi-label classification |

**Normalization**

Since not all of the parameters are from the same domain, the whole set of the parameters was normalized to range [0.1]. Every one of the 173 parameters was normalized separately. Two different types of normalization were applied, namely: Normalization I (linear unity-based normalization into range [0.1]) and Normalization II (two-step normalization where the average value becomes 0.5) (for details of normalization procedures see Section 5.1).

The important point is that the parameters were normalized according to the parameter values, which occurred in SYNAT. Therefore it is possible that the parameters of a song not included in this database may exceed range [0.1]. On the other hand, the SYNAT database contains more than 50.000 of music pieces from different time and genres, thus this kind of normalization seems to be legitimate.

Averaged assessment ratings for energy and arousal achieved from the listening test of part B were also normalized with both types of calculations.

**Correlation**

Parameters of 30 songs used in the subjective test were collected. For the Energy (positive/negative) and Arousal (high/low) descriptors, correlation analysis with 173 parameters extracted from the SYNAT database was performed. In the second step – the correlation with the normalized values of the assessment ratings and parameters was calculated. Student's t-distribution was used to assess the statistical significance of the correlation. The results of the correlation analysis for raw and normalized data are presented in Tabs. 6.7 and 6.8. Parameters have been ordered form higher to lower correlation values and separately for different types of pre-processing. Even though a great number of parameters should be listed for arousal, only some of them have been mentioned. It should be noted that for a particular way of pre-processing, a different number of statistically significant parameters has been achieved. However, the correlation is much higher for arousal and in terms of statistics is significant.

The reason behind fair and moderate correlation might be related to the students' not finding a sufficiently intuitive way of understanding the evaluation model, although arousal was moderately higher as it is easier to evaluate arousal than sad or happy instances. In the next steps described in 6.4 the MDS analysis was performed to determine a model of mood that is appropriate for emotions included in music.

**Data Distribution**

Songs from the test set were divided into two groups, namely: low and high arousal. This decomposition was derived based on the interclass inertia between classes of parameter vectors. Parameters after **Normalization I** were used in this part of data analysis. For the first analysis all of the statistically significant (according to t-student test) parameters were used. The second analysis included only parameters with the correlation coefficient larger than 0.6, which shortened the vector to 8 parameters only. The results of both calculations

are presented in Tab. 6.9. The interclass inertia is smaller for a shorter vector, which makes it more convenient for use in this type of data analysis.

*Table 6.7 Correlation between average rating for Arousal (low/high) and parameters. Parameters are ordered according to the correlation coefficient (from higher to lower values). The last presented values in table respond to the least significantly correlated parameters according to t-Student statistics*

| No. | No normalization | | Normalization I | | Normalization II | |
|---|---|---|---|---|---|---|
| | Parameter | Corr. | Parameter | Corr. | Parameter | Corr. |
| 1 | SFM13 | 0.68 | SFM13 | 0.67 | ASE16 | 0.71 |
| 2 | SFM18 | 0.65 | SFM18 | 0.65 | ASE19 | 0.71 |
| 3 | SFM14 | 0.65 | SFM15 | 0.63 | SFM15 | 0.71 |
| 4 | SFM10 | 0.64 | SFM14 | 0.63 | SFM13 | 0.71 |
| 5 | SFM_M | 0.64 | SFM10 | 0.63 | SFM18 | 0.70 |
| 6 | SFM6 | 0.64 | ASE19 | 0.62 | SFM19 | 0.68 |
| 7 | SFM7 | 0.64 | ASEV16 | 0.62 | ASE18 | 0.67 |
| 8 | SFM15 | 0.64 | SFMV3 | 0.62 | SFM16 | 0.67 |
| 9-90 | … | … | … | … | … | … |
| 91 | ASEV11 | 0.34 | ASEV3 | 0.33 | ASE29 | 0.37 |
| 92 | 2RMS_TCD_10FR_MEAN | 0.34 | | | MFCCV8 | 0.36 |
| 93 | MFCCV8 | 0.33 | | | ASEV21 | 0.36 |
| 94 | | | | | THR_1RMS_TOT | 0.35 |
| 95 | | | | | PEAK_RMS10FR_VAR | 0.35 |
| 96 | | | | | SFMV4 | 0.34 |
| 97 | | | | | MFCCV17 | 0.34 |

*Table 6.8 Correlation between average rating for Energy (negative/positive) and parameters. Parameters are ordered according to correlation coefficient (from higher to lower values). The last presented values in the table respond to the least significantly correlated parameters according to t-Student statistics*

| No. | No normalization | | Normalization I | | Normalization II | |
|---|---|---|---|---|---|---|
| | Parameter | Corr. | Parameter | Corr. | Parameter | Corr. |
| 1 | ASE29 | 0.43 | THR_1RMS_10FR_VAR | 0.45 | ASE29 | 0.42 |
| 2 | THR_1RMS_10FR_VAR | 0.42 | THR_2RMS_10FR_VAR | 0.42 | THR_3RMS_10FR_VAR | 0.40 |
| 3 | ASEV23 | 0.40 | ASEV23 | 0.41 | THR_2RMS_10FR_VAR | 0.38 |
| 4 | THR_2RMS_10FR_VAR | 0.39 | ASEV8 | 0.36 | | |
| 5 | 2RMS_TCD_10FR_VAR | 0.38 | THR_3RMS_10FR_VAR | 0.36 | | |
| 6 | MFCCV2 | 0.36 | SFMV12 | 0.36 | | |
| 7 | 3RMS_TCD_10FR_VAR | 0.35 | | | | |
| 8 | MFCC17 | 0.33 | | | | |

*Table 6.9 Interclass inertia for longer and shorter vectors of parameters*

| **Interclass inertia** | |
|---|---|
| **Long vector (91 parameters)** | **Short vector (8 parameters)** |
| 0.73 | 0.31 |

**Discussion**

Correlation between parameters and Arousal is moderate and this is a promising result to continue this study aimed for proving the Thesis no. 1, i.e.: "**It is possible to find parameters describing a musical excerpt, which are highly correlated with subjective mood labeling results.**".

The correlation coefficient for energy is not higher than 0.45 and this result leads to a conclusion that parameters contained in the SYNAT database are not sufficient to retrieve

information about positive or negative energy from. That's why some additional parameters related to temporal domain were introduced (see Section 4.6.2).

According to the classification analysis, employing shorter vectors. i.e. including only parameters with higher correlation values is more reasonable than using longer vectors (created by t-student statistical analysis).

## 6.3 TEMPO AND RHYTHM

Tempo variation has consistently been associated with differential emotional responses to music [64,108,127]. The aim of this experiment was to find relation between tempo features and mood perception. Fixed music arrangement was replaced by various rhythms and listeners were asked to describe mood of the music for different tempo and rhythm combinations. Tab. 6.10 contains general information about experiment designed to determine the influence of tempo and rhythm on mood of music.

*Table 6.10 Experiment related to influence of tempo and rhythm on mood of music*

| Title | Tempo and Rhythm |
|---|---|
| **Objectives** | The experiment was conducted to determine relationship between tempo and perceived mood of music. |
| **Protocol** | - 40 subjects<br>- One piece of music with 5 different drum rhythms was tested within range 90 and 130 BPM. |
| **General Results and Conclusions** | - Mood of music is highly correlated with tempo<br>- Values of correlation coefficients are greater than 0.8<br>- The change of rhythm for fixed tempo causes less difference in mood perception. |

**Detailed description of the experiment**

The song, which was used in experiments, consisted of the following audio tracks: female vocal (melody, no words used), electric guitar, synth bass, synths and drum set. The author of this dissertation performed whole production and recording process as a recording engineer and producer. Recording techniques were based on the previous research and experience of the author [138,245,246]

The drum set was recorded in the live room using a multi-track technique listed in Tab. 6.11. Configuration is presented in Fig. 6.9. Nine microphones were used: two for bass drum (Sennheiser e901 and Audix D6), snare drum (Shure SM57), hi-hat (Nueman TLM

103), two overheads (Schoeps MK4), and three toms (2 Audix D2 and Audix D4). Other tracks were taken from sounds library from Logic 9 Pro software [186].

*Table 6.11 Drum set recording session input list. Particular parts of the set are listed along with used microphones.*

| Instrument | Microphone |
|---|---|
| Kick Drum | Audix D6 |
| Kick Drum | Sennheiser e901 |
| Snare Drum | Shure Beta57 |
| Hi-Hat | Neuman TLM 103 |
| Rack Tom 1 | Audix D2 |
| Rack Tom 2 | Audix D2 |
| Floor Tom | Audix D4 |
| Overheads | Schoeps MK4 |

*Figure 6.9 Drum set recording setup*

The whole music arrangement remained with no changes, only the rhythmic part (drum set tracks) varied. The drummer played five different rhythms named from A to E. Characteristic features of particular rhythms are listed below:

- **Rhythm A** – tribal rock
- **Rhythm B** – shuffle rock ballad
- **Rhythm C** – pop-rock beat
- **Rhythm D** – speed blues
- **Rhythm E** – disco-funk.

Rhythms B, C and E were presented in the tempo range from 90 to 130 BPM (step of 10 BPM). The other rhythms were presented in 90 BPM, what makes 17 music samples in total. Duration of the particular samples varied from 24 to 34 seconds (depending on the tempo). The following musical excerpts were randomly chosen and ordered and the sequence was different for every listener. Samples were presented in 128 kbit/s MP3 format.

The subjects were asked to evaluate to what extent a given expression describes mood of the presented music. The mood labels used in the survey are presented in Tab. 6.12.

The survey was implemented as a HTML website. The test began with the introduction page, where subjects were instructed about their task and could playback a fragment of music to adjust the volume. Then listeners were guided through the 17 subpages in Polish (Fig. 6.10) with playback and the evaluation form. The whole test duration was approximately 15 minutes.

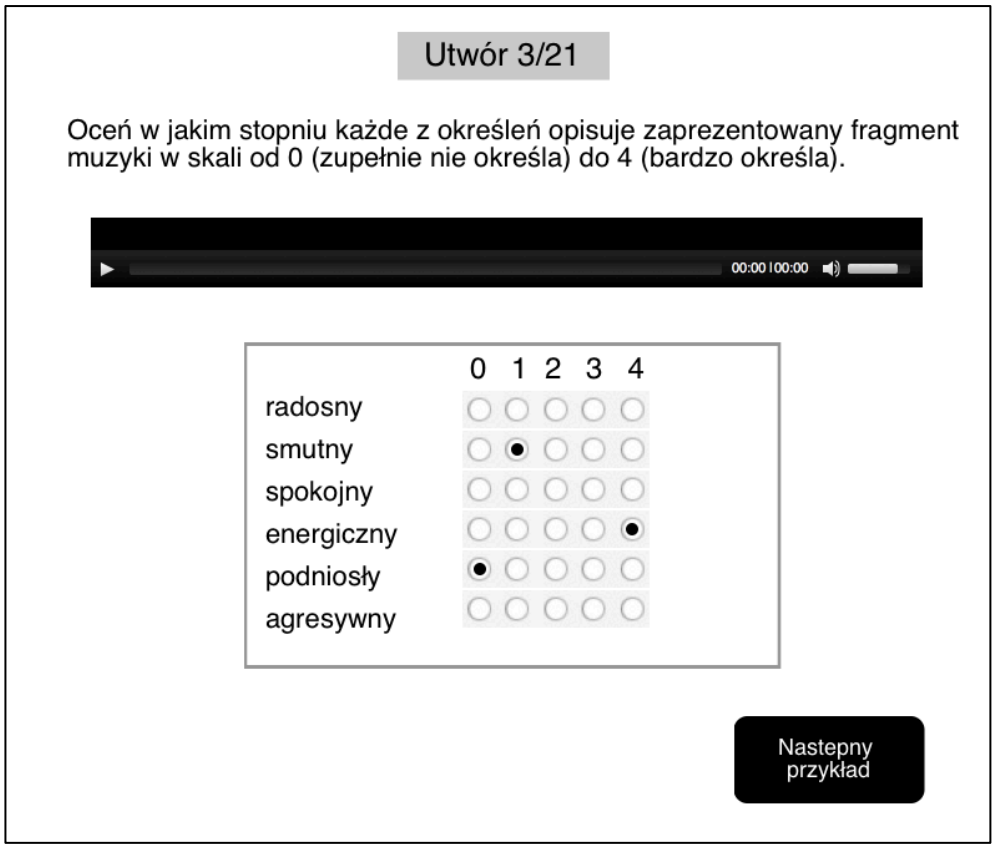



*Figure 6.10 Web interface used in the Part I of the experiment (in Polish)*

*Table 6.12 Expressions used in the survey to describe mood of music*

| Label No. | Mood label (English) | Mood label (Polish) |
|---|---|---|
| 1 | Aggressive | *Agresywny* |
| 2 | Brisk | *Energiczny* |
| 3 | Exalted | *Podniosły* |
| 4 | Joyful | *Radosny* |
| 5 | Sad | *Smutny* |
| 6 | Calm | *Spokojny* |

The panel of listeners consisted of 40 subjects: 19 men and 21 women. The average age of the subjects was 25 years. They didn't report any problems with hearing.

**Results**

Results for all of the listeners were averaged and standard deviation was calculated. Averaged grade values for particular rhythms and tempos are presented in Figures 6.11-6.17.

Results show that the influence of tempo is much more significant than different rhythms and are similar for all rhythms. As an example of the tempo impact are shown for rhythm C and E respectively in Figs. 6.11 and 6.12.

Graphs in Figs. 6.13 and 6.14 are created to show the collective comparison of all of the results for tempo 90 BPM. In Fig. 6.13 results are grouped by the labels to show which rhythm seems to be the most related to the particular expression. In Fig. 6.14 notes are grouped by rhythm to show how listeners described every rhythm. In Figs. 6.15-6.17 the dependence between intensity of different labels and the tempo for rhythms B, C and E is presented.

The correlation coefficient between particular label notes and tempo for rhythms was calculated. Results are gathered in Tab. 6.13. Moreover, correlation between particular mood labels was analyzed and these calculations are presented in Tab. 6.14. Signs "+" or "-" show positive or negative correlation, while "NO" indicates cases, when the absolute value of the correlation coefficient is smaller than 0.8.

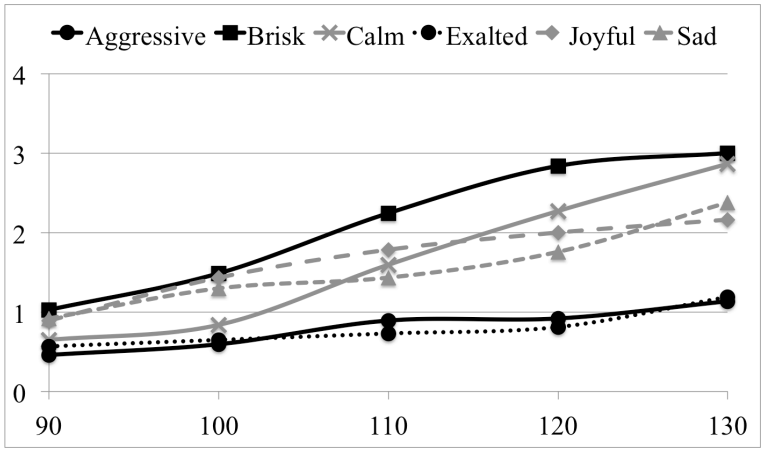



*Figure 6.11* *Relation between tempo and perceived mood of music. Averaged results for rhythm C*

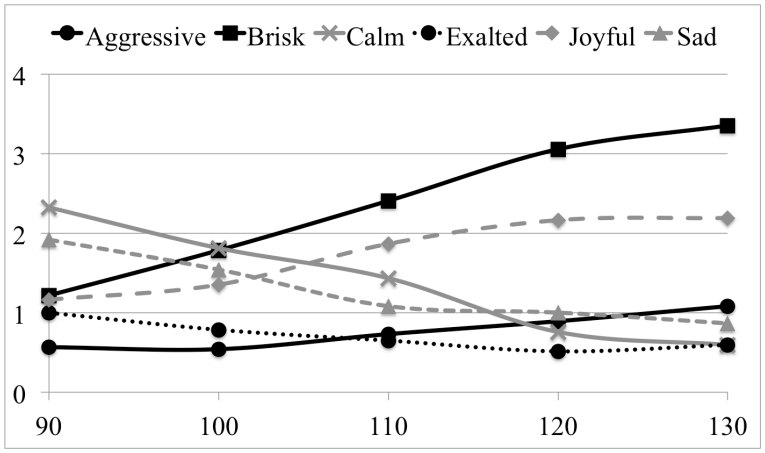



*Figure 6.12* *Relation between tempo and perceived mood of music. Averaged results for rhythm E*

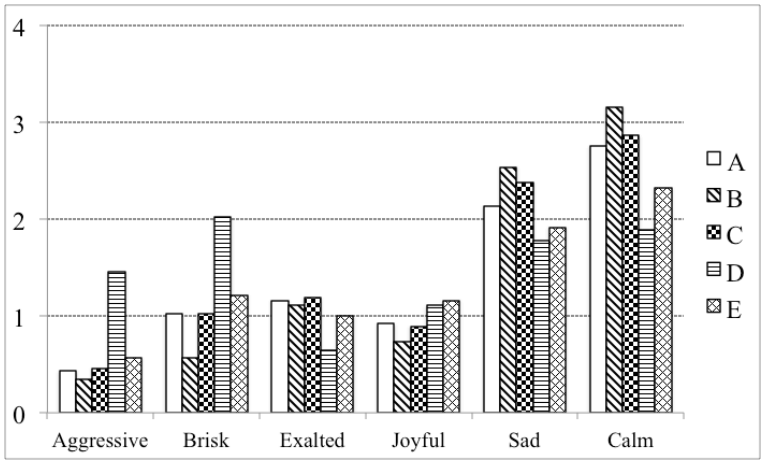



*Figure 6.13* *Evaluation of mood of music compared for different rhythms (A-E) for a fixed tempo (90 BPM)*

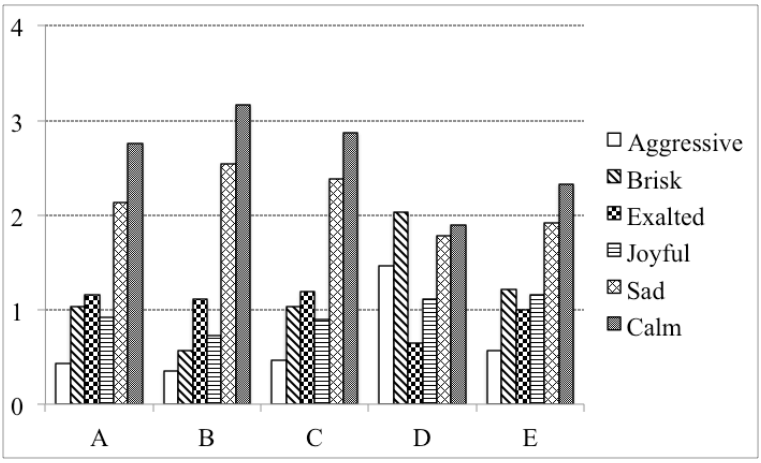



*Figure 6.14* *Mood of music description with averaged labels for different rhythms (A-E) for a fixed tempo (90 BPM)*

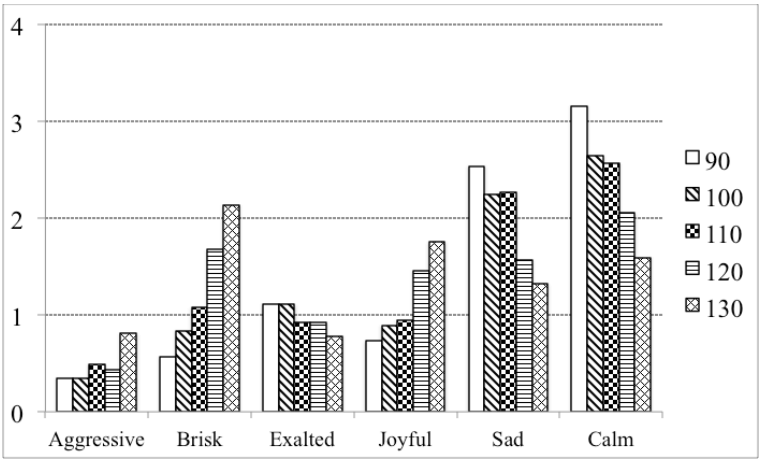



*Figure 6.15* *Averaged results for music with rhythm B at different tempos*

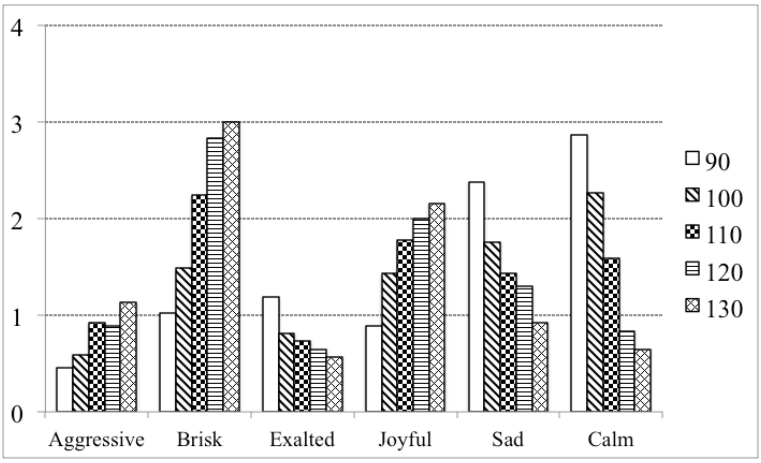



*Figure 6.16* *Averaged results for music with rhythm C at different tempos*

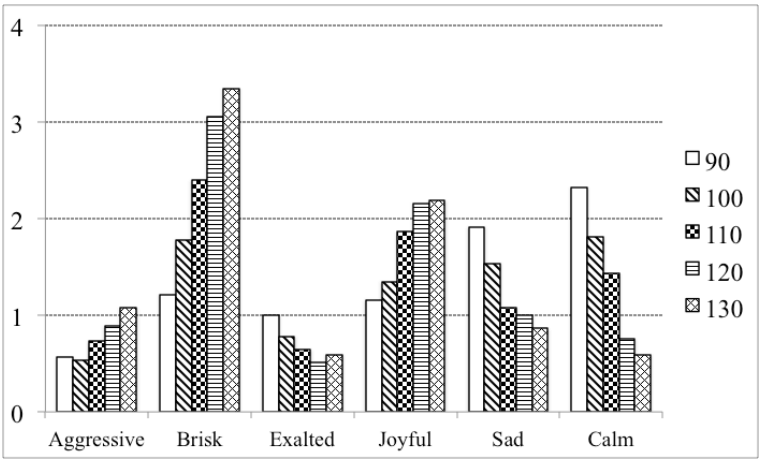



*Figure 6.17 Averaged results for music with rhythm E at different tempos*

*Table 6.13 Correlation coefficient between tempo and particular mood labels*

| Mood label | Correlation coefficient | | |
|---|---|---|---|
| | Rhythm B | Rhythm C | Rhythm F |
| Aggressive | 0.8315 | 0.9631 | 0.959 |
| Brisk | 0.9846 | 0.9835 | 0.994 |
| Exalted | -0.9486 | -0.921 | -0.8971 |
| Joyful | 0.9582 | 0.971 | 0.9656 |
| Sad | -0.9514 | -0.9734 | -0.9567 |
| Calm | -0.9832 | -0.9877 | -0.9884 |

*Table 6.14 Correlation between mood labels. Descriptions are numbered as in Tab. 6.12. The correlation was assumed as significant when modulo of the correlation coefficient was greater than 0.8. "NO" is related to not significant correlation. "+" means positive correlation and "—" negative correlation*

| No. of mood label | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1** | 1 | NO | — | NO | NO | NO |
| **2** | | 1 | — | + | — | — |
| **3** | | | 1 | — | + | + |
| **4** | | | | 1 | — | — |
| **5** | | | | | 1 | — |
| **6** | | | | | | 1 |

**Conclusions**

Results of the listening tests show that mood of music is highly correlated with tempo. This conclusion has been confirmed for three different rhythms and correlation coefficient is significant for every mood label.

The change of rhythm for fixed tempo causes less difference in mood perception (Figs. 6.13 and 6.14). This is coherent with results of Schellenberg *et al*. [278] who showed that changes in rhythm (varying vs. held constant) had little effect on listeners' emotional responses, which may be surprising to some extent.

It should be mentioned that for rhythms D and E the achieved results are different from other rhythms for tempo 90 BPM. It can be caused by specific characters of both sequences - rhythm D (speed blues) is much more dense, the drummer used more cymbal sound. On the other hand the rhythm E is a typical disco beat which can be perceived as more energetic because of its structure.

Values of the correlation coefficients are greater than 0.8, which shows a significant dependence between tempo and mood of music. This leads to the conclusion that parameters describing tempo and rhythmic content of music should be taken into consideration while creating a vector of parameters related to mood of music.

It should be noticed that the experiment was conducted for fixed arrangement with various rhythms presented in different tempo. It was intended to exclude the impact of another factors, i.e. key, harmonic progression, and overall instrumentation. The tempo was the only variable and all of the conclusions are related to tempo changes. Also no absolute values can be achieved, only the correlation between tempo and the perceived mood of music.

## 6.4 MULTIDIMENSIONAL SCALING ANALYSIS APPLIED TO MUSIC MOOD RECOGNITION

The motivation of this stage of the research comes from the earlier experiment (described in Section 6.2), where correlation results between parameters and Energy and Arousal dimensions was only fair or moderate. Moreover, as reported before, some listeners pointed out that the Energy/Arousal description was not intuitive. Therefore MDS approach was implemented to examine the model that describes mood of music in effective way. MDS allows determining significant number of dimensions to describe perceived

relations between objects. Multidimensional Scaling experiment was conducted to determine and confirm model of mood as well as check coherence between model with orthogonal dimensions and model with redundant descriptors. Tab. 6.15 includes information with regard to Multidimensional Scaling experiment related to mood of music.

*Table 6.15 Multidimensional Scaling experiment*

| Title | Multidimensional Scaling Analysis Applied to Music Mood Recognition |
|---|---|
| **Objectives** | Experiment was conducted to determine number of dimensions that allow describing mood of music. Results were confronted with evaluation with 6 descriptors. |
| **Protocol** | - Part I - 15 samples evaluated using set of 6 descriptors, each scaled from 0 to 4.<br>- Part II - collect data for MDS Analysis. 10 musical excerpts<br>- 36 subjects |
| **General Results and Conclusions** | - From MDS - sufficient number of dimensions to describe mood of music is 2; they correspond to labels „Calm" and „Joyful"<br>- Results collected in both parts of the experiment are coherent<br>- Different metrics can be used in terms of mood representation<br>- A list of parameters ordered by the correlation coefficient value with regard to the music mood description was created |

**Experiment Part I**

The initial part of the listening tests consisted of 15 samples from different music genres. For the purpose of the survey the mood labels were presented in Polish and they can be found. along with their English counterparts, in Tab. 6.16. Duration of every music excerpt was constant and remained 30 seconds. The complete and detailed list of the music tracks is listed in Tab. 6.17. Excerpts also used in the Experiment II are colored in grey. The subjects were asked to evaluate the extent to which a given label describes mood of the particular music excerpt. The labels were chosen during previous research study, which was conducted to create a dictionary associated with mood of music in Polish, reported in Section 6.1 [242].

The musical excerpts were randomly ordered and the sequence was different for each listener. Samples were presented in 128 kbit/s MP3 format. For the purpose of the experiments a survey was implemented as a HTML website in a series of simple HTML pages (Fig. 6.18, *in Polish*). The test began with the introduction page, where subjects were instructed about their task and could playback the music excerpt to adjust the volume. Then listeners were guided through the 15 subpages with playback and the evaluation form. The

entire test took approximately 15 minutes (including breaks between music samples and time needed for the answer). The panel of listeners consisted of 36 subjects: 24 men and 12 women. The average age of the subjects was 23 years. Again no hearing problems were reported.

*Table 6.16 Expressions used in the survey to describe mood of music*

| Label No. | Mood label (English) | Mood label (Polish) |
|---|---|---|
| 1 | Aggressive | *Agresywny* |
| 2 | Brisk | *Energiczny* |
| 3 | Exalted | *Podniosły* |
| 4 | Joyful | *Radosny* |
| 5 | Sad | *Smutny* |
| 6 | Calm | *Spokojny* |

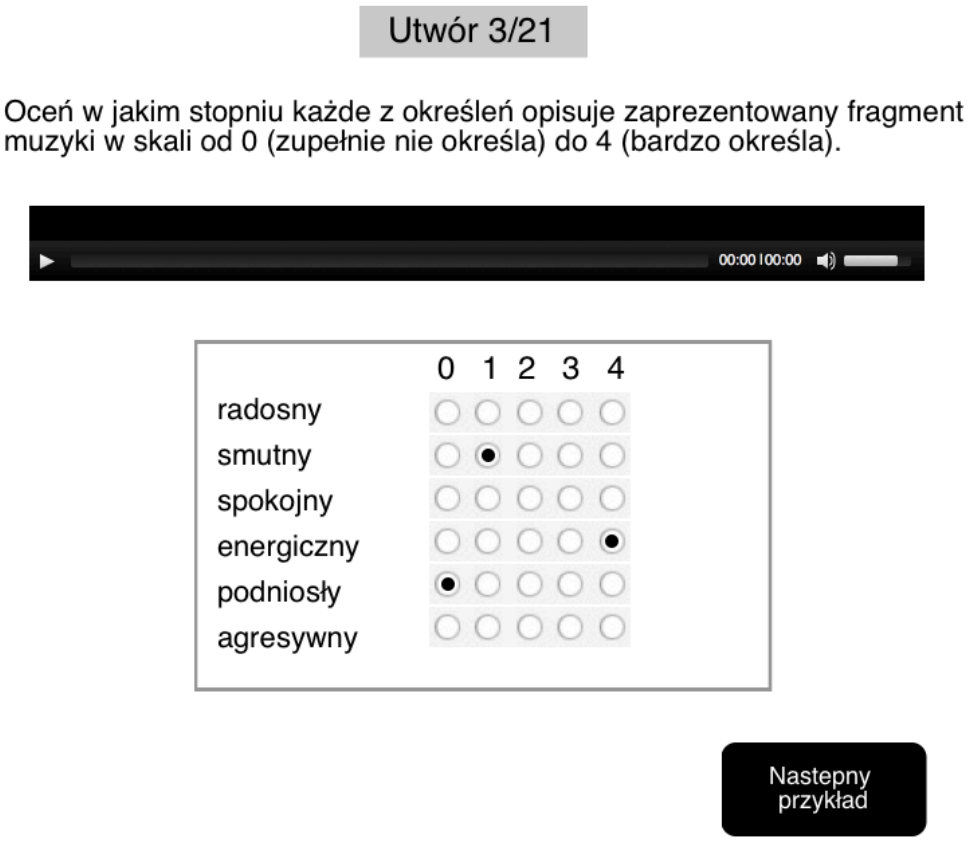



*Figure 6.18 Web interface used in the Part I of the experiment (in Polish)*

**Experiment Part II**

The second experiment was conducted to collect the similarity data for the MDS analysis. To make the duration of the entire test reasonable, only 10 of 15 music pieces were chosen (tracks used in Experiment II are marked in Tab. 6.17 in grey). These shorter excerpts were 15 seconds long and were consistent with music pieces presented in Experiment I (30-second long excerpts). Even though, this test lasted over 25 minutes because peer-to-peer similarity judgment required 45 presentations of 2 x 15 sec. pairs. The set of songs used in Experiment II was chosen according to the results of Experiment I. From every label tracks with the highest and the lowest notes were taken as well as excerpts with very similar notes were included.

The interface built for Experiment I was slightly adapted to meet the requirements of Experiment II. Each of 45 subpages contained playback and evaluation form, which includes a slider with minimum and maximum values set between 'totally different' and 'identical'. The subjects were asked to evaluate similarity between **moods** of two music pieces, one presented just after the other. Pairs were presented in random order and the sequence was different for each listener. The values of the slider were read and then written to the output CSV file format. The range of the slider was set from 1 ('totally different') to 100 ('identical') but the assigned values were not displayed (subjects' judgments were based on the position of the slider).

**Results**

In Experiment I the subjects were asked to evaluate to what extent a given expression describes mood of the presented music. Averaged results for all of the listeners are presented in Tab. 6.18.

Similarity data obtained from Experiment II were averaged. Normalized matrix of similarity is presented in Tab. 6.19. The MDS representation of data was constructed in MATLAB using Kruskal's normalized *Stress-1* criterion. Two dimensions were sufficient to create adequate representation. *Stress-1* factor reached 0.01. The MDS map is presented in Fig. 6.19.

*Table 6.17 List of the music tracks used in the experiment. All of the 15 songs were played back in Experiment I. songs marked in grey were also used in Experiment II*

| 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 | Reloaded3a Greatest Hits | Kiss |
| 5 | Alternative Rock | Pearl Jam | Ten (Legacy Edition) | Black (Remastered 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 Edition3a The Complete Electrical Recordings of Sir Edward Elgar2e | Coronation March Op2e 65 (1993 Digital Remaster) |
| 9 | Alternative Rock | Hey Champ | Star | Cold Dust Girl |
| 10 | Pop | Jennifer Lopez | Love3f (Deluxe Version) | Charge Me Up |
| 11 | Pop | Erykah Badu | Live | Tyrone (Extended Version) |
| 12 | Rock | Faith No More | This Is It3a 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 |

*Table 6.18 Averaged results of Experiment I. Columns correspond to mood labels according to Tab. 6.12 (1– Aggressive, 2 – Brisk, 3 – Exalted, 4 – Joyful, 5 – Sad, 6 – Calm) and rows represent songs (Tab. 6.17). Minimum scores for particular labels are marked in light grey, while the maximum in dark grey*

| No. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1** | 0.38 | 1.22 | 1.49 | 1.77 | 1.86 | 3.43 |
| **2** | 0.65 | 3.00 | 0.69 | 3.14 | 0.58 | 1.35 |
| **3** | 1.26 | 3.29 | 0.71 | 3.12 | 0.54 | 0.80 |
| **4** | 0.94 | 2.82 | 0.88 | 2.74 | 0.60 | 1.08 |
| **5** | 0.49 | 1.28 | 1.52 | 1.15 | 2.54 | 2.97 |
| **6** | 0.25 | 1.02 | 2.11 | 1.31 | 2.51 | 3.65 |
| **7** | 1.82 | 3.23 | 0.89 | 2.92 | 0.42 | 0.66 |
| **8** | 0.92 | 1.71 | 3.51 | 0.65 | 2.78 | 1.83 |
| **9** | 1.37 | 3.00 | 0.63 | 2.32 | 0.94 | 0.71 |
| **10** | 1.62 | 3.45 | 0.52 | 2.45 | 0.52 | 0.46 |
| **11** | 0.28 | 0.78 | 0.94 | 1.09 | 2.15 | 3.75 |
| **12** | 2.49 | 3.48 | 0.94 | 2.52 | 0.71 | 0.49 |
| **13** | 0.75 | 1.78 | 1.37 | 1.58 | 2.20 | 2.46 |
| **14** | 0.18 | 0.65 | 1.09 | 1.66 | 1.71 | 3.75 |
| **15** | 3.74 | 3.69 | 0.78 | 1.06 | 0.86 | 0.20 |

As shown in Fig. 6.19 songs with very different notes in Experiment I are placed in the opposite parts of the map. For example Track 15 evaluated as very aggressive and brisk is far from no. 1 perceived as almost not aggressive and brisk (see Tab. 6.18). Also very similar songs (nos. 1, 6, 11 and 14) are grouped in one area on the MDS map. Moreover, other distances between objects (songs) were achieved from data from Experiment I. All labels were regarded as separate dimensions and so every song was represented by 6-element vector (6D). Correlation analysis between mood labels was performed and results are presented in Tab. 6.20. Distances between objects (songs) were calculated using two metrics (Euclidean and Chebyshev). The then the MDS analysis was applied again. The average distance between the direct similarity MDS representation (MDS) and representation (6D) was slightly smaller for the Euclidean distance. The detailed results are presented in Tab. 6.21. Averaged distance between representations was calculated according to Eq. (6.1):

$$d_{av} = \frac{\sqrt{\sum_{i=1}^{m} (x_i - x_i')^2}}{m} \qquad (6.1)$$

where $x_i$ and $x_i'$ are coordinates of object $i$ and $m$ is the number of objects. As is shown in Tab. 6.22, **Dimension 1** corresponds to expression "Calm" (negative correlation) and "Brisk" (positive correlation) and **Dimension 2** to "Joyful" and to a lesser extent to "Exalted".

*Table 6.19 Similarity matrix obtained from listening tests for music tracks. Values are normalized to range [0.1]. Tracks are numbered according to Tab. 6.15*

| No. | 1 | 2 | 6 | 7 | 8 | 9 | 10 | 11 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 1 | 0.25 | 0.81 | 0.09 | 0.15 | 0.16 | 0.06 | 0.75 | 0.71 | 0.02 |
| **2** | 0.25 | 1 | 0.26 | 0.52 | 0.04 | 0.53 | 0.46 | 0.21 | 0.21 | 0.05 |
| **6** | 0.81 | 0.26 | 1 | 0.20 | 0.20 | 0.18 | 0.10 | 0.65 | 0.83 | 0.02 |
| **7** | 0.09 | 0.52 | 0.20 | 1 | 0.06 | 0.65 | 0.54 | 0.08 | 0.09 | 0.21 |
| **8** | 0.15 | 0.04 | 0.20 | 0.06 | 1 | 0.04 | 0.03 | 0.16 | 0.14 | 0.09 |
| **9** | 0.16 | 0.53 | 0.18 | 0.65 | 0.04 | 1 | 0.58 | 0.10 | 0.13 | 0.18 |
| **10** | 0.06 | 0.46 | 0.10 | 0.54 | 0.03 | 0.58 | 1 | 0.06 | 0.06 | 0.13 |
| **11** | 0.75 | 0.21 | 0.65 | 0.08 | 0.16 | 0.10 | 0.06 | 1 | 0.77 | 0.03 |
| **14** | 0.71 | 0.21 | 0.83 | 0.09 | 0.14 | 0.13 | 0.06 | 0.77 | 1 | 0.02 |
| **15** | 0.02 | 0.05 | 0.02 | 0.21 | 0.09 | 0.18 | 0.13 | 0.03 | 0.02 | 1 |

*Table 6.20 Correlation between mood labels. Descriptions are numbered as in Tab. 6.5. The correlation was assumed as significant when modulo of the correlation coefficient was greater than 0.8. "NO" is related to not significant correlation. "+" means positive correlation and "—" negative correlation*

| No. of mood label | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1** | 1 | + | NO | NO | NO | — |
| **2** | | 1 | NO | NO | — | — |
| **3** | | | 1 | + | + | NO |
| **4** | | | | 1 | — | NO |
| **5** | | | | | 1 | + |
| **6** | | | | | | 1 |

*Table 6.21 Distance between MDS (6D) representations and MDS (MDS). Average distance $d_{av}$ is calculated according to Equation 6.1*

| MDS representation | $d_{av}$ | Stress-1 |
|---|---|---|
| 6De. Euclidean distance | 0.89 | 0.02 |
| 6Dc. Chebychev distance | 0.92 | 0.00 |

*Table 6.22 Correlation between MDS dimensions and averaged notes from Experiment I. Columns correspond to mood labels according to Tab. 6.12. Maximum values of correlation coefficient for every dimension are marked in dark grey*

| No. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **D1** | 0.87 | 0.98 | -0.45 | 0.41 | -0.79 | -0.98 |
| **D2** | 0.36 | -0.08 | 0.68 | -0.78 | 0.50 | -0.04 |

**Conclusions**

According to multidimensional scaling procedure, two dimensions are sufficient to create adequate representation. MDS returns clear results coherent with the evaluation of Experiment I reported in this Section, where songs were evaluated by listeners using six mood labels.

Bigand et al. [32] 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 the presented study, therefore results may be different. Firstly, music set in Bigand's and his collaborators study consisted of only classical music (solo, chamber music, orchestra) with no involvement of other genres. Secondly, their testing procedure was based on choosing the excerpt most similar to the presented one rather than evaluating similarity or disparity. Also their study was focused on emotions of the listener while in presented research, mood of music is analyzed.

Dimensions achieved with the use of MDS correspond to labels „Calm" and „Joyful". This can lead to the conclusion that Thayer's model is accurate to describe mood of music. One of the axes can be interpreted as Valence ("Joyful" - positive or negative content) and the second as Arousal ("Calm" - energetic content). While both MDS representations (calculated from direct similarity judgments and from 6 labels similarity) are coherent, chosen mood

labels seem to be reasonable and accurate. These findings are important indications for next step of the research, where proprietary model of emotions is proposed.

Both 6D MDS representations (calculated for Euclidean and Chebychev distances in 6-dimensional labels space) return results close to MDS (direct similarity MDS map). This can lead to the conclusion that different metrics can be used in terms of mood representation. The Euclidean metric could be placed in privileged position while using linear scale during the test (e.g. data read from linear sliders).

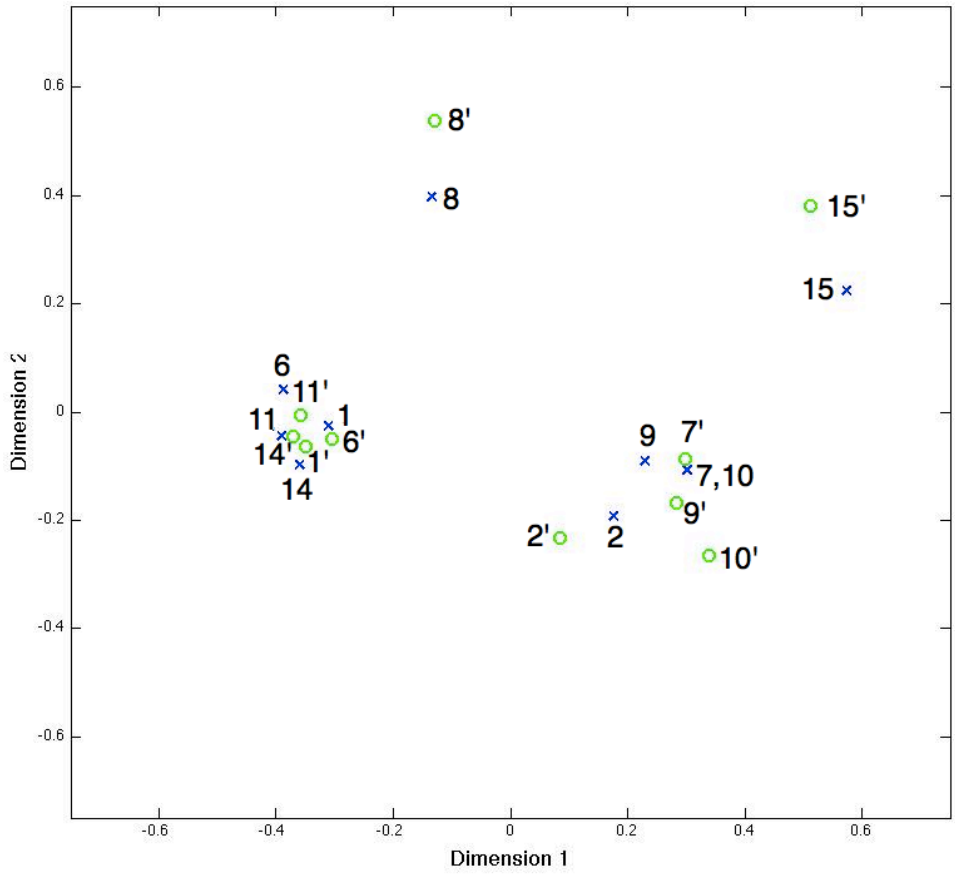



*Figure 6.19 Comparison of MDS representations based on direct similarity judgments (marked with (o) and apostrophe) and distance calculated from evaluation with 6 labels (x)*

## 6.5 MUSIC MOOD VISUALIZATION USING SOMs

As presented in Chapter 5 SOM (Self-Organizing Map) is an unsupervised neural network providing a mapping from a high-dimensional space to few-dimensional (in most of cases two-dimensional) representation. Music Mood Recognition is a task that meets

main assumptions of the method. Various mathematical and statistics methods (i.e. normalization. correlation analysis. PCA) were used to extract parameters most relevant to mood of music. These outcomes were fed into SOM algorithm to map object onto few-dimensional representation. Tab. 6.23 includes information with regard to the SOM-based procedure applied to Music Mood Recognition.

**Vector of parameters**

As before, all of the music fragments included in experiment described in Section 6.4 are elements of the SYNAT database. Parameters detailed description is included in Section 4.5. The whole parameter set has been taken into consideration but experiment was conducted to determine which of the parameters are relevant to mood recognition. To determine parameters, which are the most significant to mood descriptions, correlation analysis was applied.

*Table 6.23 Self-organizing maps experiment*

| **Title** | **Music Mood Visualization Using SOMs** |
|---|---|
| **Objectives** | Analysis performed with the results collected from experiment described in Section 6.4. Principal Component Analysis applied to vectors describing Dimension 1 (29 parameters) and 2 (12 parameters). Results processed with Self Organizing Maps. |
| **Protocol** | Described in Section 6.4. |
| **General Results and Conclusions** | The preliminary SOM analysis; methods and tools tested<br>- Ideas of graphical representation (map) |

Correlation between MDS dimensions and mood parameters was calculated and therefore a set of features strongly related to mood was created. 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 the correlation was determined according to t-student $t_{.975}$. Finally the feature vector describing mood of music consisted of 79 parameters, listed in Tab. 6.24 was obtained. For each dimension moderate and strongly correlated parameters were found, what partially proves the Thesis no. 1, which assumes the **correlation** between feature vector parameters describing mood and subjective evaluation results.

*Table 6.24 Set of parameters used for mood description. Denotations are as follows: ZCD (Zero-Crossing Rate). RMS (dedicated energy and time-related descriptor). ASE (*Audio Spectrum Envelope*). SFM (Spectral Flatness Measure). MFCC- Mel-Frequency Cepstral Coefficients (their mean and variance values)*

| No. | Dimension 1 | | No. | Dimension 2 | |
|---|---|---|---|---|---|
| | Parameter | Corr. | | Parameter | Corr. |
| 1 | ASE15 | 0.92 | 1 | ASC | 0.82 |
| 2 | ZCD | 0.92 | 2 | MFCCV4 | 0.76 |
| 3 | ZCD_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 |
| 70 | ASEV23 | 0.63 | 9 | MFCCV8 | 0.62 |

**Principal Components Analysis**

The set of 79 chosen parameters related to mood of music consisted of parameters that were strongly correlated. Principal Component Analysis was performed to achieve possible most orthogonal dimensions [293]. PCA was applied to two sets: one consisting of 79 parameters related to Dimension 1 and 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.

Therefore vector describing Dimension 1 was shortened to 7 components and Dimension 2 to 6 components.

2D visualizations of PCA for each dimension are presented in Figs. 6.20 and 6.21. Although direct interpretation of components is not possible loadings analysis indicates that particular components are associated with specific parameters. For the clarity of the presentation only some of these are shown in Tab. 6.25.

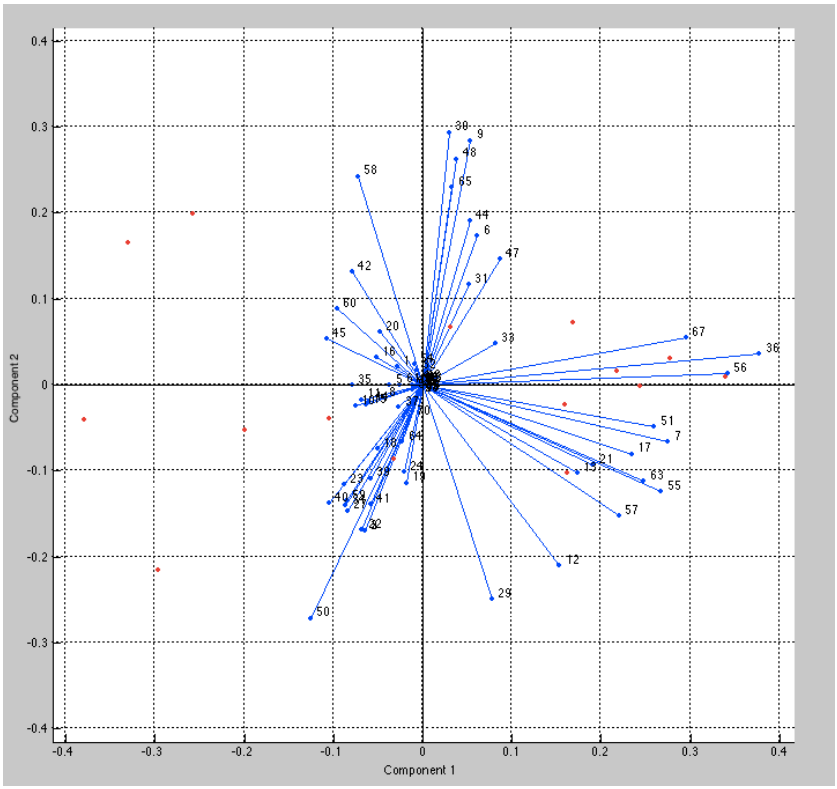



*Figure 6.20* *Graphical representation of PCA applied to 70 descriptors related to Dimension 1. Numbers refer to the parameters correlated to Dimension 1, listed in Tab. 6.24*

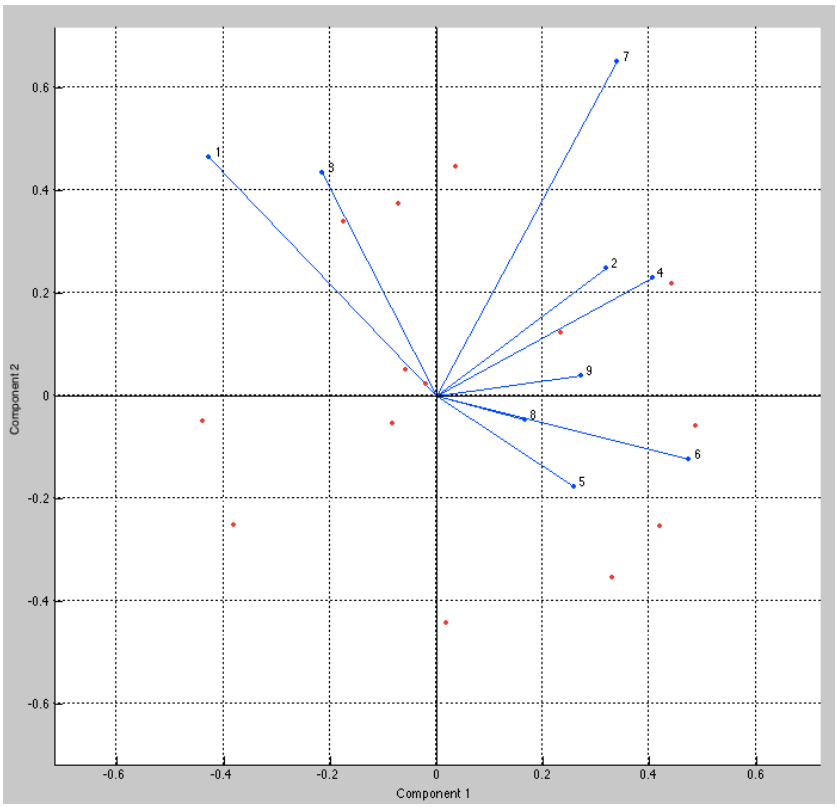



*Figure 6.21* *Graphical representation of PCA applied to 9 descriptors related to Dimension 2. Numbers refer to the parameters correlated to Dimension 2, listed in Tab. 6.24*

*Table 6.25 Maximum loading of particular components achieved from the PCA method. For clarity only values above 0.25 are presented*

| Comp. No. | Dimension 1 | | Comp. No. | Dimension 2 | |
|---|---|---|---|---|---|
| | Parameter | Load. | | Parameter | Load. |
| 1 | MFCCV7 | 0.40 | 1 | SFM8 | 0.25 |
| | MFCC5 | 0.26 | | SFM9 | 0.26 |
| | MFCC10 | 0.47 | | SFM_M | 0.27 |
| | MFCCV6 | 0.34 | | SFM15 | 0.28 |
| 2 | ASC | 0.47 | | SFM12 | 0.30 |
| | MFCCV4 | 0.25 | | SFM13 | 0.34 |
| | SC | 0.44 | | SFM14 | 0.38 |
| | MFCCV6 | 0.65 | 2 | ASE21 | -0.27 |
| 3 | MFCC5 | -0.55 | | SFM20 | -0.25 |
| | MFCC10 | 0.78 | | SFMV4 | 0.26 |
| 4 | SC | 0.56 | | SFMV3 | 0.28 |
| | MFCCV7 | -0.32 | | SFMV2 | 0.29 |
| | MFCC5 | 0.61 | 3 | SFM8 | -0.30 |
| | MFCC10 | 0.35 | | SFM12 | -0.27 |
| 5 | MFCCV4 | -0.39 | | SFM19 | 0.32 |
| | SC | 0.46 | 4 | SFM13 | -0.37 |
| | MFCCV7 | 0.45 | | SFM14 | -0.34 |
| | MFCCV8 | 0.56 | | SFM12 | -0.25 |
| 6 | ASC | 0.47 | | SFMV4 | 0.31 |
| | MFCCV4 | 0.59 | | SFM8 | 0.37 |
| | MFCCV6 | -0.47 | | SFM7 | 0.42 |
| | ASE1 | 0.31 | 5 | ASE22 | 0.25 |
| | | | | MFCC4 | 0.26 |
| | | | | SFMV3 | 0.28 |
| | | | | ASE20 | 0.34 |
| | | | | ASE21 | 0.42 |
| | | | 6 | SFM14 | -0.36 |
| | | | | SFM15 | -0.28 |
| | | | | SFMV4 | -0.26 |
| | | | | MFCC2 | 0.37 |
| | | | 7 | SFMV15 | -0.32 |
| | | | | ASE14 | -0.27 |
| | | | | MFCC2 | 0.46 |

**Results**

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 2-dimensional SOM representation the best results were achieved for grid topology with network dimensions 5x5. In this case, a vector of parameters consisted of 13 elements 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. 6.22. Songs are placed on neurons with the highest activation.

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 good representation for Dimension 1 ("Calm"), shown in Fig. 6.23. 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.

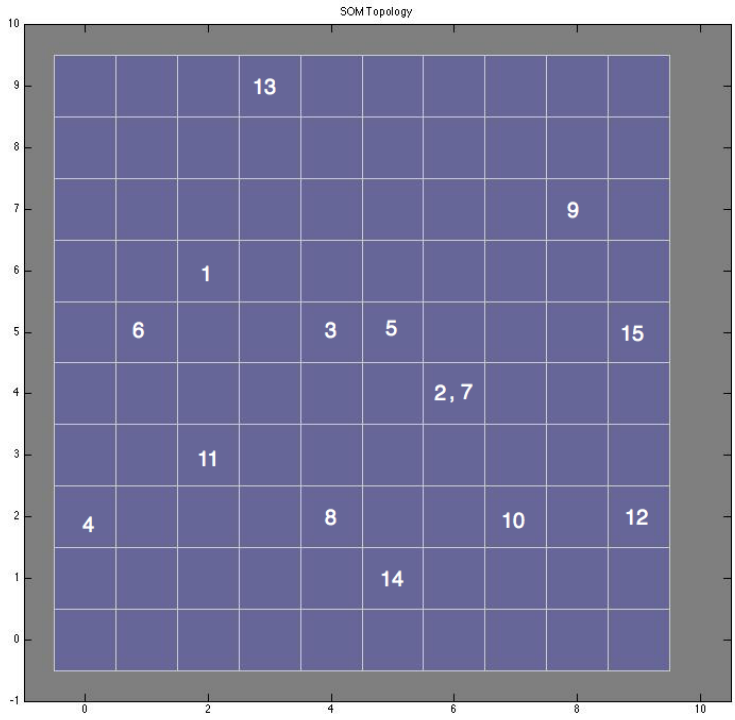



*Figure 6.22 Example of 2D SOM (5x5. grid topology) representation of 15-elements music set. Numbers represent particular songs. listed accordingly to Tab.6.17. Studies of the particular cases allow observing quite good results in one of the dimensions*

Contrarily representation of Dimension 2 ("Joyful") is less accurate but also contains correct assignments and is presented in Fig. 6.24. Wrongly placed songs are marked in the picture with the oval.

Even though, the accuracy for Dimension 1 is around 90% and for Dimension 2 around 70%, it is much higher than for 2D representation. Therefore the approach for two separate SOMs seems to be more appropriate for this task.

**Discussion**

Due to better results returned by two separate SOMs (one per dimension), this approach seems to be more appropriate for this task. Dimension 2 (Joyful) was more difficult to represent which is coherent with previous findings that it is much harder to find parameters correlated with "Valence" and responsible for "positive" or "negative" mood of music.

SOM mapping based on the 2-dimensional model of emotions is returning promising results. Therefore this model seems to be reasonable for employing computational methods and analysis. Although it should be remembered that meaning of dimensions used in this analysis was not 100% clear to the subjects, especially Dimension 2, which is correlated with "Joyful" with correlation coefficient of 0.78 (according to results achieved in Section 6.4).

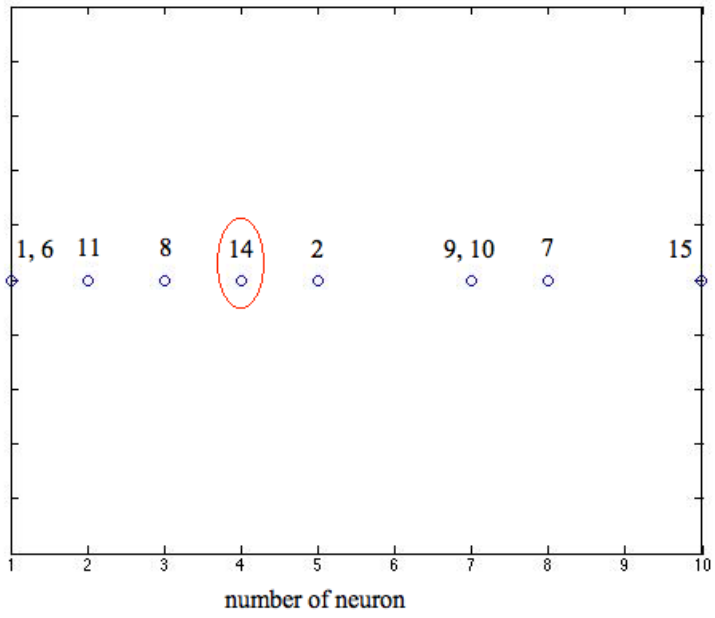



*Figure 6.23 SOM representation of 10-elements music set for Dimension 1 ("Calm"). Numbers represent particular songs, listed accordingly to Tab. 6.17. Song labeled with no. "14" is marked according to the inaccurate location*

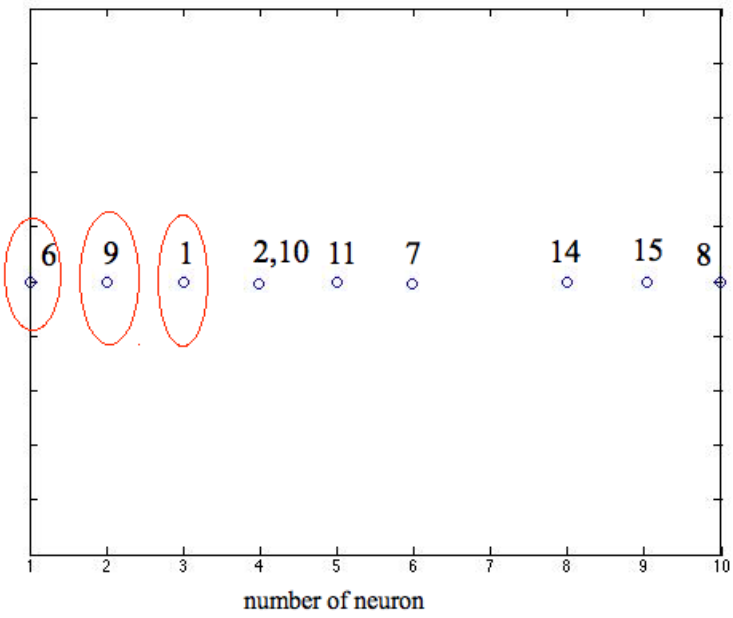



*Figure 6.24 SOM representation of 10-elements music set for Dimension 2 ("Joyful"). Numbers represent particular songs, listed accordingly to Tab.6.17. Songs located improperly are marked with ovals*

## 6.6 MOOD OF MUSIC EVALUATION BASED ON COLORS

Colors are intuitive and commonly used representation of mood. The aim of the next experiment was to determine whether the evaluation based on color intensity could be directly translated to numerical scale. Tab. 6.26 includes information with regard to experiment aimed to create a dictionary appropriate for music mood description in Polish.

*Table 6.26 Color scale experiment*

| Title | Color scale used for music mood representation |
|---|---|
| **Objectives** | To determine whether the evaluation based on color intensity could be directly translated to the numerical scale. |
| **Protocol** | - 15 samples evaluated with numerical and color scales<br>- 36 subjects |
| **General Results and Conclusions** | - Correlation between results was close to 1<br>- Color and numerical scale are equivalent and can be used alternatively |

### Detailed description of the experiment

The listening test consisted of 15 samples from different music genres. Duration of every music excerpt was constant and lasted for 30 seconds. The complete and detailed list of music tracks is listed in Tab. 6.17 (used also in the MDS Experiment described in Section

6.4). The subjects were asked to evaluate the extent to which a given label describes mood of the particular music excerpt. For the purpose of the survey mood labels were presented in Polish and they can be found, along with their English counterparts, in Tab. 6.12. The labels were chosen during previous research study, which was conducted to create a dictionary associated with mood of music in Polish (described in Section 6.1) [242].

The musical excerpts were randomly ordered and the sequence was different for each listener. Samples were presented in 128 kbit/s MP3 format. For the purpose of the experiments a survey was implemented as a HTML website in a series of simple HTML pages (Fig. 6.25, *in Polish*). The test began with the introduction page, where subjects were instructed about their task and could playback the music excerpt to adjust the volume.

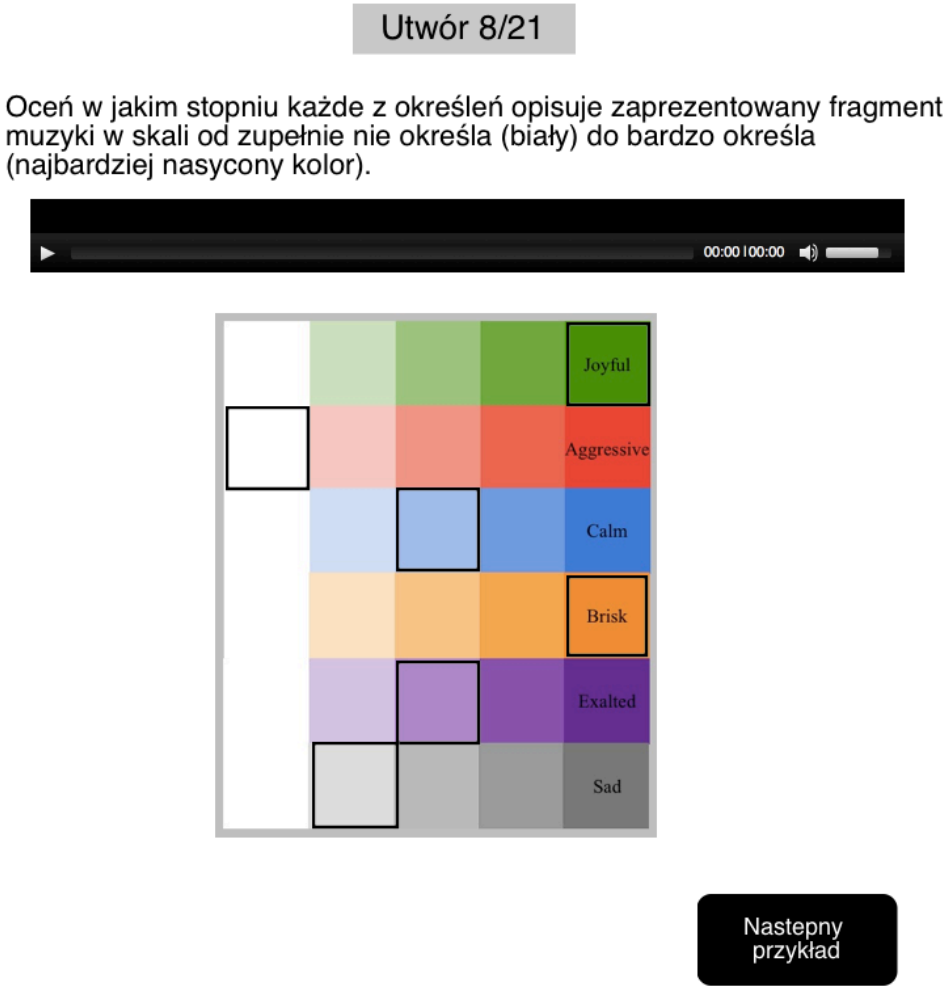



*Figure 6.25 Web interface used in the color experiment (in Polish)*

Then listeners were guided through subpages with playback and the evaluation form. This part consisted of two series, 15 samples each. In the first part scale was represented by matrix of colors with different intensity (Fig. 6.26).

Listeners were instructed as follows: "Evaluate, how particular mood label describes mood of music. Colors represent particular mood labels, intensity of color represents the extent to which label describes mood of music." In the second part matrix was replaced with numerical scale from 0 to 4. The listeners were instructed to mark notes on the scale from "**0** (does not describe at all) to **4** (extremely describes)." Both parts consisted of the same set of music samples. The order was random and different for both parts.


*Figure 6.26 Graphical representation of mood scale*

The entire test took approximately 20 minutes (including breaks between music samples and time needed for the answer). The panel of listeners consisted of 36 subjects: 24 men and 12 women. The average age of the subjects was 23 years and no problems with color perception were reported.

Comparison of the results achieved from numerical and color scale was performed. Outcomes are included in Tab. 6.27. Analysis shows that results are coherent in 99%, therefore scales applied are equivalent.

*Table 6.27 Correlation between results achieved for numerical and color scales*

| Mood label | Correlation |
|---|---|
| All (averaged) | *0.99* |
| Aggressive | *1* |
| Brisk | *0.99* |
| Exalted | *0.99* |
| Joyful | *0.99* |
| Sad | *0.99* |
| Calm | *1* |

**Conclusions**

Results achieved from both scales (color and numerical) are coherent and reach almost 100% agreement. Therefore scales can be used alternatively and results can be easily converted between them. It is worth emphasizing that listeners' comments with regard to the scale based on colors describe this way of evaluation as a very nice and intuitive. They also experienced that intensity of color reflecting the intensity of mood is a very clear and user-friendly idea. Therefore, it is reasonable to assume that the idea of colors representing emotions and color intensity representing intensity of emotions should be utilized in the user interface for evaluation and music searching.

This concept of graduate scale is also strongly related to fuzzy logic and the idea of "degree of membership". It includes various degrees of membership in between 0 (false) and 1 (truth), which is very coherent with the achieved results.

In this Chapter a series of experiments was presented. This is a research path that leads to the key experiment. Conclusions from all stages of the presented study were taken into consideration and are foundation of the key experiment. The final experiment was executed to collect subjective mood evaluation of a larger set of songs using graphical interface proposed by the author of this dissertation.

# 7 KEY EXPERIMENT

Main goals of preliminary tests were to determine appropriate vocabulary and check the methodology for the main experiment. After performing a series of preliminary tests (described in details in Chapter 6), conclusions from all stages (Fig. 6.1) were taken into consideration and the key experiment was designed and executed. The final experiment aimed to collect the subjective evaluation of a larger set of music from different music genres. Subjective evaluation is often considered as a "ground truth" for emotional content of music and is used as a reference point for automatic mood recognition [343]. It is very important what model of emotions is used for subjective evaluation how the data is collected. Author of the presented dissertation proposes model that is based on the outcomes of previous experiments and literature of the subject.

Various computational and artificial intelligence methods were used at the previous stages of this work. Preliminary experiments included analysis of parameters correlated with mood of music, different pre-processing of data and mapping based on SOMs. All these algorithms and solutions are tested on a bigger music set for which subjective data was collected and supplemented by additional methods.

The idea of the main listening test and detailed information about the procedure and results is included in subsequent sections. In next sub-sections, results of various analyses including SOM, ANN classification and fuzzy logic are presented and outcomes are compared with subjective evaluation. At the end of the chapter a tool created for intuitive presentation of the evaluation and recognition of mood of music is introduced.

## 7.1 LISTENING TEST

Tab. 7.1 includes information with regard to the main experiment that aimed for subjective emotional content evaluation of larger set music. As mentioned before, outcomes from previous experiments were taken into consideration and affected the final form of the main experiment including model of emotions, music set and procedure.

### 7.1.1 General Assumptions

The test was executed to collect subjective mood evaluation of larger set of songs using specially designed graphical interface (Fig. 7.1). The main part of the test consisted of the

series of musical excerpts presented one after another, where listeners were asked to evaluate the mood of music by clicking at the graphical mood representation. It was preceded by a short survey and level check. Experiment was performed in Polish and its total average duration was approximately 12 minutes. Translation of the mood labels used in the interface is presented in Tab. 7.2. The stages of the test are presented in Fig. 7.2 and are described in details in the subsequent section.

*Table 7.1 Main experiment*

| Title | Main experiment |
|---|---|
| **Objectives** | - subjective evaluation of larger data set in terms of mood of music<br>- proposed novel graphical model of emotions in music |
| **Protocol** | - 154 samples from 10 genres are evaluated using the proposed graphical interface<br>- 112 subjects |
| **General Results and Conclusions** | - 154 musical excerpts mapped onto graphical representation (equivalent with Energy/Arousal plane)<br>- relation between mood of music and genre examined<br>- a set of features related to mood of music |

### 7.1.2 Proposed Model of Emotions

One of objectives of this work was to create an intuitive, proprietary model used to describe the mood music Model of emotions used in graphical interface was created and designed by author of presented dissertation. During the phase of preliminary tests, numerous observations were noticed. MDS experiment indicated 2 dimensions underling the mechanism of perception of mood of music (Section 6.4), Dimension 1 related to "Calm" and Dimension 2 related to "Joyful". At the same time, SOM mapping based on this representation returned promising results using these two dimensions (Section 6.5). Experiment described in Section 6.5 leads to the conclusion that description using 6 mood labels with scaled intensity is coherent with MDS representation and they can be used alternatively. Moreover, listeners reacted very positively to the idea of a color representing particular emotion and intensity of color reflecting the intensity of the mood and found this method very intuitive (Section 6.6).

All observations from previous stages of the research (Chapter 6) were taken into consideration and thence, model presented in Fig. 7.1 was proposed. Steps of the model creation are described in subsequent Section and presented in Fig. 7.3.

Main assumptions were that model has to be intuitive for users and compatible with dimensional model consisting of two dimensions. Since listening test described in Section 6.1 showed that 2-dimensional model is not very intuitive for listeners, alternative solution was proposed. The set of mood labels was selected from the dictionary retrieved from Experiment A (Section 6.1) to use vocabulary intuitive for listeners in context of mood of music. Mood labels (originally in Polish) along with their translation can be found in Tab. 7.2.

Mood descriptors were placed on a 2-dimensional plane, with regards to dimensions retrieved from MDS experiment (Section 6.5) - Joyful and Calm (Fig. 7.4a). This placement is coherent with Thayer's model [308] and Russel's [264] emotion representation, which are described in details in Section 2.5.1. It is also consistent with findings of Brinker et al., [43], who examined relationship between Thayer's Valence/Arousal (VA) model and twelve mood labels. Their results are cited in Section 3.2.1 and shown in Fig. 3.1 and 3.2. Selected labels are also consistent with Hevner's list of adjectives [108], which can be found in Section 2.5.2.

In the next step colors were assigned to particular emotions. The selection was based on Plutchik's Color Wheel Of Emotion [249] and the idea of emotions placed close to each other being represented by similar colors (Fig. 7.3b). Simultaneously, the concept of scaling the emotion was introduced (Fig. 7.3c). The intensity of color corresponds to the intensity of particular emotion contained in music (Fig. 7.3d). "White" area placed in the center is considered as a neutral, where no emotional content is included. Graduation introduced for each label along with intensity of color was previously tested in experiment described in Section 6.6, where listeners found this kind of representation very intuitive. This concept is also strongly related to fuzzy logic and the concept of "degrees of truth". It includes various states of truth in between 0 (false) and 1 (truth), which is very intuitive, when it comes to evaluation of such a subtle substance as emotions.

Combination of labels, color representation and graduate scaling resulted in a final model presented in Fig. 7.1. This representation is intuitive for users but also compatible with 2-dimensional mood models that were successfully used at previous stages of the research.

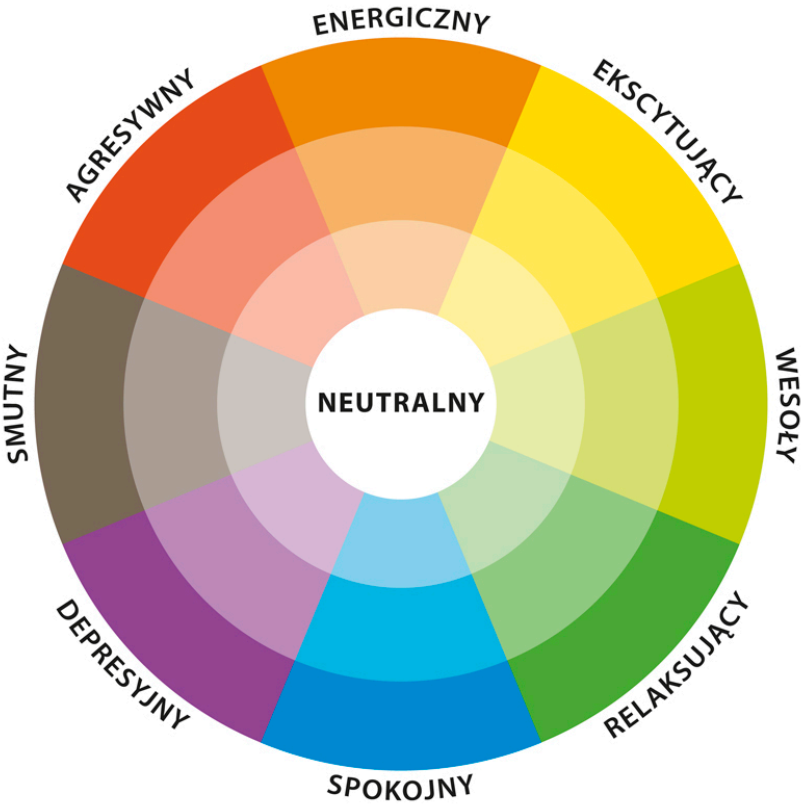



*Figure 7.1 Graphical interface dedicated for mood of music evaluation*

*Table 7.2 List of mood labels used in graphical interface designed for mood of music representation*

| No. | Mood label (Polish) | Mood label (English) |
|---|---|---|
| 1 | Agresywny | Aggressive |
| 2 | Depresyjny | Depressive |
| 3 | Ekscytujący | Exciting |
| 4 | Energiczny | Energetic |
| 5 | Neutralny | Neutral |
| 6 | Relaksujący | Relaxing |
| 7 | Smutny | Sad |
| 8 | Spokojny | Calm |
| 9 | Wesoły | Joyful |

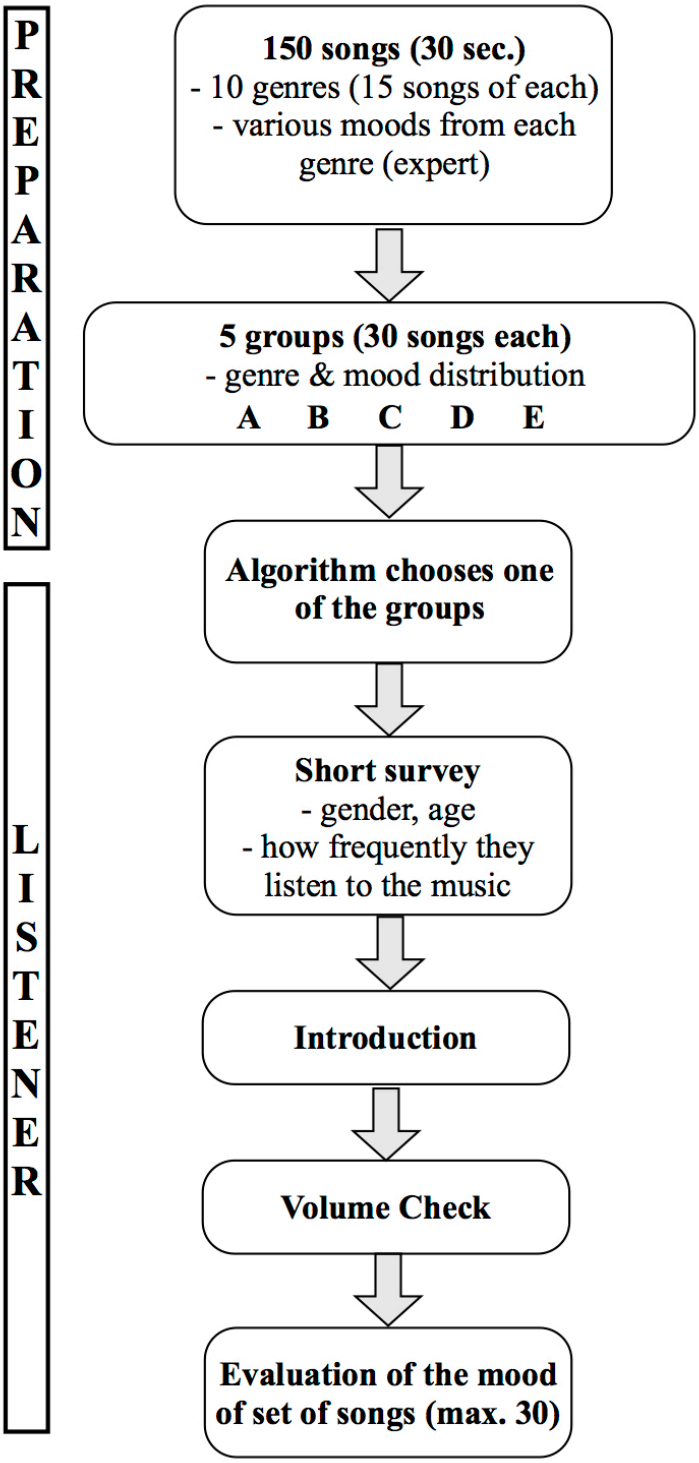



*Figure 7.2* *Main test arrangement related to music mood evaluation*

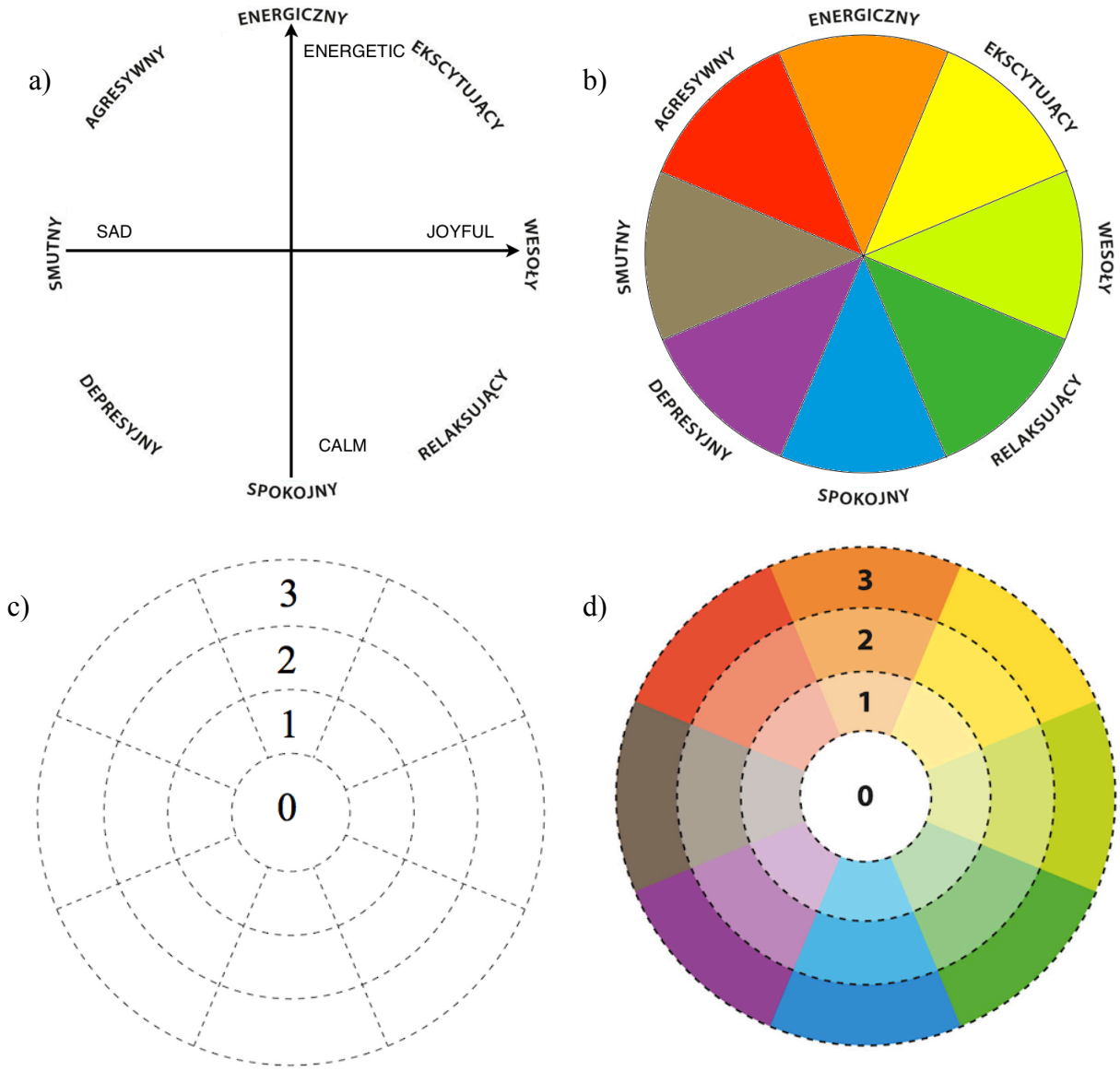



*Figure 7.3 Creation of model of emotions used in the key experiment. Different parts show particular concepts introduced in model: a) mood labels placed on the 2-dimensional model, b) colors representing emotions, c) graduation of mood, d) graduation of colors equivalent to intensity of emotion*

### 7.1.3 Listening Test

150 tracks were chosen from different 10 music genres, to obtain a diversified set. Chosen music styles are listed in Tab. 7.3. It is worth noting, that styles which were chosen are easy to distinguish between each other and cover various music material. 30-sec music pieces used in the experiment came from the SYNAT music database [151,155]. Additionally, four songs, multi-track recordings, that are analyzed in Section 4.6 are added. The complete and detailed list of the music tracks is listed in Appendix I and four additional musical excerpts are indexed as A1-A4. A musical excerpt was listened to and its mood was evaluatedby the author of this dissertation to collect data with a diversified music material.

*Table 7.3 List of music genres that were involved in the main experiment*

| **Music genre** |
|---|
| Blues |
| Classical |
| Country |
| Dance & DJ |
| Hard Rock & Metal |
| Jazz |
| Pop |
| R&B |
| Rap & Hip-Hop |
| Rock |

154 music pieces were divided into five sets (app. 30 pieces each), labeled from A to E. Songs were distributed among the set, so each of them contained all of the genres and all 'pre-judged' moods.

The test was performed using a WEB-based interface designed using Heroku platform [106]. Listeners were entering the website and were directed to a short survey, where questions about genre, age and how frequently they listen to the music, what is their main source of music were asked. Then they were guided through the introduction, where the main concept of the test and the graphical interface were introduced. Subjects were familiarized with the idea of the mood representation (Fig. 7.1) and especially with the intensity of colors that represents the intensity of the particular mood (Fig. 7.4).


*Figure 7.4* *Graphical representation used in the experiment during introduction, presenting how intensity of colors represent the intensity of particular mood*

After introduction phase, listeners were advised to enter the actual test. Then listeners were guided through subpages (Fig. 7.5) with playback and the evaluation interface. For each subject, the particular set of songs was assigned. The algorithm proposed by the Author was choosing the set (from A to E), so the same number of answers was given. The order of the tracks was random and different for each listener.

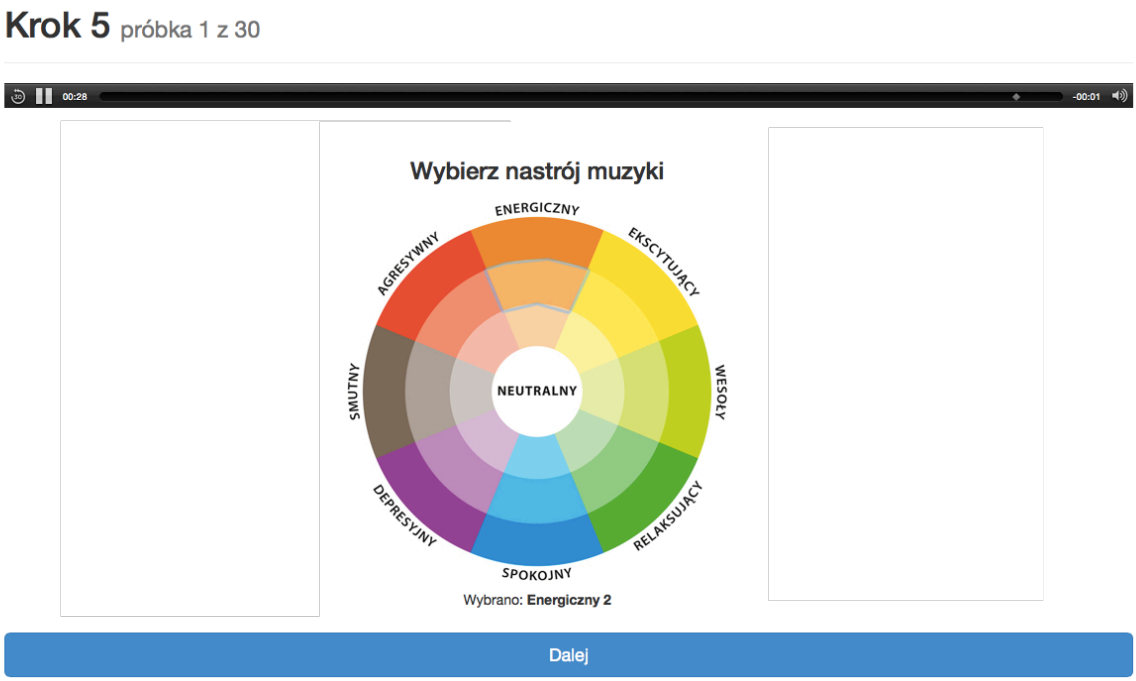



*Figure 7.5* *Web interface used in the main experiment*

112 listeners (57 women and 55 man) within age from 16 to 56 (average age 28) participated in the experiment. Majority of the audience reported that they listen to music everyday (Fig. 7.6).

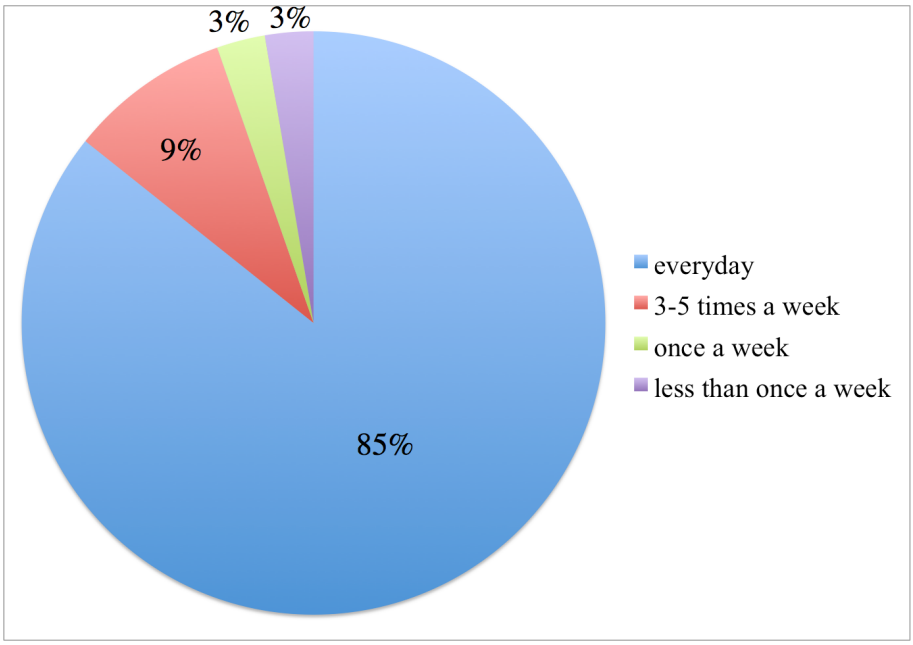



*Figure 7.6 Results of the survey in which the subjects were asked how often they listen to the music*

### 7.1.4 Results and Discussion

Answers provided by subjects were downloaded from the WEB service [106] as an XLS file. Data were pre-processed and only valid entries were included in result analysis. Submission was considered as valid if consisted of 5 to 30 evaluated songs.

Results were analyzed with both dimensional and label approaches. In the dimensional approach answers were analyzed in polar coordinates. To each field on the graphical interface, the number value was assigned according to the explanation (Fig 7.4) and angle was assigned according to the position of the label, this is presented in Fig. 7.7. This allowed mapping onto 2D Energy/Arousal plane. Achieved results were used as polar and Euclidean coordinates depending on employed method. Detailed results of the listening test are presented in Appendix II in a form of table containing averaged Energy and Arousal values along with their standard deviations and polar coordinates.


*Figure 7.7* *Value assigned to each label along with its intensity and position on the model*

Results are shown in a schematic way in Figs. 7.8 and 7.9. Representation in Fig. 7.10 includes also mood labels that are included in the model used in the tests. Translation of the mood labels (originally in Polish) can be found in Tab. 7.2. All songs are marked with red "x" signs. Labels are not included for the clarity of the presentation. The same results assigned according to music genre are included in Fig. 7.10. For discussion of results quadrants of the AV plane are numbered from I to IV as presented in Fig. 7.8.

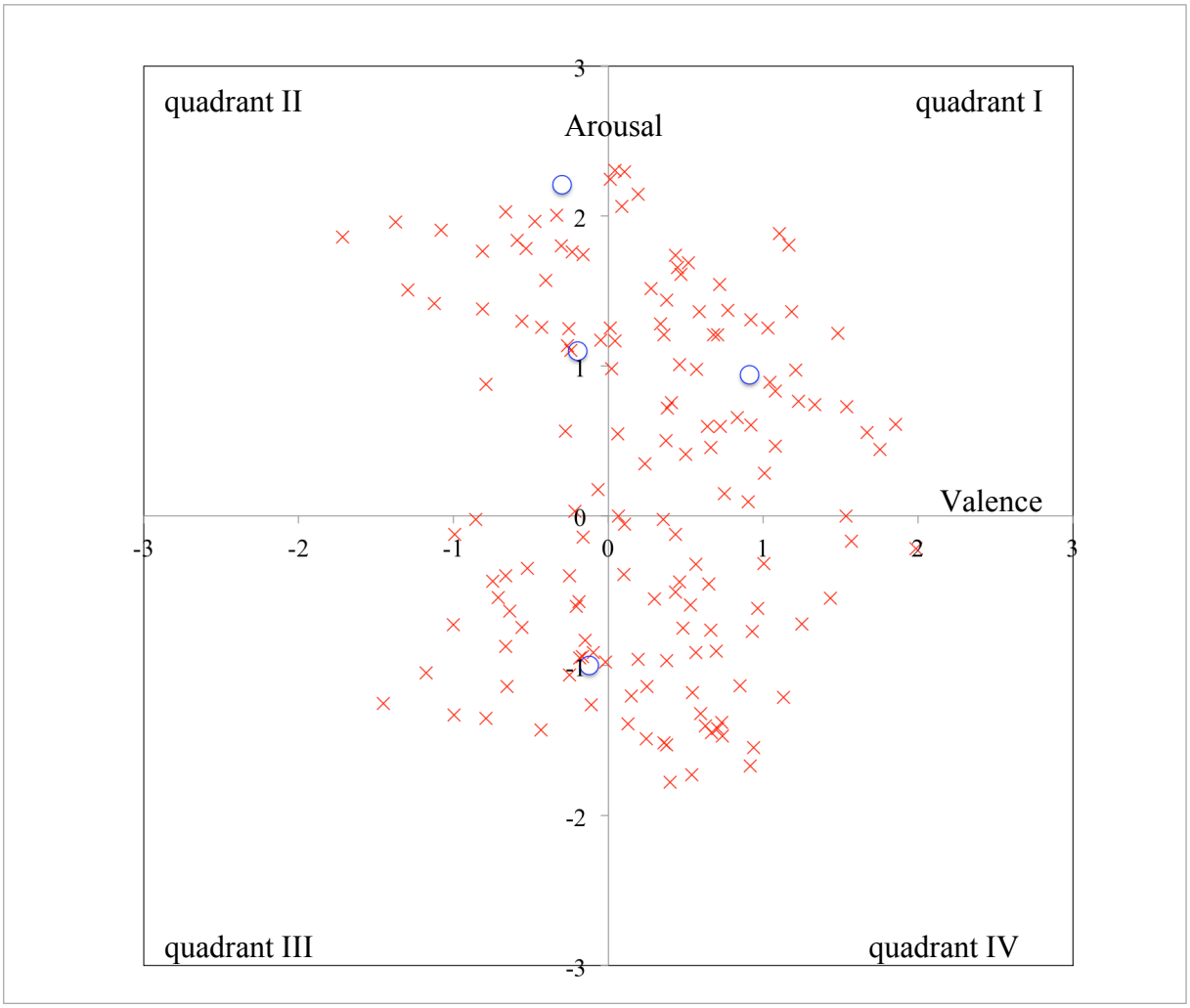



*Figure 7.8 Mapping of 154 songs onto mood plane based on the listening test results. "x" signs represent 150 songs from SYNAT and "o" represent tracks, which were very thoroughly analyzed in Section 4.6*

This map is coherent with distribution of 50000 songs on the Valence/Arousal plane presented by Kim [131] (Fig. 3.10) as well as findings of Brinker *et al.* [43] who observed that the left part of VA plane for popular music is actually rather empty. This part of the VA plane does not naturally occur in Western popular music. Details of referred works are cited in Section 3.2. It is important to remember that the music set was pre-judged and selected to cover variety of emotions included in music.

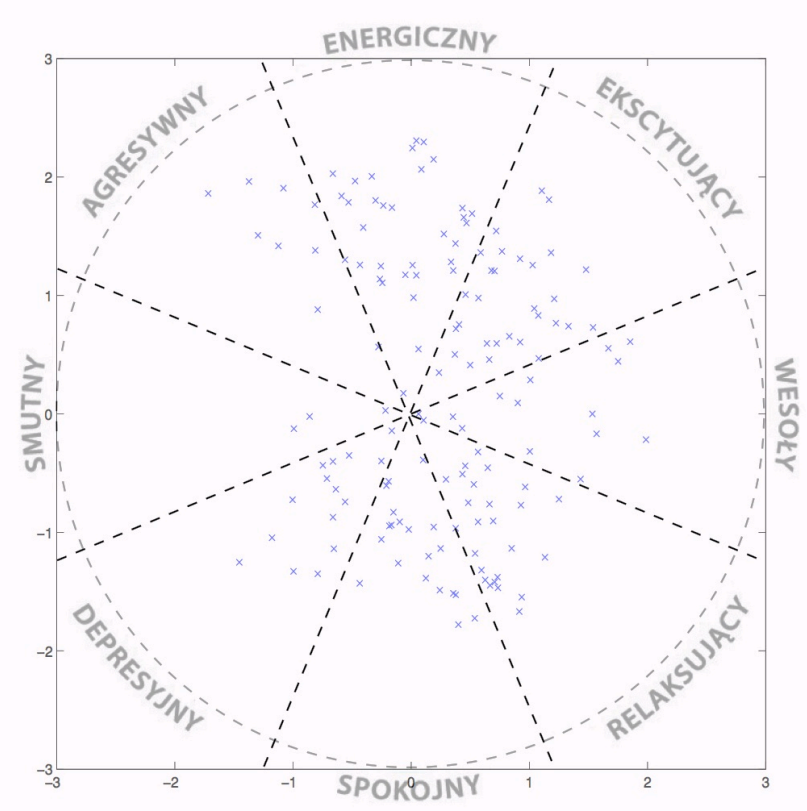



*Figure 7.9* *Mapping of 150 songs onto mood representation including mood labels (translations are listed in Tab. 7.2). "x" signs represent songs*

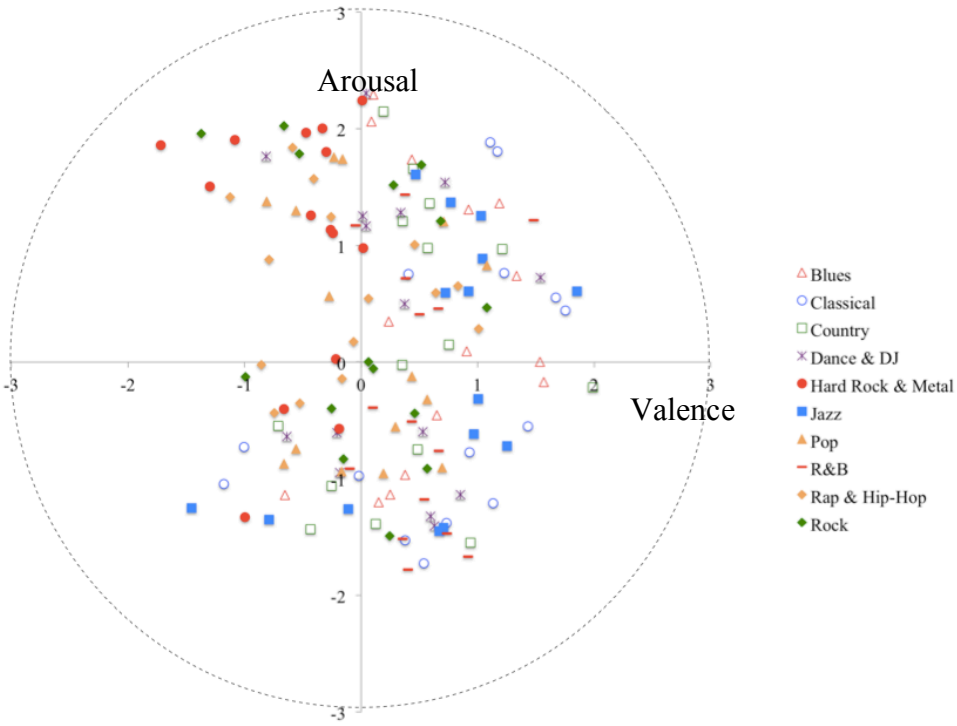



*Figure 7.10* *Mapping of 150 songs (divided by the genre) onto mood plane based on the listening test results*

Averaged ratings are mostly within range [0,2]. Only for aggressive mood a stronger result was obtained. In this case mood of music might be strongly related to the perceived emotion. Also distribution of music pieces sorted by genre is very interesting. For clarity results separated according to music style are shown in Figs. 7.11 (Jazz, Hard Rock & Metal, Pop and Rock) and 7.12 (Blues, Classical, Country, Dance & DJ, Rap & Hip-Hop, R&B), and described in subsequent paragraphs. Quadrants are discussed according to the nomenclature presented in Fig. 7.8 and songs are numbered according to Appendix I. Tab. 7.4 contains averaged results for different music genres along with values of standard deviation of Valence and Arousal within these genres.

In Fig. 7.11 genres, for which an additional analysis on separate tracks was performed in Section 4.6, are included. Excerpts, which were analyzed using tracks of separated instruments are marked with blue circles and numbered A1-A4 accordingly.

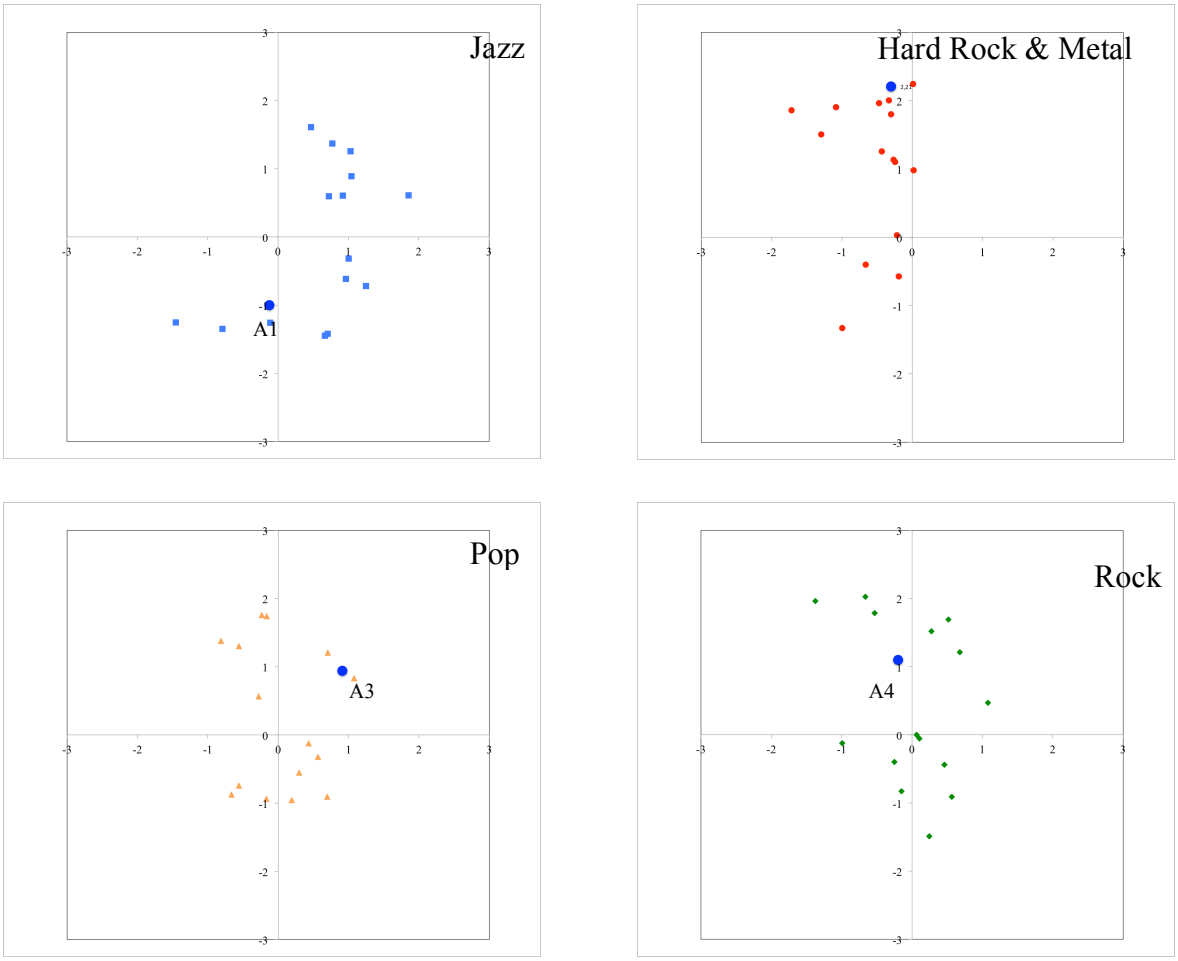



*Figure 7.11 Mapping of songs divided by the genre (Jazz, Hard Rock & Metal, Pop, Rock onto mood plane based on the listening test results. Additional tracks A1-A4 are indicated by blue circles*

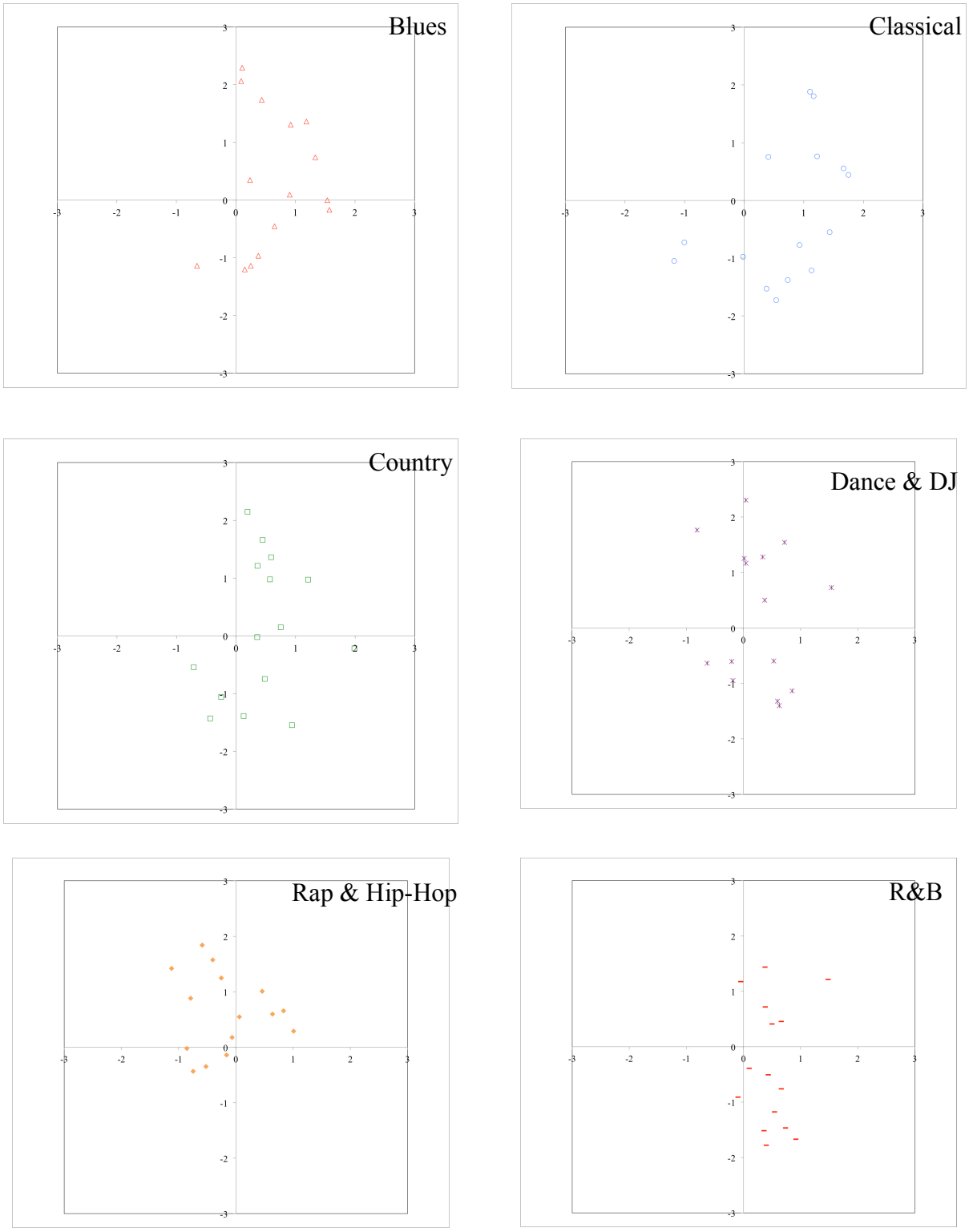



*Figure 7.12* *Mapping of songs divided by music genre (Blues, Classical, Country, Dance & DJ, Rap & Hip-Hop, R&B) onto mood plane based on the listening test results*

For pop and rock music, excerpts are distributed in all quadrants of the AV plane. There is no jazz songs in III quadrant, even though some songs could be called "Aggressive" i.e. Marcus Miller, "What Is Hip?", no. 82 or John Coltrane, "Giant Steps", no. 77. At the same

time, Hard rock & Metal genre was evaluated as containing negative energy and occurs only on the left side of the VA plane, distributed mostly over II quadrant. Additional songs A1 - A4 fit general trends within their styles.

Blues was placed by listeners mostly on the right half of the VA plane (positive) as well as classical, where only two pieces were placed in quadrant III. A similar tendency is observed for country excerpts. Rap & Hip-Hop was considered mostly as music with high arousal and appeared mostly in quadrants I and II. Some excerpts are placed in quadrant III, i.e. Tyler, The Creator, "Her", no. 124, which is rather slow and has very little energy of low frequencies.

Distribution of Blues and R&B was quite similar (mostly quadrants I and IV), which might be related to musical similarities and common roots of both genres. Although blues is often considered as very sad music, it was placed on the "positive" half of VA plane.

What is interesting, Dance & DJ excerpts are distributed among all quadrants as well as Pop and Rock. These music genres are very frequent in popular culture, therefore are not strongly related to one and only esthetics or topic but are mixtures of different trends. That is also reflected in values of the standard deviation, which are highest for these genres (Tab. 7.4).

Ratings for all songs in each genre were averaged and results are listed in Tab. 7.4 and mapped onto the AV plane shown in Fig. 7.13. Centroids for particular music genres are presented using the same marks representing music styles as in Figs. 7.11 and 7.12. Average rating for all genres except of Hard Rock & Metal are within radius of 1. Hard Rock & Metal was evaluated with strong ratings in quadrant II and that is visible also in this graph as well as a corresponding smaller value of the standard deviation of Valence (Tab. 7.4). Standard deviation for Valence is also small for R&B, because most of songs within this style were considered as positive. Standard deviations are highest for jazz (Tab. 7.4), i.e. songs are placed in quadrants I, II and IV with quite strong ratings (Fig. 7.11).

All observations listed above are important cues for conducting experiments with decision systems. **Although they are related to the specific music set, they might represent more general trends due to carefully selecting excerpts for the performed evaluation.**

*Table 7.4 Averaged results for various music genres*

| Music genre | Averaged Valence | Averaged Arousal | St. Dev. Valence | St. Dev. Arousal |
|---|---|---|---|---|
| Blues | 0.60 | 0.33 | 1.11 | 1.16 |
| Classical | 0.68 | -0.25 | 1.28 | 1.25 |
| Country | 0.44 | 0.10 | 1.05 | 1.02 |
| Dance & DJ | 0.25 | 0.26 | 1.06 | 1.15 |
| Hard Rock & Metal | -0.55 | 1.03 | 0.99 | 1.09 |
| Jazz | 0.60 | -0.10 | 1.19 | 1.33 |
| Pop | 0.04 | 0.22 | 1.04 | 1.20 |
| R&B | 0.49 | -0.32 | 0.99 | 1.12 |
| Rap & Hip-Hop | -0.17 | 0.62 | 1.10 | 1.12 |
| Rock | 0.00 | 0.43 | 1.24 | 1.13 |

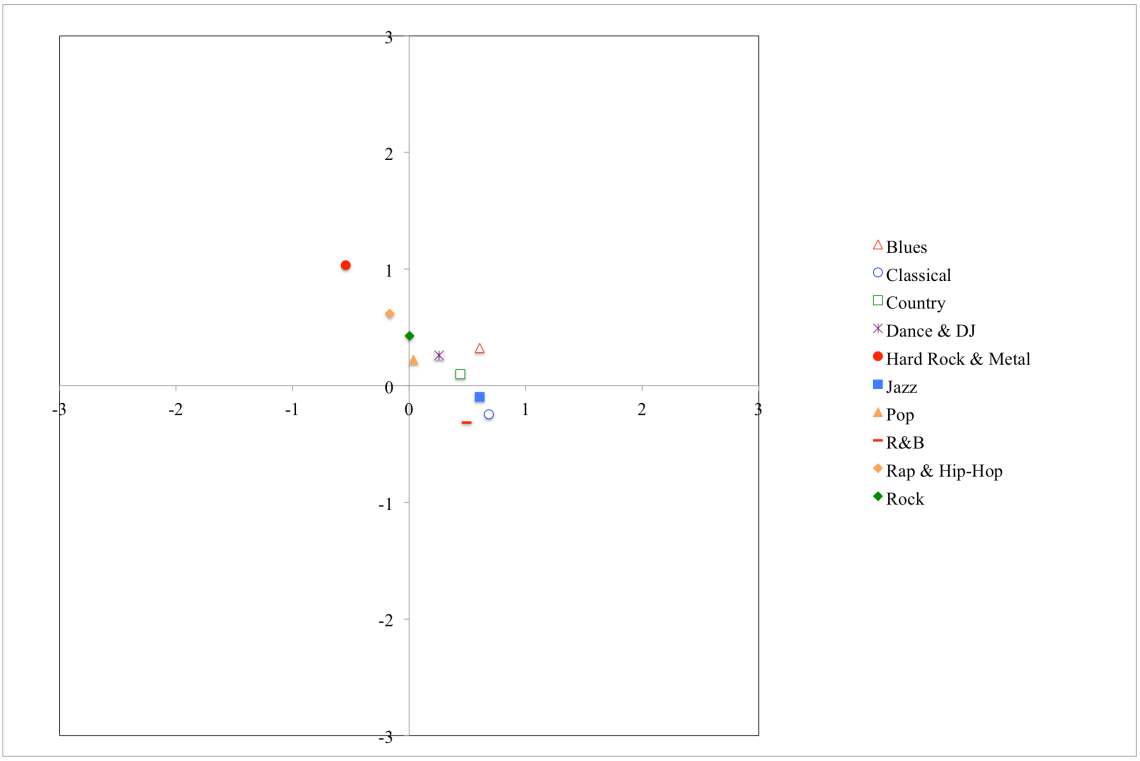



*Figure 7.13 Centroids for particular music genres*

Contrarily, the label analysis was performed, where the number of occurrences of each label was calculated for every song. As a result, each song is described with a 9-element

vector, where each position refers to mood label (Tab. 7.1). The value describes the percentage of occurrences of each label. Examples of songs along with their label description are shown in Fig. 7.14. Some songs are described by all listeners with mainly one label (i.e.: Death, "Story to tell", no. 62 in App. I and III), while for other evaluations is spread among the labels (i.e.: Guy Davis, "Watch Over Me", no. 11 in App. I and III). Detailed label results for all songs are presented in Appendix III.

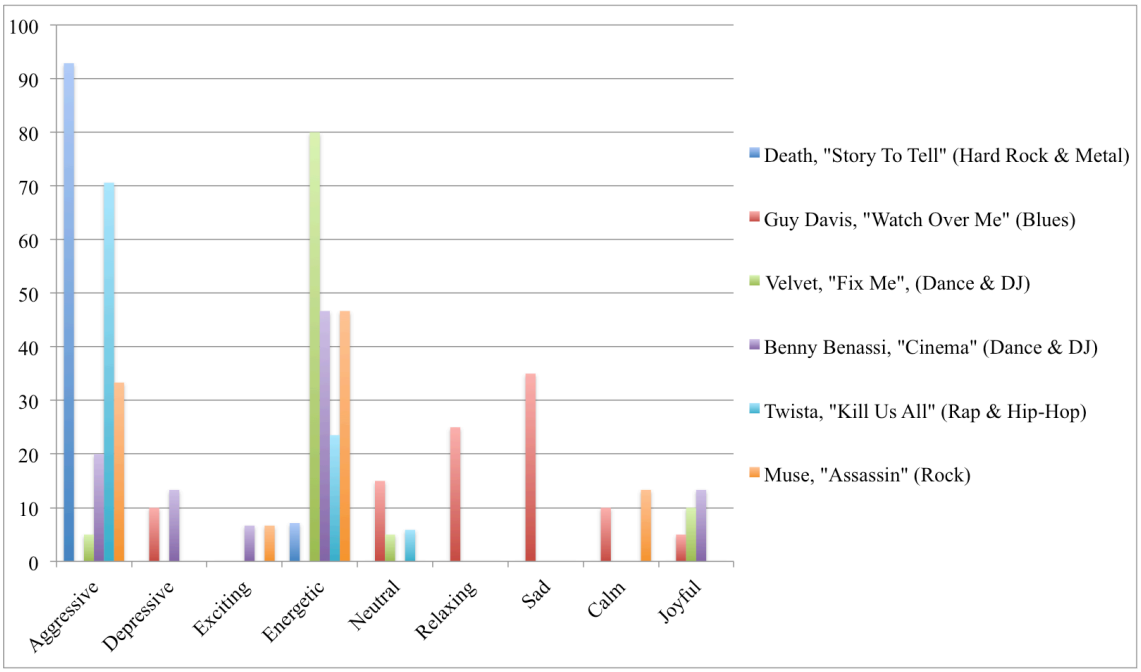



*Figure 7.14 Example of results of mood labels assigned to particular songs. The vertical axis describes the percent of occurrences of each label*

One of the main observations during analysis of results was that process of combining a point on the VA plane with music is inconvenient and distracting. Therefore a concept of a tool that would enable connection between point in the VA plane and music excerpt was introduced. Its design and implementation are described in the subsequent Section.

## 7.2 VISUALIZATION OF MOOD IN MUSIC RECOGNITION

Visualization and comparison of subjective results and automatically assigned mood of music is challenging. Therefore a tool to present the concept of musical excerpts organized according to the emotional content was created. It was designed to enable intuitive presentation of the results of listening tests and automatic mood recognition. The idea was to place objects on the graphical mood of model that was introduced in the description of the main listening test (Section 7.1).

The visualization tool is implemented and designed in Max 7 software [62]. This environment enables interactive processing of audio and image files and signals and is commonly used by audio and visual artists in various art installations [31,72,94,135]. The programming interface is designed in a "patcher " (Fig. 7.15).

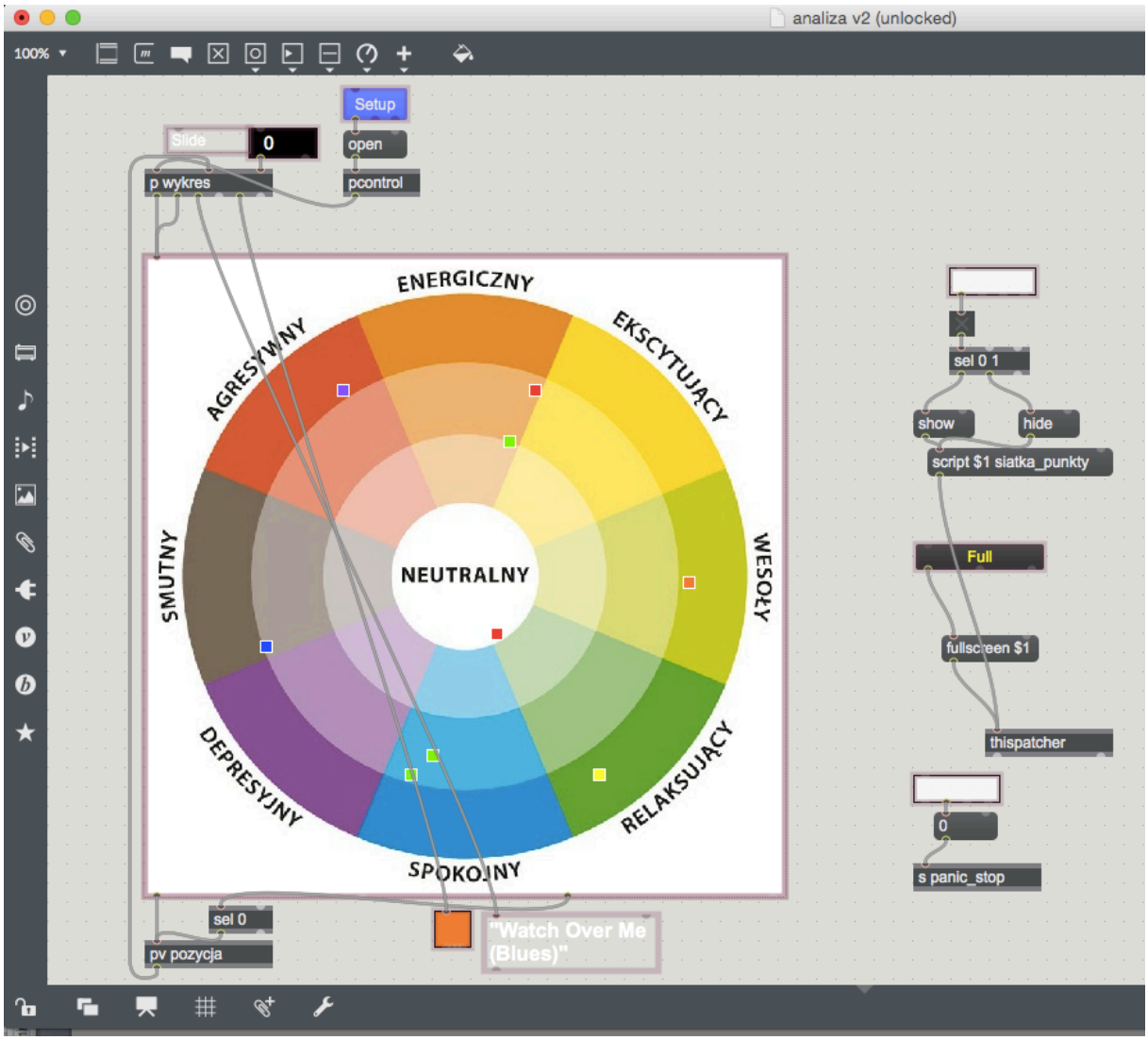



*Figure 7.15 Programming process of visualization tool in Max 7*

Musical excerpts are represented by small squares, which are placed on the mood model. After mouse click on the particular object playback of chosen music piece, which mood of music corresponds to the point on the model is started and description of the song is shown (Fig. 7.16).

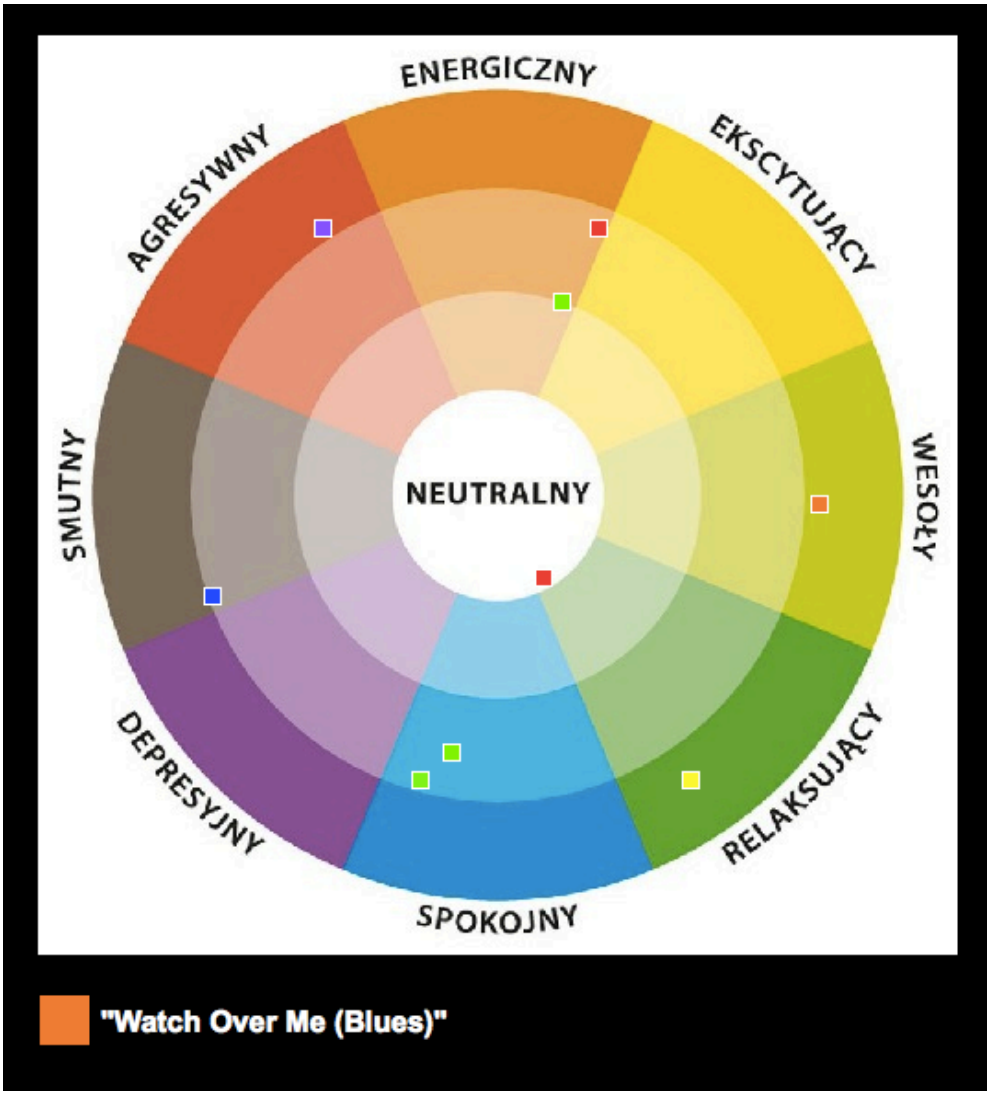



*Figure 7.16 Music fragments placed on the mood map. Mouse click on the object triggers playback of a song which mood of music corresponds to the point on the model. Detailed information about played song, including artist, title and genre, is placed in the bottom part of the interface*

Colors of the squares are related to music genres according to caption included in Fig. 7.17.

Mapping of whole set of 154 musical excerpts used in key experiment is presented in Fig. 7.18 with respective position on the VA plane and music genre annotation.

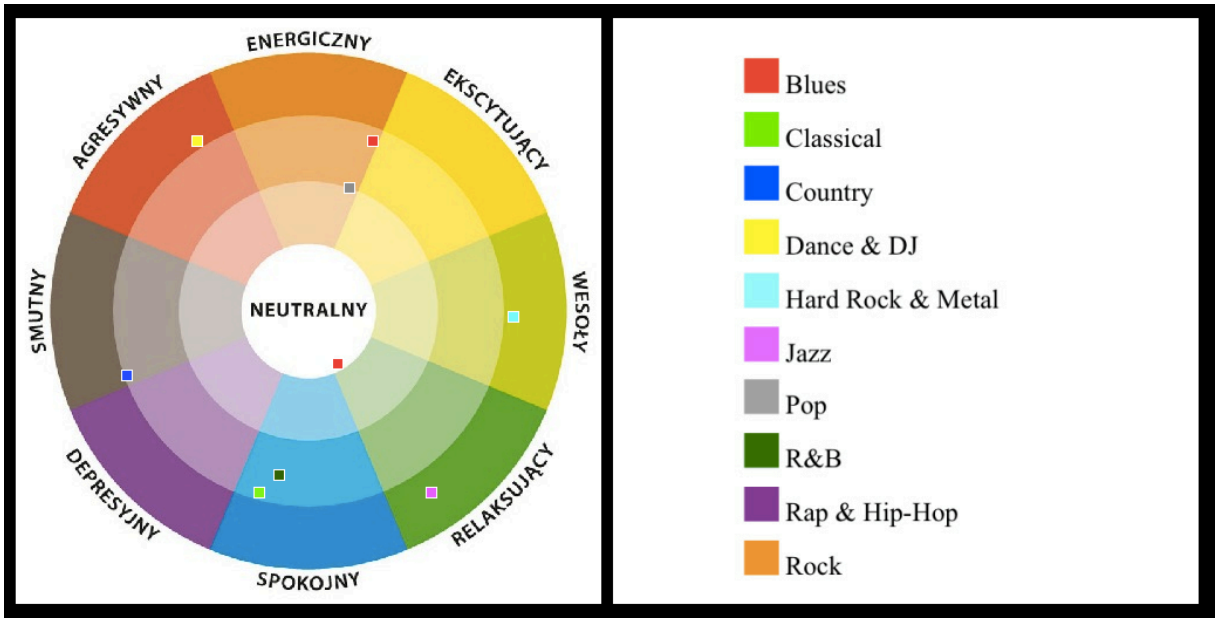



*Figure 7.17* *Visualization tool designed in MAX 7. Squares indicate songs, while color of squares represent music genre according to the legend on the right side*

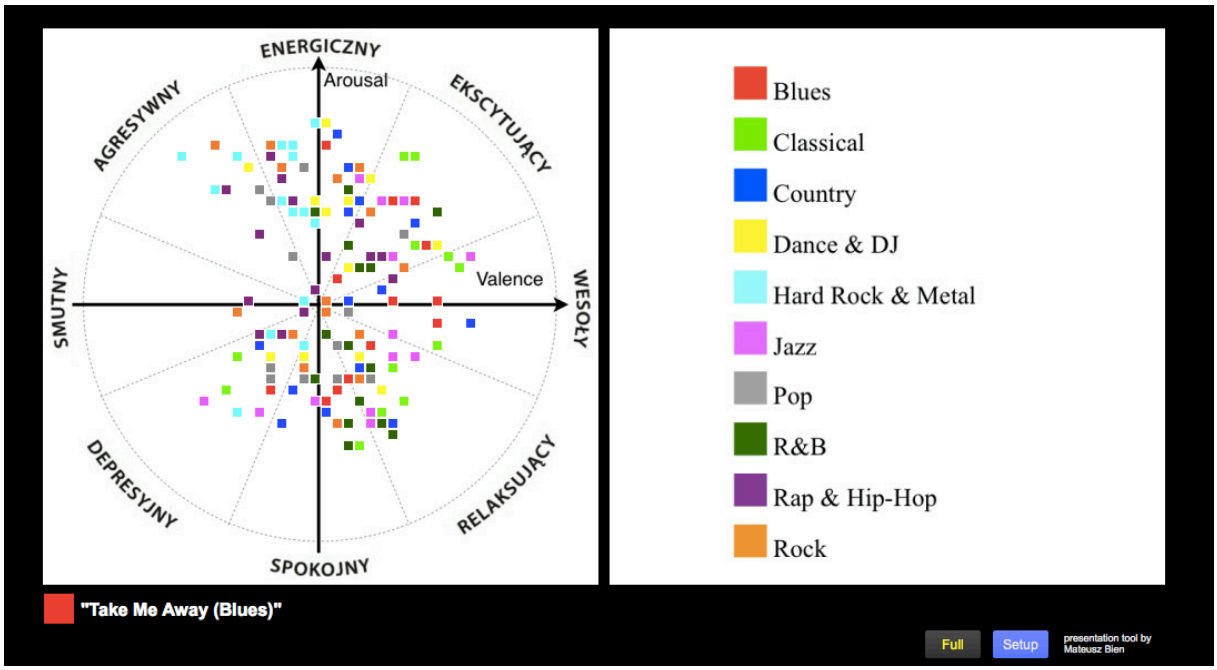



*Figure 7.18* *154 songs used in the key experiment (listed in App. I) mapped using MAX 7 visualization tool according to subjective evaluation of mood of music*

The playback interface presented in previous Section was also consulted with a group of listeners, who did not participate in evaluation of mood stage of the key experiment. The test was informal and aimed for retrieving their comments how the whole concept is intuitive in terms of exploring music. Their observations were often leading to conclusion that it is very hard to determine the intensity of mood on a crisp scale. At the same time, some of them pointed out that also transition between emotions should be blurred. All

these remarks lead to the idea of fuzzyfying the boundaries in proposed model of mood. A model with fuzzified boundaries of emotions and fuzzified intensity of mood is presented in Fig. 7.19.

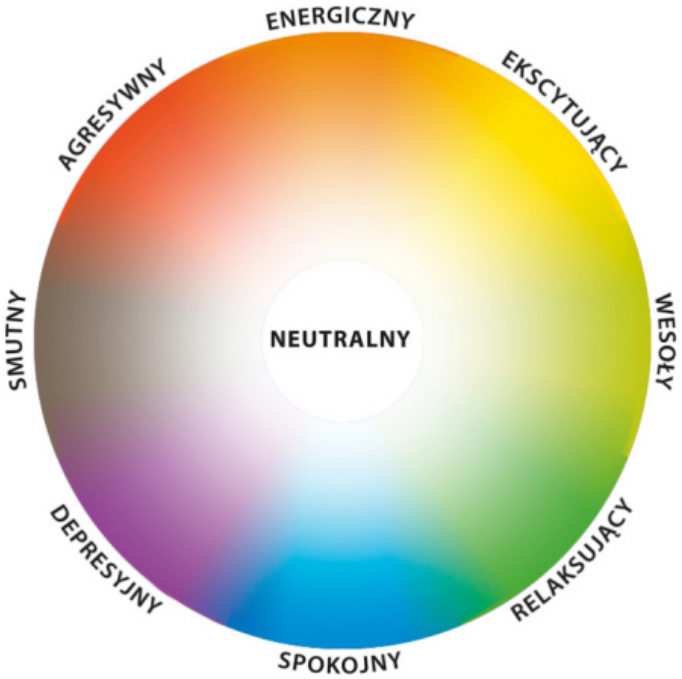



*Figure 7.19 Proposed modified model of mood with fuzzified boundries of emotions*

The set of songs used in the key experiment was mapped onto the model with fuzzified boundaries and this representation is shown in Fig. 7.20. Consultants described this projection as more intuitive and this is coherent with the intuitive concept that there is no crisp ranges representing mood and emotions, contrarily the transition area is rather blurred. This is also closely related to a human's perception of music and leaves space for interpretation. Due to fuzzification that occurs in two dimensions: intensity of mood and transition between emotions, model based on fuzzy logics should also be two-dimensional.

These observations lead to the conclusion that approach based on fuzzy logic might be appropriate for the area of Music Emotion Recognition. Therefore attempt to discussion related to employing this method in MER is presented in subsequent Section.

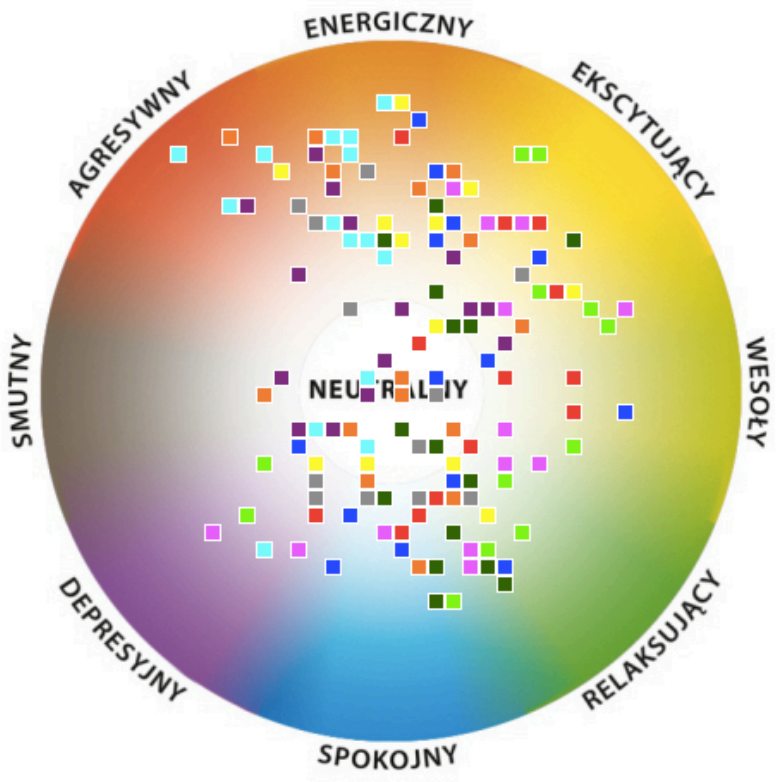



*Figure 7.20* *154 songs used in key experiment (listed in App. I) mapped according to subjective evaluation of mood of music into a model with fuzzified boundaries*

## 7.3 APPROACH BASED ON FUZZY LOGIC

Conclusions from previous Section as well as publications of Blewitt [36] and Jun and collaborators [125] imply that an approach based on fuzzy logic in the area of MER is very intuitive and might lead to interesting results. As described in Section 5.7, relations between emotions in music and music characteristics described by Hevner [108] are based on the concept of rules, which can be considered as fuzzy. According to her findings (Tab. 4.2), i.e.:

**IF tempo is LOW, and minor mode content is HIGH, THEN the mood of music should be DREAMY.**

**IF major mode content is HIGH, and flowing rhythm content is MEDIUM, THEN the mood of music should be PLAYFUL.**

**IF low pitch content is HIGH, and minor mode content is HIGH, THEN the mood of music should be PATHETIC.**

Similar rules can be created for results achieved in the course of the presented experiments. Some rules can be defined by an expert or based on the analysis of outcomes

and can result in membership function creation. For experts, rules based on futures related to music characteristics are easier to define due to more clear interpretation of their meaning. An example of conditional statement is given below:

**IF tempo is LOW and SPECTRAL BRIGHTNESS is low then mood of music is DEPRESSIVE.**

Corresponding membership functions related to tempo (BPM) and brightness (calculated according to description in Section 4.6.3) are presented in Fig. 7.21. The consequent is also considered as fuzzy, taking into consideration fuzzification between mood labels and mood intensity.

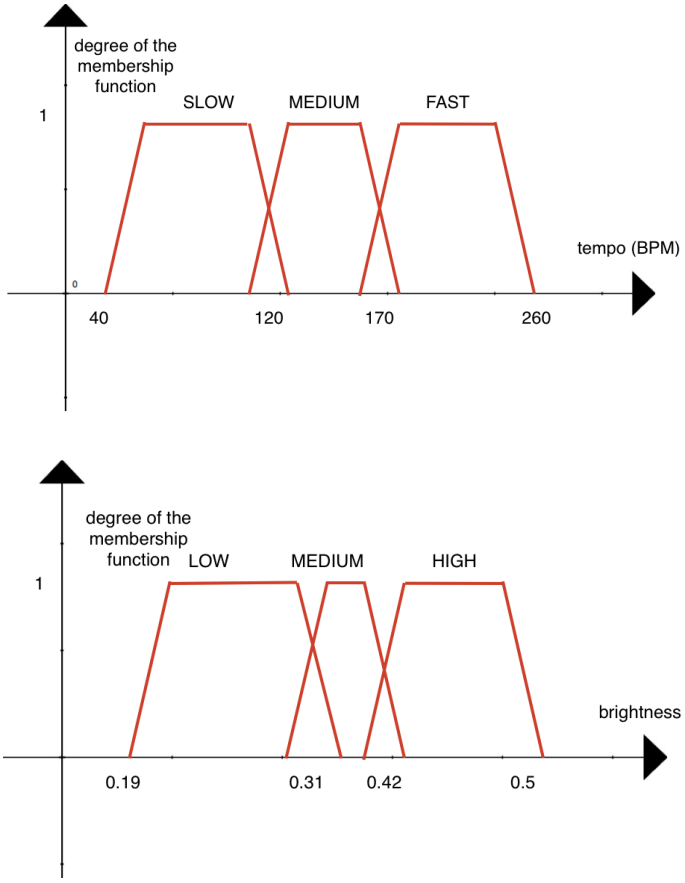



*Figure 7.21 Example of membership functions related to a rule dedicated to mood of music*

The antecedents of the proposed conditioning statement are based on tempo and brightness, therefore songs evaluated by listeners as "Depressive" are listed in Tab. 7.5 with corresponding values.

Fuzzy logic should be considered in further studies on automatic mood of music recognition and this approach can be developed into a full fuzzy logic-based system.

*Table 7.5 Objects evaluated by listeners as "Depressive along with tempo and brightness, which values are premises in the proposed conditioning statement. Tracks are named according to App. I*

| No. | Genre | Tempo (BPM) | Brightness (normalized) | Remaining FV parameters | Mood of music label |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| 3 | Blues | 110 | 0.25 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 18 | Classical | 101 | 0.17 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 25 | Classical | 134 | 0.37 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 40 | Country | 109 | 0.41 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 58 | Dance & DJ | 139 | 0.36 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 69 | Hard Rock & Metal | 105 | 0.37 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 73 | Hard Rock & Metal | 109 | 0.25 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 78 | Jazz | 105 | 0.13 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 79 | Jazz | 114 | 0.08 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 101 | Pop | 131 | 0.28 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 102 | Pop | 121 | 0.28 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 123 | Rap & Hip-Hop | 74 | 0.38 | ... | Depressive |
| 124 | Rap & Hip-Hop | 63 | 0.43 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 133 | Rap & Hip-Hop | 98 | 0.32 | ... | Depressive |
| ... | ... | ... | ... | ... | ... |
| 147 | Classic Rock | 108 | 0.40 | ... | Depressive |

## 7.4 CORRELATION ANALYSIS

For all 154 musical excerpts FVs consisting of 173 parameters from SYNAT (described in Section 4.5) were retrieved and additional parameters from MIR Toolbox (Section 4.6.3) and time-based features (TBF) proposed by the author (Section 4.6.2) were calculated. This set of parameters contains a lot of information but not necessarily related to mood of music. At previous stages of this study a selection of parameters based on correlation analysis was

performed and returned good results, therefore this approach was applied. Correlation between subjective values of Valence and Arousal and parameters was calculated and therefore a set of features strongly related to mood was created. Only parameters with correlation coefficient higher than 0.50 were included in the final feature vector. Eventually, the feature vector describing mood of music consisted of 16 parameters from SYNAT, listed in Tab. 7.6 and 17 parameters from MIR Toolbox and MTBF for different ranges are presented in Tab. 7.7. It is worth noting that correlation is slightly stronger with parameters based on music characteristic than from SYNAT, which describe general properties of an audio signal. Description of parameters included in SYNAT can be found in Section 4.5, time-based features (TBF) and features from MIR Toolbox in Sections 4.4.1 and 4.6.3 and in more detail in the MIR manual [164].

For each dimension moderate and strongly correlated parameters were found, what proves the **Thesis no. 1,** which assumes that **it is possible to find parameters describing a musical excerpt, which are highly correlated with subjective mood labeling results.**

*Table 7.6 Parameters correlated with subjective mood of music evaluation selected from 173 SYNAT FV*

| No. | Valence | | No. | Arousal | |
|---|---|---|---|---|---|
| | **Parameter** | **Corr.** | | **Parameter** | **Corr.** |
| 1 | ASE2 | -0.72 | 1 | MFCC1 | -0.79 |
| 2 | MFCC7 | -0.62 | 2 | MFCC2 | -0.78 |
| 3 | PEAK_RMS10FR_MEAN | 0.51 | 3 | SFM13 | -0.63 |
| 4 | ASE_M | -0.50 | 4 | SFM12 | -0.61 |
| 5 | ASE26 | -0.50 | 5 | SFM14 | -0.56 |
| | | | 6 | SFM15 | -0.56 |
| | | | 7 | SFM10 | -0.53 |
| | | | 8 | 1RMS_TCD | 0.52 |
| | | | 9 | SFM_M | -0.51 |
| | | | 10 | SFM11 | -0.51 |
| | | | 11 | SFM16 | -0.50 |

*Table 7.7 Parameters correlated with subjective mood of music evaluation selected from MIR Toolbox related to music characteristics and proposed time-based features (TBF)*

| No. | Valence | | No. | Arousal | |
|---|---|---|---|---|---|
| | **Parameter** | **Corr.** | | **Parameter** | **Corr.** |
| 1 | *Spectral irregularity* | 0.73 | 1 | *Brightness* | 0.83 |
| 2 | *MTBF2* | 0.54 | 2 | *Entropy of Spectrum* | 0.79 |
| 3 | *Spectral roughness* | 0.52 | 3 | *Timbre Zerocross* | 0.64 |
| | | | 4 | *Tonal Harmonic Change Detection* | 0.64 |
| | | | 5 | *Harmonic Change Detection Function* | 0.62 |
| | | | 6 | *Spectral Centroid Mean* | 0.57 |
| | | | 7 | *Key Clarity* | 0.55 |
| | | | 8 | *Tempo* | 0.55 |
| | | | 9 | *Spectral Flux Period* | 0.55 |
| | | | 10 | *Spectral irregularity* | 0.53 |
| | | | 11 | *spectral rolloff 185* | 0.52 |
| | | | 12 | *Spectral Kurtosis* | 0.51 |
| | | | 13 | *MTBF1* | 0.51 |
| | | | 14 | *Spectral roughness* | 0.50 |

As in the previous stages, less parameters are correlated with dimension corresponding to positive or negative mood of music (here Valence) than with Arousal. Also the correlation reaches higher values for Arousal (up to 0.83) comparing to ones achieved for Valence (max. 0.73). This result is coherent not only with previous studies of the author [243,244] but also with outcomes of other research studies presented in the literature [43,232,254].

## 7.5 ARTIFICIAL INTELLIGENCE METHODS USED FOR MER

Mapping of a music set based on included emotions using Self-Organizing Maps was performed at the previous stage of the research (Section 6.5). SOMs returned promising results and therefore are implemented on the larger set of music. Additionally, classification employing algorithms with supervised manner of learning is performed. Performance of Artificial Neural Networks in the mood of music classification task is described and results from both methods, self-organizing and trained networks are then compared.

Vector of parameters used in classification consists of 33 elements (as listed in Tabs. 7.6 and 7.7) and it is based on the correlation analysis with results of subjective mood evaluation. Selection of features is described in the previous Section. Principal Component Analysis was used in pre-processing of data at prior stages of experiments and returned good results for SOM mapping (Section 6.5). Therefore it was also applied to the chosen vector of parameters.

PCA was performed separately for both sub-sets of features correlated with single dimension (Valence or Arousal) as well as on the whole feature vector consisting of 33 parameters. As a result, shorter feature vectors were obtained and outcomes of these analyses are presented in Tab. 7.8. For Valence, five components cover 99% of information, for Arousal 17 elements are sufficient. PCA calculated for the whole 33-element feature vector resulted in 20 components covering 99% of information.

*Table 7.8* *Number of PCA components covering 99% of information for different vectors of parameters correlated with mood of music*

| **Vector of parameters** | **PCA result (number of components covering 99% percent of information)** |
|---|---|
| 8 parameters correlated with Valence | 5 |
| 25 parameters correlated with Arousal | 17 |
| 33 parameters (cumulative vector consisting of parameters correlated with Valence and parameters correlated with Arousal) | 20 |

For the convenience of description in further analyses, sets of features used in SOM and ANN classification are organized and named according to the nomenclature given in Tab. 7.9.

*Table 7.9* *Data sets used in SOM- and ANN-based classification*

| **Data set** | **Description** |
|---|---|
| Data_V | 8 parameters correlated with Valence |
| Data_A | 25 parameters correlated with Arousal |
| Data_VA | 33 parameters (cumulative vector consisting of parameters correlated with Valence and parameters correlated with Arousal) |
| PC_V | 5 Principal Components calculated for parameters correlated with Valence (Data_V) |
| PC_A | 17 Principal Components calculated for parameters correlated with Arousal (Data_A) |
| PC_VA | 20 Principal Components calculated for 33 parameters Data_VA |

### 7.5.1 SOM Analysis

SOM analyses were performed in Matlab using *selforgmap* module, for various topographies and sizes of the neural network. Results easier for interpretation were returned using grid topology; therefore this one was implemented in the analysis. Accuracy of various SOM representations is listed in Tab. 7.10. Based on previous experience with mood description based on SOM (Section 6.5), separate SOMs (1D SOM) for Valence and Arousal were created, but for this task the 2-dimensional SOMs (2D SOM) were much more accurate and returned better results.

For 2D SOM representation, the best results were achieved for the grid topology with network dimensions of 5x5 feature vector consisting of 33 elements describing both dimensions. Better results were achieved for data after employing PCA. Examples of SOM representations are shown in Figs. 7.22-7.25, where the number of hits for a particular neuron is marked. It is worth noting that the empty area on the left side of the plane, which is prominent in representation, achieved from subjective evaluation (Fig. 7.8), is also visible on the map returned by 5x5 SOM (Tab. 7.10, row 9), which is shown in Fig. 7.23.

*Table 7.10 Accuracy of different classification setups. "Input" column contains information about data provided into input of ANN, "SOM setup" indicates size of SOM and "Accuracy" the performance of SOM*

| No. | Input | SOM setup | Accuracy |
|---|---|---|---|
| 1 | Data_V | 1D, 1x5 | 29% |
| 2 | Data_A | 1D, 1x5 | 34% |
| 3 | Data_V | 1D, 1x7 | 15% |
| 4 | Data_A | 1D, 1x7 | 19% |
| 5 | Data_VA | 2D, 5x5 | 58% |
| 6 | PC_V | 1D, 1x5 | 28% |
| 7 | PC_A | 1D, 1x5 | 38% |
| 8 | PC_VA | 2D, 3x3 | 54% |
| 9 | PC_VA | 2D, 5x5 | **67%** |
| 10 | PC_VA | 2D, 7x7 | 49% |
| 11 | PC_VA | 2D, 9x9 | 22% |
| 12 | PC_VA | 2D, 11x11 | 20% |

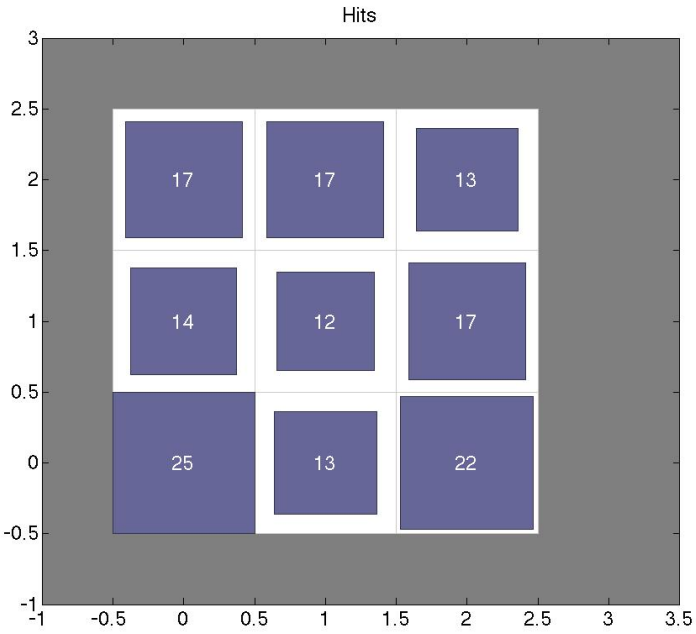



*Figure 7.22 Number of hits for each neuron for 2D SOM (3x3, grid topology) representation. 154-elements music set was mapped using PC_VA data. Accuracy achieved for this setup reached 54%*

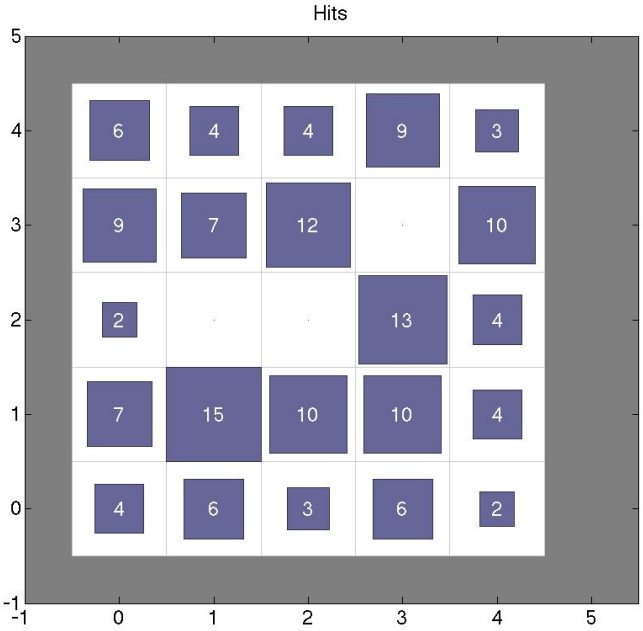



*Figure 7.23 Number of hits for each neuron for 2D SOM (5x5, grid topology) representation. 154-elements music set was mapped using PC_VA data Accuracy achieved for this setup reached 67%*

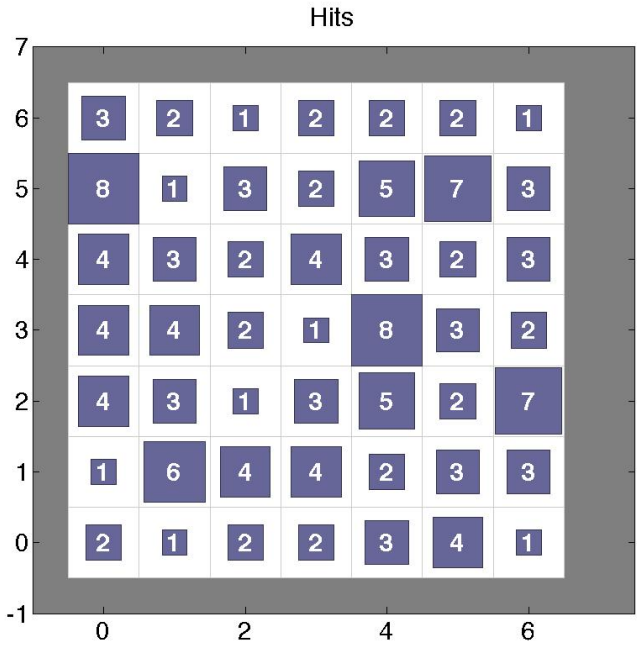



*Figure 7.24 Number of hits for each neuron for 2D SOM (7x7, grid topology) representation. 154-elements music set was mapped using PC_VA data. Accuracy achieved for this setup reached 49%*

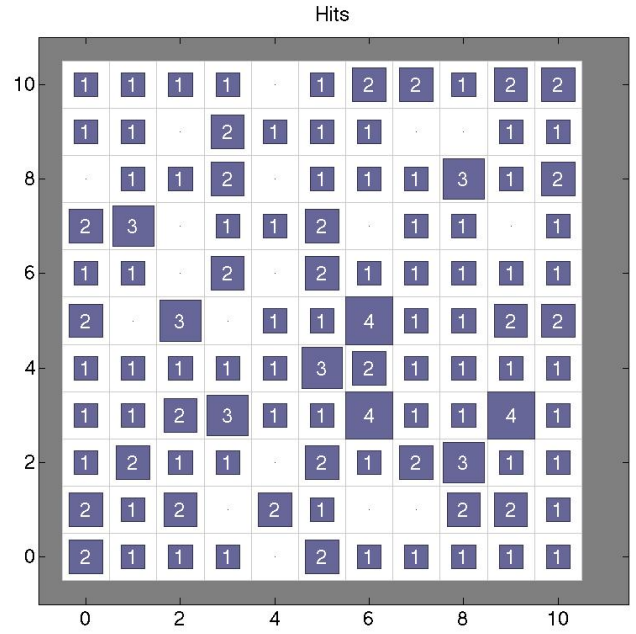



*Figure 7.25 Number of hits for each neuron for 2D SOM (11x11, grid topology) representation. 154-elements music set was mapped using PC_VA data Accuracy achieved for this setup reached 20%*

### 7.5.2 ANN-based Classification

The ANN-based classification was performed using *nntool* within the Matlab environment (MATLAB). The music set was divided into three subsets: training (70%), validation (15%) and testing (15%) randomly.

A feed-forward ANN with one hidden layer was trained to classify musical excerpts into four quadrants of Thayer's VA plane (described in Section 2.5.1). These values can vary depending on the data set. Various feature vectors listed in Tab. 7.9 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. Results strongly depend on the definition of classes. Results of various classification setups are presented in Tab. 7.11. Classification using eight classes (mood labels assigned to the parts of a circle - Fig. 7.3.b) returned low accuracy, therefore the idea of classes related to Valence and Arousal dimensions was introduced. For the purpose of the ANN-based analysis, dimensional results were normalized into [-1,1] range.

The highest accuracy was achieved for two separate networks - one dedicated to each dimension (Valence and Arousal) with input data containing only features correlated with a particular dimension. Overall, results of **83%** accuracy for Valence and **91%** for Arousal in classification into two halves of the VA plane were obtained. With increasing the number of classes, the output was fuzzified and returned results indicated position on VA plane with more detail. Outcomes of classification for four ranges in each dimension (Tab. 7.11, rows 13 and 14) also returned good results with app. 70% accuracy. Further increasing the number of compartments was decreasing the efficiency, down to 48-50% for 10 ranges. It is worth noting that data reduction based on PCA did not improve results therefore further ANN analysis using PCA were abandoned.

An interesting result was also obtained for classification described in rows 21 and 22 (Tab. 7.11). ANN was employed to separate all excerpts with low intensity of emotion (small radius in polar coordinates, objects close to the middle of VA plane). The accuracy of this classification was app. 84-85% depending on the radius. This operation slightly increased the accuracy of classification into set of eight mood rows (1 vs. 23). At the same classification for concentric classes representing similar intensity of emotion (radius) achieved low (rows 24 and 25), what shows that the dimensional-based approach enables better automatic mood recognition.

*Table 7.11 Accuracy of different classification setups. "Input" column contains information about data provided into input of ANN, "Classes" indicates number of classes and their definition and "Accuracy" the performance of ANN*

| No. | Input | Classes | Accuracy |
|---|---|---|---|
| 1 | data V+A | 8 (mood labels - Fig.7.3.b) | 45% |
| 2 | PCA_VA | 8 (mood labels - Fig.7.3.b) | 33% |
| 3 | data V+A | 4 (quadrants of VA plane) | 56% |
| 4 | PCA_VA | 4 (quadrants of VA plane) | 47% |
| 5 | data V | 2 (Valence (-1,0) and [0,1]) | **83%** |
| 6 | data A | 2 (Arousal (-1,0) and [0,1]) | **91%** |
| 7 | data V+A | 2 (Valence (-1,0) and [0,1]) | 74% |
| 8 | data V+A | 2 (Arousal (-1,0) and [0,1]) | 87% |
| 9 | PC_V | 2 (Valence (-1,0) and [0,1]) | 61% |
| 10 | PC_A | 2 (Arousal (-1,0) and [0,1]) | 52% |
| 11 | PC_VA | 2 (Valence (-1,0) and [0,1]) | 47% |
| 12 | PC_VA | 2 (Arousal (-1,0) and [0,1]) | 53% |
| 13 | data V | 4 (Valence [-1,-0.5), [-0.5, 0) , [0,0.5) and [0.5,1]) | 70% |
| 14 | data A | 4 (Arousal [-1,-0.5), [-0.5, 0) , [0,0.5) and [0.5,1]) | 71% |
| 15 | data V+A | 4 (Valence [-1,-0.5), [-0.5, 0) , [0,0.5) and [0.5,1]) | 65% |
| 16 | data V+A | 4 (Arousal [-1,-0.5), [-0.5, 0) , [0,0.5) and [0.5,1]) | 58% |
| 17 | data V | 10 (Valence [-1,-0.8), [-0.8, -0.6) ... [0.6, 0.8) and [0.8,1]) | 48% |
| 18 | data A | 10 (Arousal [-1,-0.8), [-0.8, -0.6) ... [0.6, 0.8) and [0.8,1]) | 50% |
| 19 | data V+A | 10 (Valence [-1,-0.8), [-0.8, -0.6) ... [0.6, 0.8) and [0.8,1]) | 45% |
| 20 | data V+A | 10 (Arousal [-1,-0.8), [-0.8, -0.6) ... [0.6, 0.8) and [0.8,1]) | 45% |
| 21 | data V+A | 2 (r [0,0.1) and [0.1, 1] | 84% |
| 22 | data V+A | 2 (r [0,0.2) and [0.1, 1] | 85% |
| 23 | data V+A | 8 (mood labels - Fig.7.3.b), * objects, where r [0, 0.1) were excluded | 51% |
| 24 | data V+A | 3 (r [0, 0.33), [0.33, 0.66) and [0.66, 1]) | 32% |
| 25 | data V+A | 3 (r [0.1, 0.4), [0.4, 0.7) and [0.7, 1]), * objects, where r [0, 0.1) were excluded | 35% |

## 7.6 COMPARISON OF RESULTS AND DISCUSSION

Analyses described in previous Sections clearly show, that methods based on artificial intelligence, thus closely related to a human's perception, are appropriate for the Music Emotion Recognition tasks. The size of the analyzed data set do not enable a detailed description of every single music excerpt; therefore examples of few most common trends are described in subsequent paragraphs. Musical excerpts are numbered according to Appendix I.

SOM analysis returns the 2-dimensional mapping of musical excerpts on the VA plane. Set of misclassified objects is different for every activation of the network, although some excerpts are often wrongly evaluated. This situation occurred often for songs with "neutral' mood, these with Valence and Arousal values close to 0 (i.e. Living Legends "Never Fallin' ", no. 133).

Some songs were difficult to describe in one of the dimensions. For example, Brian McKnight "The Rest of My Life", no. 108, was correctly assigned by ANN and SOM to the lower part of the VA plane, but both algorithms wrongly evaluated it as very sad.

There were also songs for which automatic description was less difficult, such as Rosin Coven, "Lion Song", no. 78, characterized by dull timbre, slow tempo and strange harmony clearly, which indicates depressive mood. Listeners evaluated this piece of music with high intensity of mood, and that is coherent with clear results returned by both algorithms. In general, most of songs, which were placed by listeners far away from 0 and axes were easier for automatic mood description, and overall they returned results that are coherent with subjective opinions.

It is also worth mentioning, that songs placed by listeners in II quadrant of the VA plane, were rarely misjudged by algorithms. At the same time pop and rock songs where often misclassified, what might lead to the conclusion that emotions contained in these genres are not as clear as for other music styles.

As a conclusion both, SOMs and ANNs, when employing the proposed feature vector, returned outcomes that enable automatic recognition of mood of music. It is important to emphasize that the approach proposed by the Author, namely using separate music tracks to introduce new features that better characterize mood specifics is very promising as it improved the accuracy of classification. Both methods returned mood recognition results, which can be used for music recommendation based on emotions included in music.

Therefore, it was proven that **Self-organizing maps (SOMs) or artificial neural networks (ANNs) trained employing designed feature vectors can effectively be applied to the automated indexing of mood of musical excerpts** (**thesis no. 2** of this dissertation).

A question arises whether it is possible to straightforward compare results obtained from unsupervised algorithms (SOM) and classification based on a supervised technique such as Artificial Neural Networks (ANN). Representation achieved using SOM reached up to 67% accuracy and ANN classification 91% accuracy for Arousal and 83% for Valence. Even though these outcomes are described by similar accuracy, their interpretation differs. SOM returns more specific information about the position of an object on the VA plane, while ANN returns less specific information but with higher accuracy. At the same time, data obtained from SOM are more difficult to interpret.

The accuracy achieved for both methods is sufficiently high, especially taking into consideration the specifics of the task. Moreover, results obtained from ANN and SOM classification are especially satisfactory, as a very high consistency of the results between listeners and a very good accuracy of the algorithms is not easy to obtain due to the subjective character of evaluation. It is as one expects a very good effectiveness from algorithms in cases when even experienced listeners' opinions differ. Also, it should be remembered that automatic mood annotation is still treated in the area of music querying as less important than search based on music genre, thus the results obtained in this dissertation, higher than those in other studies, may be one step closer to change this situation. Overall, the results achieved prove **thesis no. 3**: " **Annotations of mood of music achieved by subjective assessments and classifying based on both supervised and unsupervised learning can be coherent.**"

# 8 CONCLUSIONS AND FURTHER DIRECTIONS

This final chapter summarizes the study performed within the Ph.D. dissertation and its outcomes presented in this thesis. In addition, several extensions for future research that deserves to be explored are pointed out and briefly discussed.

The scope of this project was a framework of automatic organization of music based on the emotional content of music. A review of research studies related to music perception, MIR and especially MER pointed out issues that should be examined, such as the model of emotions used, features based on music characteristics and classification algorithms employed for automatic mood recognition.

In the course of this study numerous experiments were carried out. The first part involved quantitative and qualitative experiments and on this basis a dictionary related to mood of music in Polish was created. A Multidimensional Scaling experiment was executed to determine the dimensions underlying the perception of mood of music, hence the 2-dimensional model of mood with dimensions corresponding to "Joyful" and "Calm" was verified. In another experiment, the idea of the evaluation of mood of music using colors and their intensity was tested. A conclusion derived from that was that colors and intensity are a very intuitive method of music annotation and can directly be translated to the numerical scale and therefore can be implemented in interfaces dedicated to emotions included in music. In addition, a correlation between subjective evaluation results and objective parameters was performed and a vector of parameters related to mood of music was obtained. These results enabled verification and discussion of two classification algorithms, chosen from a variety of artificial intelligence algorithms, i.e. SOM and ANN, which were later employed in the final stage of this work.

Moreover, an overview of audio parametrization was presented with a special focus on parameters describing characteristics of music. These parameters were critically reviewed by the author, and a conclusion was that some additional parameters should be searched and examined. Therefore, an original analysis of single instrument tracks for different music genres was carried out. Sound material that enabled a deep study of information described by parameters in the case of a single instrument vs. the whole track was recorded and then selected by the author of this dissertation. The collected music material was also included in

the final, key experiment. In addition, own time-based features (TBF), describing rhythmic content in sub-bands, were introduced by the author and used in the final analysis.

The key experiment was based on conclusions derived from previous stages of the research carried out by the author. As a result, a model of emotions in music was created by the author and later implemented. The proposed model is intuitive for listeners and assigns colors as a representation of different emotions. The idea of color intensity corresponding to the intensity of emotion included in the music was incorporated into it. Moreover, the model includes mood labels achieved from the previous stage of experiments aimed at creating the dictionary related to mood of music. It is also coherent with the 2-dimensional mood representation, verified in the MDS experiment. Designing an original model of mood dedicated to the subjective evaluation of the emotional content of music was one of the partial objectives of this work.

A set of 154 songs from 10 music genres was evaluated by subjects in the listening experiment in terms of mood of music. This amount of data collected within one concise test is also an important contribution of this work, since subjective evaluation is treated as the "**ground truth**" in MER studies. Subjective data were retrieved using the author's model of mood but the model is also compatible with the commonly used 2-dimensional Valence/Arousal representation of emotions. This was shown in the course of the study. A diversified music set enabled also the genre-oriented analysis of the emotional content of music. The correlation analysis of features with Valence and Arousal values, achieved from the subjective tests, returned a vector of parameters strongly related to mood of music. This proves thesis no.1: "**It is possible to find parameters describing a musical excerpt, which are highly correlated with subjective mood labeling results.**"

A visualization tool was created to present the concept of musical excerpts organized according to emotional content. It enables linking points on the VA plane with music files in an easy and intuitive way. Taking into consideration the results of the previous analyses as well as comments from listeners, fuzzification of the mood representation was also introduced along with an approach based on fuzzy logic.

The automatic positioning of musical excerpts onto a mood plane was performed using SOMs, which were tested during the preliminary phase of experiments. A map of musical excerpts organized geometrically according to their emotional content was created with accuracy of 67%. Various setups of SOMs and data pre-processing were tested at that stage.

In addition, classification based on a supervised training, i.e. employing the ANN algorithm was performed. The achieved classification accuracy depends strongly on the number and definition of classes. The best results for classification were achieved when the VA plane was divided into halves, and these reached up to 91% accuracy for Arousal and 83% for Valence. These results outperform classification effectiveness shown in the literature sources. Analyses have shown that classification based on classes defined by values of Valence and Arousal returned higher accuracy, which demonstrated that the dimensional-based approach enables better automatic mood recognition.

The results of both algorithms, SOM and ANN, were analyzed and compared, including a case study of single musical excerpts and music genres. This resulted in the proposal of new (additional) mood descriptors that were included in the final feature vector tested with SOM and ANN. Even though some objects were misclassified by both methods, the accuracy values from both representations is very high, hence this proves thesis no. 2: " **Self-organizing maps (SOMs) or artificial neural networks (ANNs) trained employing designed feature vectors can effectively be applied to the automated indexing of mood of musical excerpts.**"

The results obtained from ANN and SOM classification are especially satisfactory, as a very high consistency of the results between listeners and a very good accuracy of the algorithms is not easy to obtain due to the subjective character of evaluation. It demonstrates a very good effectiveness of the algorithms in cases when even experienced listeners' opinions differ. The study presented by the author was focused mostly on the creation of a mood model that would be easily understood and user-friendly, as well as on the computational methods that are close to a human's reasoning and perception, which also vary a lot in terms of interpretation and decisions, and then the finding of a reasonable relationship linking both approaches together. This was demonstrated by the overall results obtained in the key experiments, therefore this proves thesis no. 3: "**Annotations of mood of music achieved by subjective assessments and classifying based on both supervised and unsupervised learning can be coherent.**". In addition, the outcomes achieved were compared with various studies presented in the literature of the topic, and the results presented in this thesis can be treated as complementary or as outperforming many of them.

# 8 CONCLUSIONS AND FURTHER DIRECTIONS

In this section, **major original contributions** of the work are listed:

- The creation of a dictionary of expressions related to mood of music in Polish;
- Executing a Multidimensional Scaling experiment dedicated to mood of music perceptions results and showing two dimensions underlying mood of music perception;
- Proposing a novel analysis of single instrument tracks vs. mix in terms of emotions included in music;
- Proposing new time-based features describing rhythmic content in frequency ranges;
- Proposing an original model of emotions dedicated to the subjective evaluation of mood of music;
- Performing a subjective evaluation of 154 musical excerpts from 10 music genres within one concise listening test of the emotional content;
- Proposing a visualization tool, which enables the presentation of a music set mapped by mood of music;
- Proposing a graphical representation of a music set (154 excerpts) organized according to the emotional content of music;
- Executing an automatic recognition of mood of music utilizing the ANN algorithm and the feature vector containing the proposed new descriptors and obtaining results for mood classification outperforming findings in the literature sources;
- Achieving an automatic mapping of music based on mood using SOM;
- Showing perspectives to utilize a fuzzy logic-based approach to MER-based studies.

The present dissertation has presented a novel model of emotions dedicated to music evaluation and recommendation. This may establish a foundation for future work aimed at the creation of a complementary system, which would allow automatic music organization by emotional content. Advanced programming tools and interfaces can be employed to achieve a fully functional tool for mood-based music organizing available to listeners. Future work on the approach based on fuzzy logic is also encouraged. It could also be used in the direction of correlating personal preferences and individual perception of mood of music.

# 8 CONCLUSIONS AND FURTHER DIRECTIONS

All results and conclusions achieved in the course of presented study can be important cues for the content recommendation based on mood of music, as emotions are what the music is truly about.

# REFERENCES


**[1]** 8 tracks, http://8tracks.com, access 20.09.2015.

**[2]** Agostini G., Longari M., Pollastri E., Musical instrument timbres classification with spectral features, IEEE Multimedia Signal Processing, 2001.

**[3]** Ahn L, Dabbish L., Labeling Images with a Computer Game, CHI, Vienna, Austria, 2004.

**[4]** Alghoniemy M., Ahmed H., Tewfik M., Rhythm and periodicity detection in polyphonic music, IEEE Third Workshop on Multimedia Signal Processing, 185-190, Denmark, 1999.

**[5]** Allamanche E., Herre J., Helmuth O., Fröba B., Kasten T., Cremer M., Content-based Identification of Audio Material Using MPEG-7 Low Level Description, International Symposium on Music Information Retrieval, Bloomington, NI, USA, 2001.

**[6]** Allen P.E., Dannenberg R.B., Tracking musical beats in Real time, Int. Comp. Music Conf. (ICMC), 140–143, Scotland, 1990.

**[7]** All music, http://www.allmusic.com, access 20.09.2015.

**[8]** Ando Y., Auditory and Visual Sensations, Springer, 2010.

**[9]** AMT Models, http://amtoolbox.sourceforge.net/models.php, access 15.09.2015.

**[10]** Audio Feature Extraction http://ifs.tuwien.ac.at/mir/audiofeatureextraction.html, access 20.09.2015.

**[11]** Auditory Modeling Toolbox, http://amtoolbox.sourceforge.net, access 15.09.2015.

**[12]** Anderson J.A., Rosenfeld E., Neurocomputing - Foundation of Research, MIT Press, Cambridge, Mass, 1988.

**[13]** Apel W., Harvard Dictionary of Music, Cambridge, Massachusetts: Harvard University Press, 1960.

**[14]** Arensburg B., Tillier A. M., Vandermeersch B., Duday H., Schepartz L. A., Rak Y., A Middle Palaeolithic human hyoid bone, Nature 338, 6218, 758–760, 1989.

**[15]** Asmus E.P., Nine affective dimensions, Tech. report, University of Miami, 1986.

**[16]** Notebook for Anna Magdalena Bach, https://en.wikipedia.org/wiki/Notebook_for_Anna_Magdalena_Bach, access 20.08.2015.

**[17]** Bachorik J., Bangert M., Loui P., Larke K., Berger J., Rowe R., Schlaug G., Emotion in Motion: Investigating the Time-Course of Emotional Judgments of Musical Stimuli, Music Perception, 26, 4, 355-364, 2009.

**[18]** Bai Y., Wang D., Fundamentals of Fuzzy Logic Control – Fuzzy Sets, Fuzzy Rules and Defuzzifications, Springer, London, 2006.

**[19]** Barbedo J. G. A., Lopes A., A New Cognitive Model for Objective Assessment of Audio Quality, JAES, 53, 1/2, 22-31, 2005.

**[20]** Barthet M., Fazekas G., Sandler M., Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-based Models, Proc. of CMMR Computer Music Modeling and Retrieval, London, UK, 492-507, 2012.

**[21]** Barras C., Did early humans, or even animals, invent music?, http://www.bbc.com/earth/story/20140907-does-music-pre-date-modern-man, access 12.8.2015

**[22]** Barrington, D., Turnbull, D. O'Malley, and G. Lanckriet, User-Centered Design of a Social Game to Tag Music, ACM KDD Workshop on Human Computation, NY, USA, 2009.

**[23]** Barrington L., Turnbull D., Yazdani M., Lanckriet G., Combining audio content and social context for semantic music discovery, In: Proc. ACM SIGIR, NY, USA, 2009.

**[24]** Bartsch M.A., Wakefield G.H., To catch a chorus: Using chroma-based representations for audio thumbnailing, IEEE Workshop Applications of Signal Processing to Audio and Acoustics, 15–18, 2001.

**[25]** Baume C., Fazekas G., Barthet M., Marston D., Sandler M., Selection of audio features for music emotion recognition using production music, 53rd International Conference: Semantic Audio, UK, London, 2014.

**[26]** Beauchamp J., Perceptually Correlated Parameters of Musical Instrument Tones, Archives of Acoustics, 36, 2, 225–238, 2011.

**[27]** Bech S., Zacharov N., Perceptual Audio Evaluation - Theory, Method and Application, Wiley, 2006.

**[28]** Berenzweig A., Logan B., Ellis D. P. W., Whitman B., A Large-Scale Evaluation of Acoustic and Subjective Music- Similarity Measures, Computer Music Journal, 28, 2, 63–76, 2004.

**[29]** Becker J., Rohlfing C., A Segmental Spectral Flatness Measure for Harmonic-Percussive Discrimination, in Proc. of International Student Conference on Electrical Engineering POSTER '13, Prague, 2013.

**[30]** Bello J.P., Pickens J., A robust mid-level representation for harmonic content in music signals, Proc. Int. Conf. Music Information Retrieval, London, UK, 2005.

**[31]** Międzynarodowy Festiwal Kompozytorów Krakowskich, http://zkp.krakow.pl/27, access 24.09.2015.

**[32]** Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., Dacquet, A.: 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, 2005.

**[33]** Bischoff K., Firan C.S., Paiu R., Nejdl W., Laurier C., Sordo M., Music mood and theme classification - a hybrid approach, ISMIR, 657–662, 2011.

**[34]** Bischoff K., Claudiu S. Firan , Raluca Paiu , Wolfgang Nejdl , Cyril Laurier , Mohamed Sordo. Music Mood and Theme Classification - a Hybrid Approach, ISMIR, 2009.

**[35]** Blacking, J., Music, culture, and experience, University of Chicago Press, 1995.

**[36]** Blewitt W., Exploration of Emotion Modelling through Fuzzy Logic, PhD Thesis, Montfort University, 2012.

**[37]** Blood B., Music Theory Online: Lesson 21: Phrasing & Articulation, Dolmetsch Organisation, http://www.dolmetsch.com/musictheory21.htm, access 12.08.2015.

**[38]** Borg I., Groenen P., Modern Multidimensional Scaling: Theory and Applications, Springer, Germany, 2007.

**[39]** Boulez P., Boulez on Music Today, Harvard University Press, London, Faber, 1971.

**[40]** Bowman W., Philosophical perspectives on music. New York, Oxford University Press, 1998.

**[41]** Bray S., Tzanetakis G., Distributed Audio Feature Extraction for Music, Int. Conf. on Music Information Retrieval., 2005.

**[42]** Bregman A. S., Auditory Scene Analysis: The Perceptual Organization of Sound, Cambridge, MA, MIT Press, 1990.

**[43]** Brinker B., Dinther R., Skowronek J., Expressed music mood classification compared with valence and arousal ratings, EURASIP J. Audio, Speech, and Music Processing, 1, 2012, http://link.springer.com/journal/13636/2012/1/page/1, access 6.10.2015

**[44]** Brixen E. B., Sadolin C., Kjelin, H**.,** Acoustical Characteristics of Vocal Modes in Singing, 134th AES Convention, Paper No. 8897, 2013.

**[45]** Burger B., Thompson M.R., Luck G., Saarikallio S., Toiviainen P., Frontiers in Psychology, 4,183, 2013, http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00183/abstract, access 25.05.2015

**[46]** Burns Edward M., Intervals, Scales, and Tuning, In The Psychology of Music, New York, Academic Press, 1998.

**[47]** Byrd G., A similarity scale for content-based music IR, Available: www.informatics.indiana.edu/donbyrd/MusicSimilarityScale.html, access 07.08.2015.

**[48]** Byrd D., Crawford T., Problems of music information retrieval in the real world, Information Processing and Management, 38, 2, 249- 272, 2002.

**[49]** Caetano M., Wiering F., Theoretical Framework Of A Computational Model Of Auditory Memory For Music Emotion Recognition, ISMIR 2014, 331 - 336, 2014.

**[50]** Casey M.A., Veltkamp R., Goto M., Leman M., Rhodes C., Slaney M., Content-Based Music Information Retrieval: Current Directions and Future Challenges, Proceedings of the IEEE, 96, 4, 668-696, April 2008.

**[51]** Castro S.L., Lima C.F., Age and Musical Expertise Influence Emotion Recognition in Music, Music Perception, 32, 2, 2014.

**[52]** Cemgil A. T., Gürgen F., Classification of Musical Instrument Sounds Using Neural Networks, In. Proc. of SIU97, 1997.

**[53]** Cemgil A., Kappen B., Monte Carlo methods for tempo tracking and rhythm quantization, J. Artificial Intell. Res., 18, 45–81, 2003.

**[54]** Chang W., C., Su A.W.Y., Synthesizing Coupled-String Musical Instruments with a Mulit-Channel Recurrent Network, 116th AES Convention, Paper No. 6042, 2004.

**[55]** Chen Y.-A., Yang Y.-H., Wang J.-C., Chen H., The Amg1608 Dataset For Music Emotion Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

**[56]** Chen, Y.-A., Wang, J.-C., Yang, Y.-H, Chen, H.: Linear regression-based adaptation of music emotion recognition models for personalization, ICASSP IEEE International Conference, 2014.

**[57]** Chithra S., Sinithv M.S., Gayathri A., Music Information Retrieval for Polyphonic Signals Using Hidden Markov Model, Procedia Computer Science, 46, 381–387, Elsevier, 2015.

**[58]** Clayton M., Herbert T., Middleton R., The cultural study of music: A critical introduction, New York, Routledge, 2003.

**[59]** Covaciu-Pogorilowski A., Musical Time Theory & A Manifesto, http://www.zeuxilogy.home.ro/media/manifesto.pdf, access 12.08.2015.

**[60]** Cremers H. R., Demenescu L. R., Aleman A., Renken R., Van Tol M., Van Der Wee N., Neuroticism modulates amygdala-prefrontal connectivity in response to negative emotional facial expressions, NeuroImage, 49, 963-970, 2010.

**[61]** Cruse H., Neural Networks as Cybernetic Systems, Thieme, Stuttgart, 1996.

**[62]** Max 7, https://cycling74.com, access 24.09.2015.

**[63]** Czyzewski, A., Krolikowski, R., Kostek, B.: Encoding Spatial Information for Advanced Teleconferencing, 19th AES International Conference: Surround Sound - Techniques, Technology, and Perception, Paper No. 1890, 2001.

**[64]** Dalla Bella S., Peretz I., Rousseau L., Gosselin N., A developmental study of the affective value of tempo and mode in music, Cognition, 80, B1–B10, 2001.

**[65]** Dang T.T., Shirai K., Machine learning approaches for mood classification of songs toward music search engine, In: Proc. ICKSE, 2009.

**[66]** Davies S., Allen P., Mann M., Cox, T., Musical Moods: A Mass Participation Experiment for Affective Classification of Music, *12th International Society for Music Information Retrieval Conference*, USA, pp. 741--746, 2011.

**[67]** Decibel, https://decibel.net, access 24.09.2015.

**[68]** Denisch, B.: Music Theory Handbook: Getting started with counterpoint, Berklee, avaiable at https://welcome.online.berklee.edu/music-theory-handbook.html, Access 12.08.2015.

**[69]** Deutsch D., The Psychology of Hearing, http://deutsch.ucsd.edu/pdf/SVC_1998_Sept.pdf, access 09.08.2015.

**[70]** Dowling W. J., Harwood D. L., Music cognition. New York, Academic Press, 1986.

**[71]** Downie J. S., The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research, Acoustical Science and Technology, 29, 4, 247–255, 2008.

**[72]** DuBois L., ART && CODE SYMPOSIUM: Max/MSP/Jitter, https://vimeo.com/5480466, access 24.09.2015.

**[73]** Duong N., Duong H-T., A Review of Audio Features and Statistical Models Exploited for Voice Pattern Design, PATTERNS, 2015.

**[74]** Eerola T., Toiviainen P., MIR IN MATLAB: THE MIDI TOOLBOX, 5th International Conference on Music Information Retrieval, 2004.

**[75]** Eerola T., Lartillot O., Toiviainen P., Prediction of multidimensional emotional ratings in music from audio using multivariate regression models, Int. Society for Music Information Conf., Japan, 2009.

**[76]** Ellis D., Poliner G., Identifying cover songs with chroma features and dynamic programming beat tracking, Proc. Int. Conf. Acoustic, Speech and Signal Processing, Honolulu, 2007.

**[77]** Elman J., Finding Structure in Time, Cognitive Science 14, 179-211, 1990.

**[78]** Elman J., Learning and development in neural networks: the importance of starting small, Cognition, 48, 71-99, 1993.

**[79]** Eronen A., Klapuri A., Musical instrument recognition using cepstral coefficients and temporal features, Proceedings of the Acoustics, Speech, and Signal Processing, IEEE International Conference, 02, II753-II756, 2000.

**[80]** ESP Game http://en.wikipedia.org/wiki/ESP_game, 24.09.2015.

**[81]** Ellis D., Whitman B., Berenzweig A., Lawrence S., The quest for ground truth in musical artist similarity, 3rd International Conference on Music Information Retrieval, 170–177, 2002.

**[82]** Etzkorn B., Data Normalization and Standardization, http://www.benetzkorn.com/wp-content/uploads/2011/11/Data-Normalization-and-Standardization.pdf, access 29.07.2015.

**[83]** Evans M., Interactive Beat Tracking for Assisted Annotation of Percussive Music, 123 AES Convention, Paper No. 7247, 2007.

**[84]** Everitt B. S., Skrondal A., The Cambridge Dictionary of Statistics, Cambridge University Press, 2010.

**[85]** Fahlman S., An empirical study of learning speed in back- propagation networks, 1988.

**[86]** Farnsworth P. R., A study of the Hevner adjective list, J. Aesthetics and Art Criticism, 13, 97-103, 1954.

**[87]** Feiten, B., Günzel, S.: A Sound-Retrieval Index Based on Two-Dimensional Similarity Maps, AES Paper No. 3542, Presented at the 94th AES Convention, 1994.

**[88]** Feng Y., Zhuang Y., Pan Y., Popular music retrieval by detecting mood, ACM SIGIR, 375–376, 2003.

**[89]** Frühwirth M., Rauber A., Self-Organizing Maps for Content-Based Music Clustering, 12th Italian Workshop on Neural Nets, 2001.

**[90]** Foote J., Uchihashi S., The beat spectrum: A new approach to rhythm analysis, IEEE Int. Conf. on Multimedia and Expo, 2001.

**[91]** Fornari J., Eerola T., The pursuit of happiness in music: retrieving valence with contextual music descriptors, Proc. CCMR 2008, 119–133, 2008.

**[92]** Friberg A., Hedblad A., A comparison of perceptual ratings and computed audio features, 8th Sound and Music Computing Conference, 122–127, 2011.

**[93]** Galiana M., Llinaresa C., Pageb Á., Subjective evaluation of music hall acoustics: Response of expert and non-expert users, Building and Environment, 58, 1–13, 2012.

**[94]** Gemisatos, http://www.usc.edu/dept/polish_music/news/july09.html, access 24.09.2015.

**[95]** Gibson E.J., Improvements in perceptual judgments as a function of controlled practice or training, Psychological Bulletin, 50, 401–431, 1953.

**[96]** Grossman L., If You Liked This…, Time, June 14 – June 21, 50-54, 2010.

**[97]** Gutheil E. A., Music and Your Emotions, Liveright, 1970.

**[98]** Gracenote, https://gracenote. com, access 24.09.2015.

**[99]** Goto M., An audio-based real-time beat tracking system for music with or without drum-sounds, J. New Music Res., 30, 2, 159–171, 2001.

**[100]** Gray A.H., Markel J.D. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis, IEEE Trans. Acoust. Speech Signal Process., 22, 207 -217, 1974.

**[101]** Gunawan D., Sen S., Separation of Harmonic Musical Instrument Notes Using Spectro-Temporal Modeling of Harmonic Magnitudes and Spectrogram Inversion with Phase Optimization, J. AES,1004-1014, 2012.

**[102]** Hallam S., Cross I., Thaut M., The Oxford handbook of music psychology, 45–118, New York, Oxford University Press, 2001.

**[103]** Hamann S., Canli T., Individual differences in emotion processing, Current Opinian in Neurobiology, 14, 233-238, 2004.

**[104]** Han B., Rho S., Dannenberg R., Hwang E., SMERS: Music Emotion Recognition Using Support Vector Regression, 10th International Society for Music Information Retrieval Conference, 651–656, 2009.

**[105]** Helen M., Virtanen T., Separation of Drums From Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine, 13th European Signal Processing Conference, Antalaya, Turkey, 2005.

**[106]** Heroku, https://www.heroku.com, access 24.09.2015.

**[107]** Herz J., Krogh A., Palmer R.G.: Wstęp do teorii obliczeń neuronowych (In Polish), Wydawnictwa Naukowo-Techniczne, Warszawa, 1995.

**[108]** Hevner K., The affective value of pitch and tempo in music, American Journal of Psychology, 49, 621-630, 1937.

**[109]** Hirvonen T., **Classification of Spatial Audio Location and Content Using Convolutional Neural Networks,** 138th AES Convention, AES Paper No. 9294, 2015.

**[110]** Hoffmann P., Kaczmarek A.,Spaleniak P. , Kostek B., Music Recommendation System, Journal of Telecommunications and Information Technology, 2, 59-69, 2014.

**[111]** Honing H., Structure and Interpretation of Rhythm and Timing, Psychology of Music, Academic Press, 369-404, 2002.

**[112]** Howard D., Angus J., Acoustics and psychoacoustics, Oxfordm, Focal Press, 2006.

**[113]** Hu X., Downie J., Exploring mood metadata: relationships with genre, artist and usage metadata, 8th International Conference on Music Information Retrieval, 67-72, 2007.

**[114]** Hu X., Downie S. J., Laurier C., Bay M., Ehmann A. F., The 2007 MIREX audio mood classification task: Lessons learned, ISMIR, 462–467, 2008.

**[115]** Hu X., Lee J., A Cross-cultural Study of Music Mood Perception between American and Chinese Listeners, *13th International Society for Music Information Retrieval Conference*, 2012.

**[116]** Hu X., Yu B., Exploring the Relationship Between Mood and Creativity in Rock Lyrics, 12th International Society for Music Information Retrieval Conference, 789-794, 2011.

**[117]** Huron D., Sweet anticipation: Music and the psychology of exectation, Cambridge, MA: MIT Press, 2006.

**[118]** Husain G., Thompson W. F., Schellenberg E. G., Effects of musical tempo and mode on arousal, mood, and spatial abilities, Music Perception, 20, 151–171, 2002.

**[119]** In The Mood Music, http://inthemoodmusic.com, access 24.09.2015.

**[120]** Irtel H, PXLab Self-Assessment-Manikin Scales, http://www.webcitation.org/69cFmMce1, access 13.08.2015.

**[121]** iTunes, https://itunes.apple.com, access 8.10.2015.

**[122]** ITU-R BS.1534.2 Recommendation, Method for the subjective assessment of intermediate quality levels of coding systems, 2014.

**[123]** Jolliffe I.T., Principal Component Analysis, Springer Series in Statistics, Springer, NY, 2002.

**[124]** Jones R., Fay R.R, Popper A., Music Perception, Springer-Verlag, New York, 2010.

**[125]** Jun S., Rho S., Han B., Hwang E., A Fuzzy Inference-based Music Emotion Recognition System, Visual Information Engineering, 2008.

**[126]** Juslin P. N., From mimesis to catharsis: expression, perception, and induction of emotion in music, Musical communication, pp. 85-115, Oxford University Press, 2005.

**[127]** Juslin P., Sloboda J. A., Handbook of music and emotion: Theory, research, applications, New York, Oxford University Press, 2011.

**[128]** Justus T., Bharucha J., Music perception and cognition, Handbook of experimental psychology,Sensation and perception, 453–492, New York, Wiley & Sons, 2002.

**[129]** Kaminsky I., Czaszejko T., Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using *k* NNC, Journal of Intelligent Information Systems, 24, 2-3, 199-221, 2005.

**[130]** Kennedy M., Portato, The Oxford Dictionary of Music, 1994.

**[131]** Kim J.H., Lee S., Kim S.M., Yoo W.Y., Music mood classification model based on Arousal-Valence values, ICACT, 292–295, 2011.

**[132]** Kim H-G., Moreau N., Sikora T., MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval, Wiley, 2005.

**[133]** Kim Y., Schmidt E., Emelle L., MOODSWINGS: A Collaborative Game For Music Mood Label Collection, ISMIR, 2008.

**[134]** Kim Y.E., Schmidt E.M., Migneco R., Morton B.G., Richardson P., Scott J., Speck J.A., Turnbull D., Music Emotion Recognition: A State of the Art Review, 11th International Society for Music Information Retrieval Conference, ISMIR 2010, 255-266, 2010.

**[135]** Kirn P., Jamie Lidell, Artists Talk Max Inspiration, Write Musical Odes to Max, http://createdigitalmusic.com/2006/07/jamie-lidell-on-maxmsp-artists-talk-max-inspiration-write-musical-odes-to-max, access 24.09.2015.

**[136]** Kleczkowski P., Perception of Mixture of Musical Instruments with Spectral Overlap Removed, Archives of Acoustics, 37, 3, 355–363, 2012.

**[137]** Kleczkowski P., Król A., Małecki P., Multichannel Sound Reproduction Quality Improves with Angular Separation of Direct and Reflected Sounds, J. AES, 63, 6, 427-442, June 2015.

**[138]** Kleczkowski P., Plewa M., Pyda G., *Localization of a sound source in Double MS recordings*, Archives of Acoustics, 33, 4 (Suppl.), 147–152, 2008.

**[139]** Kleczkowski P., Plewa M., Pluta M., Masking a frequency band in a musical fragment played by a single instrument, Acta Physica Polonicam, 119, 6–A, 991–995, 2011.

**[140]** Kleczkowski P., Plewa M., Pluta M., Can selective mixing enhance multi-instrumental music?, ICSV'18, Brasil, 2011.

**[141]** Klickstein G., The Musician's Way: A Guide to Practice, Performance, and Wellness, Oxford University Press , 2009.

**[142]** Knees, P. Schedl, M., Music Retrieval and Recommendation: A Tutorial Overview, 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1133-1136, 2015.

**[143]** Koch K., Pauly K., Kellermann T., Seiferth R., Reske M., Backes V., Gender differences in the cognitive control of emotion: An fMRI study, Neuropsychologia, 45, 2744-2754, 2007.

**[144]** Kohonen T., Honkela T., Kohonen Network, *Scholarpedia*, 2007.

**[145]** Korbicz J., Obuchowicz A., Uciński D.: Sztuczne sieci neuronowe - Podstawy i zastosowania (In Polish), Akademicka Oficyna Wydawnicza PLJ, Warszawa, 1994.

**[146]** Kosta K., Song Y., Fazekas G., Sandler, M., A study of cultural dependence of perceived mood in greek music. *14th International Society for Music Information Retrieval Conference*, 2013.

**[147]** Kostek B., Application of Learning Algorithms to Musical Sound Analysis, 97th AES Convention, Paper No: 3873, 1994.

**[148]** Kostek B., Musical Data Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques, Proceedings of the IEEE 92, 4, 712-729, 2004.

**[149]** Kostek B., Perception-Based Data Processing in Acoustics, Springer-Verlag GmbH, 2005.

**[150]** Kostek B., Perception-Based Data Processing in Acoustics: Applications to Music Information Retrieval and Psychophysiology of Hearing, Springer, 2010.

**[151]** Kostek B., Content-Based Approach to Automatic Recommendation of Music, 131st AES Convention, Paper No: 8506, New York, 2011.

**[152]** Kostek B., Czyzewski A., Krolikowski R., Neural Networks Applied to Sound Localization Detection, 110th AES. Convention, Paper No: 5375, 2011.

**[153]** Kostek B., Hoffmann P., Spaleniak P., Kaczmarek A., Creating a Realible Music Discovery and Recomendation System, Intelligent Tools for Building a Scientific Information Platform, Springer Verlag, 2014.

**[154]** Kostek B., Królikowski R., Application of artificial neural networks to the recognition of musical sounds, Archives of Acoustics, **22**, 1, 27-50, 1997.

**[155]** Kostek B., Kupryjanow A., Żwan P., Jiang W., Ras Z., Wojnarski M., Swietlicka J., Report of the ISMIS 2011 Contest: Music Information Retrieval, Foundations of Intelligent Systems, ISMIS 2011.

**[156]** Kostek B., Plewa M., Parametrization and correlation analysis applied to music mood classification, International J. of Computational Intelligence Studies, 2, 1, 4-25, 2013.

**[157]** Kostek B., Szczuko P., Żwan P., Dalka P., Processing of Musical Data Employing Rough Sets and Artificial Neural Networks, Transactions on Rough Sets, III, 112-133, Springer, 2005.

**[158]** Kostek B., Wójcik J., Forming and Ranking Musical Rhythm Hypotheses, LNAI 3213, Springer Verlag, Berlin, Heidelberg, 750 - 756, 2004.

**[159]** Kulkarni A., Iyer D., Sridharan S.R., Audio Segmentation, In IEEE International Conference on Data Mining, ICDM, 2001.

**[160]** Lamere P.: Social tagging and music information retrieval, J. New Music Research, 37, 2, 101-114, 2008.

**[161]** Landman Y., Explanation of the origin of musical scales clarified by a string division method, furious.com, access 10.06.2014.

**[162]** Last.fm, http://www.last.fm, access 10.06.2014.

**[163]** Laroche J., Estimating tempo, swing and beat locations in audio recordings, IEEE Workshop on Application of Signal Proc. to Audio and Acoust., 135–138, New Paltz, 2001.

**[164]** Lartillot O., MIRtoolbox 1.4: User's Manual, Finnish Centre of Excellence in Interdisciplinary Music Research Swiss Center for Affective Sciences, 2012.

**[165]** Lartillot O., Eerola T., Toiviainen P., Fornari J., Multi-feature modeling of pulse clarity: Design, validation, and optimization, International Conference on Music Information Retrieval , 2008.

**[166]** Lartillot O., Toiviainen P., A Matlab Toolbox for Musical Feature Extraction From Audio, International Conference on Digital Audio Effects, 2007.

**[167]** Laurier C., Lartillot O., Eerola T., Toiviainen P., Exploring Relationships between Audio Features and Emotion in Music, Conference of European Society for the Cognitive Sciences of Music, 2009.

**[168]** Laurier C., Sordo M., Serra J., Herrera P., Music Mood Representations from Social Tags, Proceedings of the 10th International Society for Music Information Conference, Kobe, Japan, 381-386, 2009.

**[169]** [Lee 2015] Lee J. H., Cho H., Kim J.-S., Users' music information needs and behaviors: Design implications for music information retrieval systems, Journal of the Association for Information Science and Technology, 2015.

**[170]** Lee S., Kim J.H., Kim S.M., Yoo W.Y., Smoodi: Mood-based music recommendation player. In: Proc. IEEE ICME, 1–4, 2011.

**[171]** Lee K., Slaney M., Automatic Chord Recognition from Audio Using an HMM with Supervised Learning, AMCMM, New York, 2006.

**[172]** Leimeister M., Gaertner, D., Dittmar C., Rhythmic Classification of Electronic Dance Music, 53rd AES International Conference: Semantic Audio, Paper No. 8607, 2014.

**[173]** Leman M., Vermeulen V., De Voogdt L., Moelants D., Lesaffre M., Prediction of Musical Affect Using a Combination of Acoustic Structural Cues, J. of New Music Research, 34, 1, 39-67, 2005.

**[174]** Levitin D. J., This Is Your Brain on Music: The Science of a Human Obsession, London, Grove/Atlantic, 2008.

**[175]** Li T., Ogihara M., Shao B. Wang D., Machine Learning Approaches for Music Information Retrieval, Theory and Novel Applications of Machine Learning, http://www.intechopen.com/books/theory_and_novel_applications_of_machine_learning/machine_learning_approaches_for_music_information_retrieval, access 11.09.2014

**[176]** Liang S. F., Su A. W. Y., Recurrent Neural-Network-Based Physical Model for the Chin and Other Plucked-String Instruments, JAES, 48., 11,1045-1059, 2000.

**[177]** Lidy T., Rauber A., Evaluation of Feature Extractors and Psychoacoustic Transformations for Music Genre Classification, ISMIR 2005, 34-41, London, UK, 2005.

**[178]** Lidy T., Rauber A., Computing Statistical Spectrum Descriptors for Audio Music Similarity and Retrieval, Music Information Retrieval Evaluation eXchange (MIREX), *2006*.

**[179]** Liebetrau J., Nowak J., Sporer T., Krause M., Rekitt M., Schneider S., Paired Comparison as a Method for Measuring Emotions, AES Convention, Paper No. 9016, 2013.

**[180]** Liebetrau J., Schneider S., Music and Emotions: A Comparison of Measurement Methods, 134th AES Convention, Paper No. 8875, 2010.

**[181]** Lima M.F.M., Machado J.A.T., Costa A.C., A Multidimensional Scaling Analysis of Musical Sounds Based on Pseudo Phase Plane, Appl. Anal., 2012, Special Issue, 2012.

**[182]** Lin Y., Yang Y., Chen H. H., Liao I., Ho Y., Exploiting Genre for Music Emotion, in Proc. ICME, 618-621, 2009.

**[183]** Lipkind G., http://www.lipkind.info/results/s-single-voice-polyphony/s01, access 10.08.2015.

**[184]** Liu H., Yu L., Toward Integrating Feature Selection Algorithms for Classification and Clustering, In IEEE Transactions on Knowledge and Data Engineering, 17, 491–502, 2005.

**[185]** Logan B., Mel Frequency Cepstral Coefficients for Music Modeling, ISMIR, 2000.

**[186]** Logic, SoundOnSound: Apple Logic Pro 9, http://www.soundonsound.com/sos/oct09/articles/logic9.html, access 07.08.2015.

**[187]** Lohri A., Carral S., Chatziioannou V., Combination Tones in Violins, Archives of Acoustics, 36, 4, 727–740, 2012.

**[188]** Lu L., Liu D., Zhang H.-J., Automatic Mood Detection and Tracking of Music Audio Signals, IEEE Transactions on Audio, Speech and Language Processing, 14, 5-18, 2006.

**[189]** Madhu N., Note on measures for spectral flatness. In Electronics Letters, 45, 23, pp. 1195 - 1196, 2009.

**[190]** MajorMiner, http://majorminer.org/info/intro, access 03.09.2015.

**[191]** Malecki P., Evaluation of objective and subjective factors of highly reverberant acoustic field (in Polish), Doctoral Thesis, AGH University of Science and Technology, Krakow, 2013.

**[192]** Mandel M., Ellis D., A Web-Based Game for Collecting Music Metadata, 8th International Conference on Music Information Retrieval, 365-366, 2007.

**[193]** Mann M., Cox T.J., Li F.F., Music mood classification of television theme tunes, ISMIR, 735–740, 2011.

**[194]** Markov K., Matsui T., Dynamic Music Emotion Recognition Using State-Space Models, MediaEval Workshop, 2014.

**[195]** Mathworks, http://www.mathworks.com, access 07.08.2015.

**[196]** MATLAB, Neural Network Toolbox, available at http://www.mathworks.com/help/pdfdoc/nnet/nnetug.pdf, access 05.12.2013.

**[197]** Mayer R., Merkl D., Rauber A., Mnemonic SOMs: Recognizable Shapes for Self-Organizing Maps, Proceedings of the Fifth International Workshop on Self-Organizing Maps, 2005.

**[198]** Mazur Z., Wiklak K., Modification of Page Rank Algorithm for Music Information Retrieval Systems, New Research in Multimedia and Internet Systems, IV, 227-237, Springer, 2015.

**[199]** Mermelstein P., Distance measures for speech recognition, psychological and instrumental, Pattern Recognition and Artificial Intelligence, 374–388, Academic, New York, 1976.

**[200]** Miell D., MacDonald R., Hargreaves D., Musical Communication, Oxford University Press, 2005.

**[201]** Mika D., Kleczkowski P., ICA-based Single Channel Audio Separation: New Bases and Measures of Distance, Archives of Acoustics, 36, 2, 311–331, 2011.

**[202]** Miller F., Stiksel M., Jones R., Last.fm in numbers, Last.fm press material, 2008.

**[203]** Wu M.-J., Jang J.-S. R., Confidence-based late Fusion for Music Genre Classification, MIREX, 2014.

**[204]** Mion L., Application of Bayesian networks to automatic recognition of expressive content of piano improvizations, Stockholm Music Acoustics Conf., 2, 557–560, 2003.

**[205]** MIR Toolbox, http://www.cc.jyu.fi/~lartillo/mirtoolbox, access 05.10.2014.

**[206]** Moelants D., McKinney M., Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous? 8th Conference on Music Perception and Cognition, 2004.

**[207]** Molau S., Pitz M., Schluter R., Ney H., Computing Mel-frequency cepstral coefficients on the power spectrum, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, 2001.

**[208]** Moodswings, http://music.ece.drexel.edu/research/emotion/moodswings, access 05.10.2014.

**[209]** Moore B. C. J., An Introduction to the Psychology of Hearing, New York, Academic Press, 1997.

**[210]** Moore J., Tian L., Lai C., Word-Level Emotion Recognition Using High-Level Features, CICLing 2014, 17–31, Springer-Verlag, Berlin Heidelberg, 2014.

**[211]** [Morando et al.]Morando M., Museli M., Guarino M., Musical rhythm recognition with neural networks, Expert Systensm abd Beural Networks, 229-232, 1996.

**[212]** Moreau A., Flexer A., Drum transcription in polyphonic music using non-negative matrix factorisation. 8th International Conference on Music Information Retrieval (ISMIR), 2007.

**[213]** Morton B. G., Speck J. A., Schmidt E. M., Kim Y. E., Improving music emotion labeling using human computation. *Workshop on Human Computation*, 2010.

**[214]** Mourjopoulos J., Tsoukalas D., Neural Network Mapping to Subjective Spectra of Music Sounds, JAES, 40, 4, 253-259, 1992.

**[215]** MPEG-7, Information Technology — Multimedia Content Description Interface — Part 4: Audio, ISO/IEC JTC 1/SC 29, 2001.

**[216]** MPEG-7 Audio, http://mpeg.chiariglione.org/standards/mpeg-7/audio, access 01.08.2015.

**[217]** Mufin, http://www.mufin.com/us, access 01.08.2015.

**[218]** MusicBrainz, http://musicbrainz.org, access 05.12.2014.

**[219]** MusicID2, http://www.musicid2.com, access 05.12.2014.

**[220]** Musicovery, http://musicovery.com, access 04.09.2015.

**[221]** Müller M., Information Retrieval for Music and Motion. Springer. 65-75, 2001.

**[222]** Müller M., Grosche P., Tempo and Beat Analysis, Music Processing course materials, http://resources.mpi-inf.mpg.de/departments/d4/teaching/ss2010/mp_mm/2010_MuellerGrosche_Lecture_MusicProcessing_BeatTracking_handout.pdf, access 05.10.2015.

**[223]** Myint E.E.P., Pwint M., An approach for multi-label music mood classification, ICSPS., VI, pp. 290–294, 2010.

**[224]** Napiorkowski S., Music mood recognition: State of the Art Review, University of Aachen report, http://hpac.rwth-aachen.de/teaching/sem-mus-15/reports/Napiorkowski.pdf, access 04.09.2015.

**[225]** Nikunen J., Virtanen T., Vilermo M., Multichannel Audio Upmixing by Time-Frequency Filtering Using Non-Negative Tensor Factorization, JAES, 794-806, 2012.

**[226]** Novello A., McKinney, M.M.F., Kohlrauschab A., Perceptual Evaluation of Inter-song Similarity in Western Popular Music, J. New Music Research, 40, 1, 1-26, 2011.

**[227]** Ntalampiras S., Audio Pattern Recognition of Baby Crying Sound Events, JAES, 63, 5, 358-369, 2015.

**[228]** Osendorfer C., Schlüter K., Schmidhuber J., Van der Smagt P., Unsupervised learning of low-level audio features for music similarity estimation, 28th International Conference on Machine Learning, 2011.

**[229]** O'Shaughnessy D., Speech communication: human and machine, Addison-Wesley, 1987.

**[230]** Osmalskyj J., Van Droogenbroeck M., Embrechts J.-J., Performances of low-level audio classifiers for large-scale music similarity, International Conference on Systems, Signals and Image Processing, 2014.

**[231]** Palomäki K., Pulkki V., Karjalainen M., Neural Network Approach to Analyze Spatial Sound, AES 16th International Conference on Spatial Sound Reproduction, Paper No. 16-022, 1999.

**[232]** Pampalk E., Islands of Music, Analysis, Organization, and Visualization of Music Archives, Master Thesis, Vienna Technical University, 2001.

**[233]** Pampalk E., Rauber A., Merkl D., Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Map, Proceedings of the International Conference on Artificial Neural Network, 871-876, 2002.

**[234]** Panda R., Paiva R.P., Using Support Vector Machines for Automatic Mood Tracking in Audio Music, 130th AES Convention, Paper No. 8378, 13–16, 2011.

**[235]** Panda R., Rocha B., Paiva R. P., Music Emotion Recognition with Standard and Melodic Audio Features, Publishing models and article dates explained, 313-334, 2015.

**[236]** Pandora, http://www.pandora.com/corporate, access 09.10.2014.

**[237]** Park S.-H., Ihm S.-Y., Jang W.-I., Z Nasridinov A., Park Y.-H., A Music Recommendation Method with Emotion Recognition Using Ranked Attributes, Computer Science and its Applications, 1065-1070, 2015.

**[238]** Patel A. D., Music, language, and brain, Oxford, Oxford University Press, 2008.

**[239]** Paulus J, Virtanen T., Drum transcription with non-negative spectrogram factorization, Proceedings of 13th European Signal Processing Conference (EUSIPCO), 1059-1062, 2005.

**[240]** Peretz I., Zatorre R.J., The Cognitive Neuroscience of Music, Oxford University Press, 2003.

**[241]** Pistone D., Tempo, Dictionnaire de la Musique. Science de la Musique. Formes, Technique, Instruments, Paris, 1977.

**[242]** Plewa M., Kostek B., Creating Mood Dictionary Associated with Music, 132 AES Conventionm, Paper No. 8607, 2012.

**[243]** Plewa M., Kostek B., Multidimensional Scaling Analysis Applied to Music Mood Recognition, 134th AES Convention, Paper No. 8876, 2013.

**[244]** Plewa M., Kostek B., Plewa, M., Kostek, B.: Music Mood Visualization Using Self-Organizing Maps, Archives of Acoustics, 50, 4, 2015.

**[245]** Plewa M., Kleczkowski P., Choosing and configuring a stereo microphone technique based on localisation curves, Archives of Acoustics, 36, 2, 347–363, 2011.

**[246]** Plewa M., Piotrowski S., Practical application of theoretical model to forecasting localization curves in percussion recordings, ISSET'2009, 124–129, 2009.

**[247]** Plewa, M., Pyda, G., Localization Curves in Stereo Microphone Techniques-Comparison of Calculations and Listening Test Results, 128th AES Convention, Paper No. 7978, 2010.

**[248]** Pluta M., Badanie interakcji słuchacza ze sprzętowo-programowym systemem stymulowania błędów intonacyjnych (in Polish), Doctoral Thesis, AGH University of Science and Technology, Krakow, 2008.

**[249]** Plutchik's Color Wheel Of Emotion http://www.writeincolor.com/2011/03/26/plutchiks-color-wheel-of-emotion, access 26.09.2015.

**[250]** Pouyanfar S., Sameti H., Music emotion recognition using two level classification, Intelligent Systems (ICIS), 1-6, 2014.

**[251]** Preisach H., Burkhardt L., Schmidt-Thieme R., Decker Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2008.

**[252]** Pruett J. W., Slavens T. P., Research guide to musicology, Chicago, American Library Association, 1985.

**[253]** Rabiner L., Juang B., Fundamentals of Speech Recognition, Prentice-Hall, 1993.

**[254]** Rauber A., Frühwirth M., Automatically Analyzing and Organizing Music Archives, 5. European Conference on Research and Advanced Technology for Digital Libraries, Springer, 2001.

**[255]** Rauber A., Pampalk M., Merkl D. Content-based Music Indexing and Organization, Proc. 25 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 409-410, 2002.

**[256]** Rauber A., Pampalk M., Merkl D. Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structuring of Music by Musical Styles, 3rd International Symposium on Music Information Retrieval (ISMIR 2002), 71-80, 2002.

**[257]** Rojas R., Neural Networks: A Systematic Introduction, Berlin, Springer, 1996.

**[258]** D. F. Rosenthal, Machine Rhythm: Computer Emulation of Human Rhythm Perception, Ph.D. thesis, Massachusetts Institute of Tech., 1992.

**[259]** Rosner A, Weninger F, Schuller B, Michalak M, Kostek B. Influence of Low-Level Features Extracted from Rhythmic and Harmonic Sections on Music Genre Classification. International Conference on Man-Machine Interactions, 467-473, 2013.

**[260]** Recommender systems, https://en.wikipedia.org/wiki/Recommender_system, access 04.09.2015.

**[261]** Riedmiller M., Braun, H., A direct adaptive method for faster backpropagation learning: the RPROP algorithm, IEEE International Conference on Neural Networks, 1993.

**[262]** Riemann H., Musikalische Dynamik und Agogik, https://archive.org/stream/ldpd_6769763_000/ldpd_6769763_000_djvu.txt, access 11.08.2015.

**[263]** Rumsey F., Loudness revisited, JAES, 62, 12, 906-910, 2014.

**[264]** Russel, J. A., A circumplex model of affects, Journal of personality and Social Psychology, 39, 1161-1178, 1980.

**[265]** Saari P., Eerola T., Fazekas G., Barthet M., Lartillot O., Sandler, M., The role of audio and tags in music mood prediction: a study using semantic layer projection, *14th International Society for Music Information Retrieval Conference*, 2013.

**[266]** Sacks O., Musicophilia: Tales of Music and the Brain, Vintage, 2008.

**[267]** Salembier P., Avaro O., MPEG-7: Multimedia Content Description interface, http://gps-tsc.upc.es/imatge/_Philippe/demo/MPEG21_MPEG7.pdf, access 01.08.2015.

**[268]** [Sandhan]Sandhan T., Sonowal S., Choi, J-Y., Audio Bank: A high-level acoustic signal representation for audio event recognition, Control, Automation and Systems (ICCAS), 2014.

**[269]** Sanden C., Zhang J.: An empirical study of multi-label classifiers for music tag annotation, ISMIR, 717–722, 2011.

**[270]** Sarroff A., Casey M., Groove Kernels as Rhythmic-Acoustic Motif Descriptors, 14th International Society for Music Information Retrieval Conference (ISMIR), 2013.

**[271]** Schedl M., Kepler J., Music Information Retrieval: Recent Developments and Applications, now publishers, Boston, 2014.

**[272]** Scheirer E.D., Tempo and beat analysis of acoustic musical signals, JAES, 103, 1, 588–601, 1998.

**[273]** Schmidt E. M., Kim Y. E., Projection of acoustic features to continuous valence-arousal mood labels via regression. International Society for Music Information Retrieval Conference (ISMIR), 2009.

**[274]** Schmidt E. M., Kim, Y. E, Prediction of time-varying musical mood distributions using Kalman filtering, IEEE International Conference on Machine Learning and Applications, 2010.

**[275]** Schmidt E. M., Turnbull, D. Kim, Y. E., Feature selection for content-based, time-varying musical emotion regression. ACM SIGMM International Conference on Multimedia Information Retrieval, 2010.

**[276]** Scholes P., "Metre" and "Rhythm", The Oxford Companion to Music, London and New York, Oxford University Press, 1977.

**[277]** Scheirer E. D, Tempo and beat analysis of acoustic musical signals, JAES, 103, 1, 588–601, 1998.

**[278]** Schellenberg E. G., Krysciak A. M., Campbell, R. J., Perceiving emotion in melody: Interactive effects of pitch and rhythm. Music Perception, 18, 155–171, 2000.

**[279]** Scherer, K. R., Oshinsky J. S., Cue utilization in emotion attribution from auditory stimuli, Motivation and Emotion, 1, 331–346, 1977.

**[280]** Schubert E., Update of the Hevner adjective checklist, Perceptual and Motor Skills, 96, 1117-1122, 2003.

**[281]** Schuller B., Dorfner J., Rigoll G., Determination of nonprototypical valence and arousal in popular music: Features and performances, EURASIP Journal on Audio, Speech, and Music Processing, 2010, 1–20, 2010.

**[282]** Schuller B., Lehmann A., Weninger F., Eyben F., Rigoll G., Blind Enhancement of the Rhythmic and Harmonic Sections by NMF: Does it help? International Conference on Acoustics including the 35th German Annual Conference on Acoustics, 2009.

**[283]** Serra J., Gomez A., A cover song identification system based on sequences of tonal descriptors, Int. Conf. Music Information Retrieval, 2007.

**[284]** [Sethares]Sethares W., Local Consonance and the Relationship between Timbre and Scale, Journal of the Acoustical Society of America, 94, 1, 1993.

**[285]** [Seung] Baek S.-E., Kim M. Y., Music Genre Classification MIREX 2014 submissions, 2014.

**[286]** [Sheh] Sheh A., Ellis D.P., BChord segmentation and recognition using EM-trained hidden Markov models, Int. Conf. Music Information Retrieval, 2003.

**[287]** [Skowronek 2006]Skowronek J., McKinney M., Features for audio classification: Percussiveness of sounds, Intelligent Algorithms in Ambient and Biomedical Computing, 7, 119–132, Springer, Dordrecht, 2006.

**[288]** Skowronek J., Par S., McKinney M.: Groundtruth for automatic music mood classification, 7th International Conference on Music Information Retrieval, 2006.

**[289]** Slacker, http://www.slacker.com, access 10.10.2014.

**[290]** Sloboda J. A., Music structure and emotional response: Some empirical findings, Psychology of Music, 19, 110-120, 1991.

**[291]** Sloboda J., Exploring the musical mind, New York, Oxford University Press, 2005.

**[292]** Smaragdis P., Brown J.C., Non-negative matrix factorization for polyphonic music transcription, WASPAA, 177-180, 2003.

**[293]** Smith L., A tutorial on Principal Components Analysis, http://www.cs.otago.ac.nz/cosc453/studenttutorials/principalcomponents.pdf, access 5.12.2013.

**[294]** Sofianos S., Ariyaeeinia A., Polfreman R., Sotudeh R., H-Semantics: a Hybrid Approach to Singing Voice Separation. JAES, 831-841, 2012.

**[295]** Song Y., Dixon S., Pearce M., Halpern A., Do online social tags predict perceived or induced emotional responses to music?, 14th International Society for Music Information Retrieval Conference, 2013.

**[296]** Sotiropoulos D.N., Lampropoulos A.S., Tsihrintzis G.A., MUSIPER: a system for modeling music similarity perception based on objective feature subset selection, User Modeling and User-Adapted Interaction, 18, 315-348, 2008.

**[297]** Spotify, https://www.spotify.com, access 04.09.2015.

**[298]** Stereomood, http://www.stereomood.com, access 5.12.2013.

**[299]** Stevens S. S., Volkmann J., Newman, E. B., A scale for the measurement of the psychological magnitude pitch, Journal of the Acoustical Society of America 8, 3, 185–190, 1937.

**[300]** Sturges M., Rhytmic structure in auditory temporal pattern perception and immediate memory, J. Exp. Psychol., 102, 377-383, 1974.

**[301]** Szczerba M., Recognition and Prediction of Music - A Machine Learning Approach, 106th AES Convention, Paper No: 4904, 1999.

**[302]** Szczuko P., Dalka P., Dabrowski M., Kostek B., MPEG-7-based Low-Level Descriptor Effectiveness in the Automatic Musical Sound Classification, 116th AES Convention, Paper No. 6105, Berlin, 2004.

**[303]** Tadeusiewicz R., Sieci neuronowe, Akademicka Oficyna Wydawnicza, Warszawa, 1993.

**[304]** Tadeusiewicz R., Chaki R., Chaki, N., Exploring Neural Networks with C#, CRC Press, 2014.

**[305]** Tanguiane A.S., Time Determination by Recognizing Generative Rhythmic Patterns, Musikometrika, 4, 83-99, Bochum: Brockmeyer, 1922.

**[306]** TC Electronic, http://www.tcelectronic.com/loudness/loudness-explained, access 11.08.2015.

**[307]** Tellegen A., Watson D., Clarck L. A., On the dimensional and hierarchical structure of affect, Psychological science, 10, 4, 297-303, 1999.

**[308]** Thayer R. E., The Biopsychology of Mood and Arousal, Oxford University Press, 1989.

**[309]** Thompson W. F., Music, thought, and feeling. Understanding the psychology of music, Oxford, Oxford University Press, 2009.

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

**[311]** Trohidis K., Tsoumakas G., Kalliris G, Vlahavas I., Multi - Label Classification of Music Into Emotions, International Symposium on Music Information Retrieval (ISMIR), 2008.

**[312]** Tsunoo E., Akase T., Ono N., Sagayama S., Music mood classification by rhythm and bass-line unit pattern analysis, ICASSP, 265–268, 2010.

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

**[314]** Tzanetakis G., Cook P., MARSYAS A Framework for Audio Analysis, Organized Sound, 4, 3, 73-80, 1999.

**[315]** [Tzanetakis 2007] Tzanetakis G., Marsyas a case study in implementing Music Information Retrieval Systems, Intelligent Music Information Systems Tools and Methodologies, 2007.

**[316]** Uhle C., Dittmar C., Sporer T., Extraction of drum tracks from polyphonic music using independent subspace analysis. Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation, 843-848, 2003.

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

**[318]** Umetrix, http://umetrics.com/sites/default/files/books/sample_chapters/multimega_parti-3_0.pdf, access 10.06.2015.

**[319]** Unehara M., YamadaK., Shimada T., Subjective evaluation of music with brain wave analysis for interactive music composition by IEC, Soft Computing and Intelligent Systems (SCIS), 66 - 70, 2014.

**[320]** Upton G., Cook I., A Dictionary of Statistics, Oxford University Press, 2014.

**[321]** Vaizman Y., Granot R.Y., Lanckriet G., Modeling dynamic patterns for emotional content in music, ISMIR, 747–752, 2011.

**[322]** Vatolkin I., Rudolph G., Interpretable Music Categorisation Based on Fuzzy Rules and High-Level Audio Features, Data Science, Learning by Latent Structures, and Knowledge Discovery, 423-432, Springer, Berlin Heidelberg, 2015.

**[323]** Vedala K., Mel-Hz plot, Own work, Licensed under CC BY-SA 3.0 via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Mel-Hz_plot.svg#/media/File:Mel-Hz_plot.svg, access 02.10.2015.

**[324]** Vickhoff B., A Perspective Theory of Music Perception and Emotion, Doctoral dissertation in musicology at the Department of Culture, Aesthetics and Media, University of Gothenburg, Sweden, 2008.

**[325]** Wagenaars W. M., Houtsma A. J., van Lieshout, R. A., Subjective Evaluation of Dynamic Compression in Music, JAES, 34, 1/2, 10-18, 1986.

**[326]** Wallin N. L., Brown S., Merker B., The Origins of Music, Cambridge, MIT Press, 2001.

**[327]** Wallis I., Ingalls T., Campana E., Goodman J., A rule-based generative music system controlled by desired valence and arousal, 8th International Sound and Music Computing (SMC) Conf., 2011.

**[328]** Wang J., Anguerra, X. Chen X., Yang D., Enriching music mood annotation by semantic association reasoning. Int. Conf. on Mult., 2010.

**[329]** Wang X., Chen X., Yang D., Wu Y., Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme, ISMIR, 765–770, 2011.

**[330]** Wang J.-C., Wang H.-M., Lanckriet G., A histogram density modeling approach to music emotion recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

**[331]** Watson, D., Mandryk, R., Modeling Musical Mood From Audio Features and Listening Context on an In-Situ Data Set, 13th International Society for Music Information Retrieval Conference, 2012.

**[332]** Webster G. D., Weir C. G.: Emotional Responses to Music: Interactive Effects of Mode, Texture, and Tempo, Motivation and Emotion, 29, 1, 2005.

**[333]** Weissenberger L., Toward a Universal, Meta-Theoretical Framework for Music Information Classification and Retrieval, Journal of Documentation, 71, 5, 2015.

**[334]** White J. D., The Analysis of Music, Englewood Cliffs, N.J., Prentice-Hall, 1976.

**[335]** Wieczorkowska A., Synak P., Ras Z.W., Multi-label classification of emotions in music, Intel. Info. Proc. and Web Min., 307–315, 2006.

**[336]** WiMP, http://www.wimpmusic.com, access 10.06.2015.

**[337]** Wójcik J., Kostek B., Intelligent Methods for Musical Rhythm Retrieval, Advanced Knowledge International, International Series on Advanced intelligence, 187 - 201, 2004.

**[338]** Wright B.A., Fitzgerald M. B., Sound – discrimination learning and auditory displays, International Conference on Auditory Display, 228–232, 2003.

**[339]** Wu B., Zhong E., Horner A., Yang Q., Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning, ACM International Conference on Multimedia, 117-126, 2014.

**[340]** XLStat, http://www.xlstat.com, access 10.06.2015.

**[341]** Xu R., Wunsch D.C.: Clustering, IEEE Press Series on Computational Intelligence, Wiley, New York, 2009.

**[342]** Xu J., Li X., Hao Y., Yang G., Source Separation Improves Music Emotion Recognition, International Conference on Multimedia Retrieval, New York, 2014.

**[343]** Yang Y. H., Chen H. H., Music Emotion Recognition, CRC Press, 2011.

**[344]** Yang Y., Hu X., Cross-cultural Music Mood Classification: A Comparison on English and Chinese Songs, *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 2012.

**[345]** Yang Y.-H., Lin Y.-C., Su Y.-F, Chen H., A Regression Approach to Music Emotion Recognition, IEEE Transactions on Audio, Speech, and Language Processing, 16, pp. 448-457, 2008.

**[346]** Zaanen, M., Kanters, P., Automatic Mood Classification Using TF*IDF Based on Lyrics, *11th International Society for Music Information Retrieval Conference*, 75-80, 2010.

**[347]** Zadeh L. A., Fuzzy Logic and its Application to Approximate Reasoning. Information Processing, 74, 591–594, 1974.

**[348]** Zadeh, L. A., The Concept of a Linguistic Variable and its Application to Approximate Reasoning, Information Sciences, 8, 199–249, 1975.

**[349]** Zentner M., Grandjean D., Scherer K., Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion, 8, 494-521, 2008.

**[350]** Zhang T., Kou J.C.-C., Heuristic approach for generic audio data segmentation and annotation, ACM Multimedia Conference, 67-76, 2009.

**[351]** Zhao Z., Xie L., Liu J., Wu W., The analysis of mood taxonomy comparison between Chinese and Western music, In: Proc. ICSPS, VI, 606–610, 2010.

**[352]** Zimmerman music paradox, [http://www.quora.com/The-Zimmerman-music-paradox-Music-is-not-sound-How-absurd-do-you-find-the-above-statement], access 08.10.2015.

**[353]** Żurada J., Barski M., Jędruch W., Sztuczne sieci neuronowe. Podstawy teorii i zastosowania (in Polish), Wydawnictwo PWN, Warszawa, 1996.

**[354]** Żwan P., Expert System for Automatic Classification and Quality Assessment of Singing Voices, 121st AES Convention, Paper No. 6898, San Francisco, USA, 2006.

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **1** | Blues | Joe Bonamassa | A New Day Yesterday Live | Cradle Rock |
| **2** | Blues | 13, Lester Butler | 13 Featuring Lester Butler | Boogie Disease |
| **3** | Blues | The Holmes Brothers | Feed My Soul | Take Me Away |
| **4** | Blues | Kenny Wayne Shepherd | 10 Days Out: Blues From The Backroads | Red Rooster |
| **5** | Blues | Ray Charles | Ray Charles - 93 Essential Tracks | Leave My Woman Alone |
| **6** | Blues | Cephas & Wiggins | Shoulder To Shoulder | Seattle Rainy Day Blues |
| **7** | Blues | Forrest Lee Jr. | Telethon | Telethon |
| **8** | Blues | Ronnie Baker Brooks | Golddigger | Make These Blues Survive |
| **9** | Blues | The Holiday Band | Sweet Love | She Sure Got A Way With My Heart |
| **10** | Blues | Lucky Peterson | Heart Of Pain | He's the Answer |
| **11** | Blues | Guy Davis | Give In Kind | Watch Over Me |
| **12** | Blues | Paul Butterfield's Better Days | Paul Butterfield's Better Days | Baby Please Don't Go |
| **13** | Blues | Robin Rogers | Back In The Fire | You Don't Know |
| **14** | Blues | Clarence "Gatemouth" Brown | Gates's On The Heat | Three Weeks And A Suitcase |
| **15** | Blues | Phantom Blues Band | Footprints | Leave Home Girl |
| **16** | Classical | Sarah Chang Charles Dutoit Royal Concert gebouw Orchestra | Vieuxtemps Lalo Violin Concertos | Symphonie espagnole Op. 21: I. Allegro non troppo |
| **17** | Classical | Berlin Symphony Orchestra and Eduardo Marturet | The 99 Darkest Pieces Of Classical Music | Tragic Overture In D Minor, Op. 81: Allegro Ma Non Troppo - Molto Pi Moderato - Tempo Primo Ma Tranquillo |
| **18** | Classical | Simone Kermes | Purcell: Dido & Aeneas | Dido & Aeneas, Act III: Dido's Lament |
| **19** | Classical | Ton Koopman | J.S. Bach Cantatas Vol. 1 | "Ich hatte viel Bekdmmernis" BWV 21: Prima Parte - Aria (Soprano): "Seufzer, Troenen, Kummer, Not" |
| **20** | Classical | Sir Neville Marriner Academy of St Martin-in-the-Fields | Bizet: Symphony in C L'Arlesienne Suites Nos. 1 & 2 | Symphony in C: Final : Allegro vivace |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **21** | Classical | Balazs Szokolay | Classical Piano Music | The Tale of Tsar Saltan: Flight of the Bumble-bee |
| **22** | Classical | Wiener Philharmoniker | Tchaikovsky: Ballet Suites - Swan Lake; Sleeping Beauty; The Nutcracker | Nutcracker Suite, Op.71a - Russian Dance (Trepak) |
| **23** | Classical | Pachelbel Orchestra | J.S. Bach: Air on the G String | Orchestral Suite No. 3 in D Major, BWV 1068: II. Air |
| **24** | Classical | Vaclav Neumann, Leipzig Gewandhaus Orchestra | Mussorgsky, M.P.: Pictures At An Exhibition Liszt, F.: Les Preludes Smetana, B.: Moldau (Kegel, Neumann) | Ma vlast (My Fatherland): Ma vlast (My Fatherland): No. 2. Vltava (Moldau) |
| **25** | Classical | Gidon Kremer | Mozart: Sinfonia Concertante K.364; Violin Concerto No.1 | Sinfonia concertante for Violin, Viola and Orchestra in E flat, K.364 - 3. Presto |
| **26** | Classical | Columbia Symphony Orchestra;Bruno Walter | Bruno Walter Conducts and Talks About Mahler: Symphony No.9 | Symphony No. 9 in D minor IV. Adagio. Sehr langsam und noch zuruckhaltend |
| **27** | Classical | Choir of King's College, Cambridge Sinfonia of London Stephen Cleobury | John Rutter: The Platinum Collection | Requiem: 6. The Lord is my shepherd (Christopher Hooker, oboe) |
| **28** | Classical | John Bayless | The Movie Album (Classical Pictures) | Il Postino from "The Postman"Visi d'Arte, from "Tosca" |
| **29** | Classical | Murray Perahia;English Chamber Orchestra | Mozart: Concerto for Piano and Orchestra No. 26 & Rondos in D & A Major | Rondo in D Major for Piano and Orchestra, K.382 (Instrumental) |
| **30** | Classical | John Rutter | The Handel Collection | Hornpipe (from the Water Music) |
| **31** | Country | The Bone Collector | The Brotherhood Album | Duck Blind (Feat. Rhett Akins & Dallas Davidson) |
| **32** | Country | John Rich | Rich Rocks | Let Somebody Else Drive (Feat. Hank Williams, Jr.) |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **33** | Country | LeAnn Rimes | You Light Up My Life | Amazing Grace |
| **34** | Country | Reba McEntire | All The Women I Am | If I Were A Boy |
| **35** | Country | Kellie Pickler | Kellie Pickler | Don't You Know You're Beautiful |
| **36** | Country | Hank Williams III | Rebel Within | Rebel Within |
| **37** | Country | Ryan Shupe & The Rubberband | Live | The Devil Went Down to Georgia |
| **38** | Country | Brad Paisley | Hits Alive | I'm Still A Guy |
| **39** | Country | Dolly Parton & Kenny Rogers | 16 Biggest Hits | Islands In The Stream |
| **40** | Country | Dana Lyons | Three Legged Coyote | Sweet New Orleans |
| **41** | Country | Allison Moorer | The Definitive Collection | A Soft Place To Fall |
| **42** | Country | Billy Currington | Little Bit Of Everything | No One Has Eyes Like You |
| **43** | Country | Hank Williams, Jr. | American Legends: Best Of The Early Years | The Last Love Song |
| **44** | Country | Marty Robbins | The Essential Marty Robbins | The Cowboy In The Continental Suit (Single Version) |
| **45** | Country | Yonder Mountain String Band | Mountain Tracks: Vol. 3 | Too Late Now |
| **46** | Dance & DJ | Mstrkrft | Bounce - Single | Bounce (feat. N.O.R.E.) [Extended Version] |
| **47** | Dance & DJ | M83 | Dead Cities, Red Seas & Lost Ghosts | Gone |
| **48** | Dance & DJ | Bonobo | Days To Come | Between The Lines (feat. Bajka) |
| **49** | Dance & DJ | Velvet | Fix Me | Fix Me (Radio Edit) |
| **50** | Dance & DJ | To Kool Chris | Droppin' That Old Skool Vol 1 | The Love I Lost - Mark Imperial |
| **51** | Dance & DJ | Benny Benassi | Electroman | Cinema (Feat. Gary Go) (Radio Edit) |
| **52** | Dance & DJ | Kelis | Flesh Tone (Amazon MP3 Exclusive Version) | Acapella |
| **53** | Dance & DJ | Royksopp | Melody AM | In Space |
| **54** | Dance & DJ | Vanessa Daou | Zipless | Near the Black Forest |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **55** | Dance & DJ | Bonobo, Andreya Triana | Black Sands | The Keeper |
| **56** | Dance & DJ | Morcheeba | Dive Deep | Enjoy The Ride |
| **57** | Dance & DJ | Daft Punk | Daft Club | Something About Us (Love Theme From Interstella 5555) |
| **58** | Dance & DJ | Moby | Destroyed | After |
| **59** | Dance & DJ | Chic | Dance, Dance, Dance: The Best Of Chic | Le Freak |
| **60** | Dance & DJ | Change | Journey Into Paradise: The Larry Levan Story | Paradise |
| **61** | Hard Rock & Metal | Nightwish | End Of An Era | Slaying The Dreamer |
| **62** | Hard Rock & Metal | Death | The Sound Of Perseverance - Reissue | Story To Tell |
| **63** | Hard Rock & Metal | T.T. Quick | Metal Of Honor | Hell To Pay |
| **64** | Hard Rock & Metal | Tyr | By the Light of the Northern Star | Into The Storm |
| **65** | Hard Rock & Metal | Skrewdriver | Blood & Honour | Street Fight (1996) |
| **66** | Hard Rock & Metal | Rhapsody | Symphony Of Enchanted Lands | Emerald sword |
| **67** | Hard Rock & Metal | Trivium | Ascendancy | Dying In Your Arms |
| **68** | Hard Rock & Metal | Motorhead | Overkill | Metropolis |
| **69** | Hard Rock & Metal | Kamelot | Poetry For The Poisoned | House On A Hill |
| **70** | Hard Rock & Metal | Kamelot | Ghost opera | EdenEcho |
| **71** | Hard Rock & Metal | Van Canto | Tribe of Force | Last Night Of The Kings |
| **72** | Hard Rock & Metal | Tyr | By the Light of the Northern Star | Turid Torkilsdottir |
| **73** | Hard Rock & Metal | Nightwish | Wishmaster | Two For Tragedy |
| **74** | Hard Rock & Metal | Reedom | World's Greatest Tribute To System Of A Down | Sugar |
| **75** | Hard Rock & Metal | Trust Company | Dreaming In Black And White | Almost There |
| **76** | Jazz | Martijn Schok | Solo Boogie Woogie And Blues Piano | Happy Piano Shuffle |
| **77** | Jazz | John Coltrane | Giant Steps | Giant Steps |
| **78** | Jazz | Rosin Coven | Penumbra | Lion Song |
| **79** | Jazz | Tord Gustavsen Trio | Being There | At Home |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **80** | Jazz | Chuck Mangione | Greatest Hits: Chuck Mangione | Children Of Sanchez |
| **81** | Jazz | George Benson | Breezin' | Six To Four |
| **82** | Jazz | Marcus Miller | Marcus | What Is Hip? |
| **83** | Jazz | Les Demerle | Cookin' at the Corner Vol. 1 | Cute |
| **84** | Jazz | Regina Carter | Motor City Moments | Don't Mess With Mr. T |
| **85** | Jazz | Herbie Hancock | Empyrean Isles (The Rudy Van Gelder Edition) | Oliloqui Valley |
| **86** | Jazz | Chet Baker | My Funny Valentine | My Funny Valentine |
| **87** | Jazz | Diana Krall | When I Look In Your Eyes | I've Got You Under My Skin |
| **88** | Jazz | Paul Hardcastle | The Collection | Moments In Time |
| **89** | Jazz | Preservation Hall Jazz Band | Best of Preservation Hall Jazz Band | Tiger Rag |
| **90** | Jazz | Les McCann & Eddie Harris | Hommage Nesuhi | Compared To What [Live at Montreux Jazz Festival, June 21 or 22, 1969] |
| **91** | Pop | P!nk | I'm Not Dead | I'm Not Dead |
| **92** | Pop | Avril Lavigne | The Best Damn Thing | The Best Damn Thing |
| **93** | Pop | Sting | The Dream Of The Blue Turtles | Fortress Around Your Heart |
| **94** | Pop | Shivaree | Breach | I Close My Eyes |
| **95** | Pop | Travie McCoy | Lazarus | We'll Be Alright |
| **96** | Pop | N Sync | 'N Sync | Tearin' Up My Heart |
| **97** | Pop | Selena Gomez & The Scene | A Year Without Rain | Summer's Not Hot |
| **98** | Pop | Jessie and The Toy Boys | Push It Feat. Yelawolf - Single | Push It Feat. Yelawolf |
| **99** | Pop | Finley Quaye;William Orbit | Dice | Dice |
| **100** | Pop | Paula Abdul | Spellbound | Rush, Rush |
| **101** | Pop | Smashing Pumpkins | Rarities & B-Sides | Landslide |
| **102** | Pop | Jann Arden | Living Under June | I Would Die For You |
| **103** | Pop | Alex Bugnon | Love Season | Around 12:15 AM |
| **104** | Pop | Holly & The Italians | The Right To Be Italian | Means To A Den |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **105** | Pop | Colbie Caillat | Breakthrough | You Got Me |
| **106** | R&B | Cam'ron | S.D.E. | What Means The World To You |
| **107** | R&B | George Clinton & Parliament Funkadelic | Live At Montreux 2004 | Not Just Knee Deep (Reprise) |
| **108** | R&B | Brian McKnight | Ten | The Rest Of My Life |
| **109** | R&B | Quincy Jones | Q's Jook Joint | Moody's Mood For Love (I'm In The Mood For Love) |
| **110** | R&B | Jackie Wilson | The Ultimate Jackie Wilson | Reet Petite |
| **111** | R&B | Stevie Wonder | At The Close Of A Century | Happy Birthday |
| **112** | R&B | Jody Watley | Greatest Hits | Some Kind Of Lover |
| **113** | R&B | Tamia | Between Friends | Christmas Medley |
| **114** | R&B | Charlie Wilson | Charlie, Last Name Wilson | Let's Chill |
| **115** | R&B | The Cotillionaires | Summer - EP | Otis Redding |
| **116** | R&B | Kandi | Kandi Koated | Me And U |
| **117** | R&B | Janet Jackson | Janet Jackson's Rhythm Nation 1814 | Someday Is Tonight |
| **118** | R&B | Aaliyah | Age Ain't Nothing But A Number | Back & Forth |
| **119** | R&B | Patti Labelle | The Best Of Patti LaBelle 20th Century Masters The Millennium Collection | The Right Kinda Lover |
| **120** | R&B | Surface | The Best Of Surface... A Nice Time 4 Lovin | Happy |
| **121** | Rap & Hip-Hop | Twista | Kamikaze | Kill Us All |
| **122** | Rap & Hip-Hop | T.I. | No Mercy | Yeah Ya Know (Takers) |
| **123** | Rap & Hip-Hop | Jadakiss | Kiss Of Death | By Your Side |
| **124** | Rap & Hip-Hop | Tyler, The Creator | Goblin | Her |
| **125** | Rap & Hip-Hop | Sho Baraka | Lions and Liars | I.T.W.N.O.I. Ft. R-swift, Tedashii, Honey Larochelle, Benjah |
| **126** | Rap & Hip-Hop | 2lanez | Logged In - Single | Logged In |
| **127** | Rap & Hip-Hop | Stuey Rock | Shinin (feat. Future) - Single | Shinin (feat. Future) |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **128** | Rap & Hip-Hop | Cypress Hill featuring Pitbull and Marc Anthony | Armada Latina (Feat. Pitbull And Marc Anthony) | Armada Latina (Feat. Pitbull And Marc Anthony) |
| **129** | Rap & Hip-Hop | Macklemore | The Unplanned Mixtape | American |
| **130** | Rap & Hip-Hop | Lloyd Banks feat. Pusha T | H.F.M. 2 (Hunger For More 2) | Home Sweet Home (Feat. Pusha T) |
| **131** | Rap & Hip-Hop | The Notorious B.I.G. | Life After Death | The World Is Filled... (Featuring Too Short & Puff Daddy) |
| **132** | Rap & Hip-Hop | T.I. | Got Your Back (Feat. Keri Hilson) | Got Your Back (Feat. Keri Hilson) |
| **133** | Rap & Hip-Hop | Living Legends | Classic | Never Fallin' |
| **134** | Rap & Hip-Hop | Young MC | Stone Cold Rhymin' | Principal's Office |
| **135** | Rap & Hip-Hop | Salt-N-Pepa | The Best Of Salt-N-Pepa 20th Century Masters The Millennium Collection | None Of Your Business |
| **136** | Rock | Sentenced | Frozen (Deluxe Reissue) | Digging The Grave |
| **137** | Rock | Muse | Black Holes And Revelations (Updated 09 version) | Assassin |
| **138** | Rock | Austra | Feel It Break | The Choke |
| **139** | Rock | Iced Earth | Something Wicked This Way Comes (Limited Edition w Bonus Tracks) | Melancholy (Holy Martyr) |
| **140** | Rock | Deep Purple | Deepest Purple (30th Anniversary Edition) | Stormbringer (2009 remix) |
| **141** | Rock | Lmnt | Juliet Greatest Gift | Juliet (Single Edit) |
| **142** | Rock | Bloc Party | Intimacy | Talons |
| **143** | Rock | Relient K | The First Three Gears (2000-2003) [+Digital Booklet] | Pressing On |
| **144** | Rock | Genesis | Platinum Collection | In Too Deep |

APPENDIX I List of tracks used in key experiment (Chapter 7)

| No. | Genre | Artist | Album | Title |
|---|---|---|---|---|
| **145** | Rock | Buffalo Springfield | Buffalo Springfield Again | Everydays |
| **146** | Rock | My Chemical Romance | The Black Parade | Cancer |
| **147** | Rock | Lacuna Coil | Enjoy the Silence - EP | Virtual Environment |
| **148** | Rock | The Lovemongers | Whirlygig | Sand |
| **149** | Rock | Sky Sailing | An Airplane Carried Me To Bed | Brielle |
| **150** | Rock | The Submarines | Honeysuckle Weeks | You, Me and the Bourgeoisie |

| Multi-track recordings used for "Analysis by synthesis " (Section 4.6) | | | | |
|---|---|---|---|---|
| **No** | **Genre** | **Artist** | **Album** | **Title** |
| **A1** | Jazz | Gosia Guja with band | None | Artystka |
| **A2** | Metal | A.N. | Unknown | Unknown |
| **A3** | Pop | Annalie Wilson | Open Heart Circus | This Time is Different |
| **A4** | Rock | TCB | Unknown | Unknown |

APPENDIX II Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. *R* and *φ* are corresponding to the position of songs on the VA plane in polar coordinates

| No. | Valence | Arousal | St. Dev. Valence | St. Dev. Arousal | R | φ |
|---|---|---|---|---|---|---|
| **1** | 0.09 | 2.06 | 0.90 | 0.99 | 2.07 | 1.53 |
| **2** | 0.43 | 1.74 | 0.86 | 1.39 | 1.79 | 1.33 |
| **3** | -0.65 | -1.14 | 1.20 | 0.95 | 1.31 | -2.09 |
| **4** | 0.24 | 0.35 | 0.70 | 1.40 | 0.42 | 0.97 |
| **5** | 1.33 | 0.74 | 1.03 | 1.14 | 1.52 | 0.51 |
| **6** | 1.53 | 0.00 | 1.07 | 0.49 | 1.53 | 0.00 |
| **7** | 0.11 | 2.29 | 0.88 | 1.14 | 2.30 | 1.52 |
| **8** | 1.18 | 1.36 | 1.34 | 1.32 | 1.80 | 0.85 |
| **9** | 1.57 | -0.17 | 1.12 | 1.04 | 1.58 | -0.11 |
| **10** | 0.90 | 0.09 | 1.04 | 1.63 | 0.91 | 0.10 |
| **11** | 0.15 | -1.20 | 1.36 | 0.88 | 1.21 | -1.45 |
| **12** | 0.38 | -0.97 | 1.60 | 1.25 | 1.04 | -1.20 |
| **13** | 0.65 | -0.46 | 1.11 | 1.25 | 0.79 | -0.61 |
| **14** | 0.25 | -1.14 | 1.34 | 1.32 | 1.16 | -1.35 |
| **15** | 0.92 | 1.31 | 1.13 | 1.17 | 1.60 | 0.96 |
| **16** | 0.41 | 0.75 | 1.40 | 1.41 | 0.86 | 1.08 |
| **17** | 1.23 | 0.77 | 0.86 | 1.54 | 1.45 | 0.56 |
| **18** | -1.18 | -1.05 | 1.69 | 1.08 | 1.57 | -2.42 |
| **19** | 0.38 | -1.53 | 1.56 | 0.90 | 1.57 | -1.33 |
| **20** | 1.75 | 0.44 | 1.21 | 1.33 | 1.81 | 0.25 |
| **21** | 1.17 | 1.81 | 1.11 | 0.52 | 2.15 | 1.00 |
| **22** | 1.11 | 1.88 | 1.19 | 1.14 | 2.18 | 1.04 |
| **23** | 1.13 | -1.21 | 0.94 | 1.75 | 1.66 | -0.82 |
| **24** | 0.93 | -0.77 | 1.03 | 1.57 | 1.21 | -0.69 |
| **25** | -1.00 | -0.73 | 2.01 | 0.90 | 1.24 | -2.52 |
| **26** | -0.02 | -0.98 | 1.54 | 1.46 | 0.98 | -1.59 |
| **27** | 0.54 | -1.73 | 1.15 | 1.09 | 1.81 | -1.27 |
| **28** | 0.73 | -1.38 | 1.44 | 1.35 | 1.56 | -1.08 |
| **29** | 1.43 | -0.55 | 1.34 | 1.21 | 1.54 | -0.37 |
| **30** | 1.67 | 0.55 | 0.80 | 1.50 | 1.76 | 0.32 |
| **31** | 0.36 | 1.21 | 1.20 | 1.16 | 1.26 | 1.28 |
| **32** | 0.19 | 2.15 | 0.85 | 0.87 | 2.16 | 1.48 |
| **33** | -0.25 | -1.06 | 1.40 | 1.21 | 1.09 | -1.80 |
| **34** | 0.57 | 0.98 | 0.97 | 1.22 | 1.13 | 1.04 |
| **35** | 0.75 | 0.15 | 0.91 | 0.88 | 0.76 | 0.20 |
| **36** | 0.59 | 1.36 | 0.99 | 1.26 | 1.48 | 1.16 |

APPENDIX II Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. *R* and *φ* are corresponding to the position of songs on the VA plane in polar coordinates

| No. | Valence | Arousal | St. Dev. Valence | St. Dev. Arousal | R | φ |
|---|---|---|---|---|---|---|
| **37** | 0.45 | 1.66 | 1.01 | 1.17 | 1.72 | 1.31 |
| **38** | 0.35 | -0.02 | 1.27 | 1.18 | 0.35 | -0.07 |
| **39** | 0.48 | -0.75 | 0.57 | 0.63 | 0.89 | -1.00 |
| **40** | -0.71 | -0.55 | 1.13 | 1.13 | 0.90 | -2.49 |
| **41** | -0.44 | -1.43 | 1.21 | 0.99 | 1.50 | -1.87 |
| **42** | 0.94 | -1.55 | 1.10 | 0.78 | 1.81 | -1.03 |
| **43** | 0.12 | -1.39 | 0.98 | 0.91 | 1.39 | -1.48 |
| **44** | 1.99 | -0.22 | 0.85 | 0.60 | 2.00 | -0.11 |
| **45** | 1.21 | 0.97 | 1.27 | 1.25 | 1.55 | 0.68 |
| **46** | -0.81 | 1.77 | 1.07 | 1.21 | 1.94 | 2.00 |
| **47** | -0.21 | -0.61 | 1.46 | 1.01 | 0.64 | -1.90 |
| **48** | 0.37 | 0.50 | 1.21 | 1.76 | 0.62 | 0.93 |
| **49** | 0.04 | 2.31 | 0.71 | 1.08 | 2.31 | 1.55 |
| **50** | 0.34 | 1.28 | 0.80 | 0.92 | 1.33 | 1.31 |
| **51** | 0.01 | 1.26 | 0.89 | 1.26 | 1.26 | 1.56 |
| **52** | 0.04 | 1.17 | 0.75 | 1.47 | 1.17 | 1.53 |
| **53** | 0.59 | -1.32 | 1.37 | 0.80 | 1.45 | -1.15 |
| **54** | 0.53 | -0.60 | 1.05 | 1.45 | 0.80 | -0.84 |
| **55** | 0.63 | -1.40 | 1.30 | 1.13 | 1.54 | -1.15 |
| **56** | -0.18 | -0.95 | 0.94 | 0.81 | 0.96 | -1.76 |
| **57** | 0.85 | -1.14 | 1.11 | 1.24 | 1.42 | -0.93 |
| **58** | -0.64 | -0.64 | 0.97 | 0.97 | 0.90 | -2.36 |
| **59** | 1.54 | 0.73 | 1.35 | 1.08 | 1.70 | 0.44 |
| **60** | 0.72 | 1.54 | 0.94 | 1.02 | 1.70 | 1.13 |
| **61** | -1.30 | 1.51 | 1.11 | 1.18 | 1.99 | 2.28 |
| **62** | -1.72 | 1.86 | 0.66 | 0.44 | 2.53 | 2.32 |
| **63** | -0.22 | 0.03 | 0.84 | 1.54 | 0.22 | 3.01 |
| **64** | -0.47 | 1.97 | 0.93 | 0.88 | 2.02 | 1.81 |
| **65** | -0.33 | 2.00 | 0.75 | 0.93 | 2.03 | 1.74 |
| **66** | 0.01 | 2.24 | 1.34 | 0.95 | 2.24 | 1.57 |
| **67** | -0.30 | 1.80 | 0.55 | 1.03 | 1.83 | 1.74 |
| **68** | 0.02 | 0.98 | 0.94 | 1.44 | 0.98 | 1.55 |
| **69** | -0.66 | -0.40 | 0.77 | 1.44 | 0.77 | -2.60 |
| **70** | -0.24 | 1.11 | 1.45 | 1.36 | 1.13 | 1.79 |
| **71** | -0.19 | -0.57 | 1.18 | 1.19 | 0.60 | -1.89 |
| **72** | -0.43 | 1.26 | 0.84 | 1.11 | 1.33 | 1.90 |

APPENDIX II Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. *R* and *φ* are corresponding to the position of songs on the VA plane in polar coordinates

| No. | Valence | Arousal | St. Dev. Valence | St. Dev. Arousal | R | φ |
|---|---|---|---|---|---|---|
| **73** | -1.00 | -1.33 | 1.19 | 1.06 | 1.66 | -2.21 |
| **74** | -1.08 | 1.90 | 0.87 | 0.80 | 2.19 | 2.09 |
| **75** | -0.26 | 1.14 | 1.38 | 1.08 | 1.17 | 1.80 |
| **76** | 0.77 | 1.37 | 1.36 | 1.31 | 1.57 | 1.06 |
| **77** | 0.72 | 0.59 | 1.21 | 1.52 | 0.94 | 0.69 |
| **78** | -1.45 | -1.25 | 1.35 | 1.14 | 1.92 | -2.43 |
| **79** | -0.79 | -1.35 | 1.37 | 1.37 | 1.56 | -2.10 |
| **80** | 1.04 | 0.89 | 0.89 | 1.65 | 1.37 | 0.71 |
| **81** | 1.00 | -0.32 | 1.15 | 1.08 | 1.05 | -0.31 |
| **82** | 0.47 | 1.61 | 1.06 | 1.42 | 1.68 | 1.29 |
| **83** | 1.03 | 1.26 | 1.09 | 1.77 | 1.62 | 0.88 |
| **84** | 1.25 | -0.72 | 0.82 | 1.49 | 1.44 | -0.52 |
| **85** | 0.96 | -0.62 | 1.05 | 1.57 | 1.15 | -0.57 |
| **86** | -0.11 | -1.26 | 2.05 | 1.08 | 1.26 | -1.66 |
| **87** | 0.67 | -1.45 | 1.42 | 0.90 | 1.59 | -1.14 |
| **88** | 0.71 | -1.42 | 0.93 | 0.85 | 1.59 | -1.11 |
| **89** | 1.85 | 0.61 | 1.10 | 1.05 | 1.95 | 0.32 |
| **90** | 0.92 | 0.61 | 1.04 | 1.79 | 1.10 | 0.58 |
| **91** | -0.56 | 1.30 | 0.78 | 1.43 | 1.42 | 1.98 |
| **92** | -0.16 | 1.74 | 0.99 | 1.12 | 1.75 | 1.66 |
| **93** | 0.57 | -0.32 | 0.94 | 1.58 | 0.65 | -0.52 |
| **94** | 0.29 | -0.55 | 1.36 | 0.94 | 0.63 | -1.08 |
| **95** | 1.08 | 0.83 | 1.32 | 1.14 | 1.36 | 0.66 |
| **96** | -0.28 | 0.56 | 0.95 | 1.56 | 0.63 | 2.03 |
| **97** | -0.81 | 1.38 | 1.09 | 1.58 | 1.60 | 2.10 |
| **98** | -0.24 | 1.76 | 0.94 | 1.24 | 1.77 | 1.70 |
| **99** | -0.17 | -0.94 | 1.04 | 0.95 | 0.95 | -1.75 |
| **100** | 0.19 | -0.96 | 0.91 | 0.98 | 0.97 | -1.37 |
| **101** | -0.66 | -0.87 | 1.11 | 1.15 | 1.09 | -2.22 |
| **102** | -0.56 | -0.74 | 1.51 | 0.93 | 0.93 | -2.22 |
| **103** | 0.43 | -0.12 | 0.77 | 1.24 | 0.45 | -0.27 |
| **104** | 0.71 | 1.21 | 0.99 | 1.06 | 1.40 | 1.04 |
| **105** | 0.69 | -0.90 | 0.92 | 1.13 | 1.14 | -0.92 |
| **106** | -0.05 | 1.17 | 0.78 | 1.08 | 1.17 | 1.61 |
| **107** | 0.38 | 0.72 | 1.33 | 1.28 | 0.81 | 1.08 |
| **108** | -0.10 | -0.91 | 1.18 | 1.41 | 0.92 | -1.68 |

APPENDIX II Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. *R* and *φ* are corresponding to the position of songs on the VA plane in polar coordinates

| No. | Valence | Arousal | St. Dev. Valence | St. Dev. Arousal | R | φ |
|---|---|---|---|---|---|---|
| **109** | 0.91 | -1.67 | 1.21 | 0.76 | 1.90 | -1.07 |
| **110** | 1.48 | 1.22 | 1.39 | 1.40 | 1.92 | 0.69 |
| **111** | 0.50 | 0.41 | 1.00 | 1.10 | 0.65 | 0.69 |
| **112** | 0.37 | 1.44 | 0.96 | 1.31 | 1.49 | 1.32 |
| **113** | 0.66 | -0.76 | 0.79 | 1.27 | 1.01 | -0.85 |
| **114** | 0.73 | -1.47 | 0.90 | 1.14 | 1.64 | -1.11 |
| **115** | 0.40 | -1.78 | 1.15 | 0.71 | 1.82 | -1.35 |
| **116** | 0.10 | -0.39 | 0.72 | 1.42 | 0.40 | -1.32 |
| **117** | 0.54 | -1.18 | 0.92 | 0.66 | 1.30 | -1.14 |
| **118** | 0.36 | -1.51 | 0.75 | 0.91 | 1.56 | -1.34 |
| **119** | 0.66 | 0.46 | 0.94 | 1.22 | 0.80 | 0.61 |
| **120** | 0.43 | -0.51 | 0.76 | 1.13 | 0.67 | -0.86 |
| **121** | -1.12 | 1.42 | 0.83 | 0.58 | 1.81 | 2.24 |
| **122** | -0.59 | 1.84 | 1.33 | 0.75 | 1.93 | 1.88 |
| **123** | -0.52 | -0.35 | 1.08 | 1.26 | 0.63 | -2.55 |
| **124** | -0.75 | -0.43 | 1.08 | 0.98 | 0.86 | -2.61 |
| **125** | -0.79 | 0.88 | 0.86 | 1.44 | 1.18 | 2.30 |
| **126** | -0.41 | 1.57 | 1.20 | 0.83 | 1.62 | 1.82 |
| **127** | 0.06 | 0.55 | 1.40 | 1.55 | 0.55 | 1.46 |
| **128** | 0.46 | 1.01 | 1.18 | 1.14 | 1.11 | 1.14 |
| **129** | 1.01 | 0.29 | 1.11 | 0.90 | 1.05 | 0.28 |
| **130** | -0.86 | -0.02 | 1.05 | 1.33 | 0.86 | -3.12 |
| **131** | -0.07 | 0.17 | 1.07 | 1.09 | 0.19 | 1.93 |
| **132** | 0.64 | 0.59 | 1.06 | 1.24 | 0.87 | 0.75 |
| **133** | -0.16 | -0.14 | 1.25 | 1.33 | 0.22 | -2.43 |
| **134** | 0.83 | 0.65 | 0.96 | 1.07 | 1.06 | 0.67 |
| **135** | -0.26 | 1.25 | 1.04 | 1.32 | 1.27 | 1.77 |
| **136** | -1.37 | 1.96 | 0.85 | 0.51 | 2.39 | 2.18 |
| **137** | -0.53 | 1.79 | 0.94 | 1.09 | 1.86 | 1.86 |
| **138** | 0.10 | -0.05 | 1.54 | 1.34 | 0.12 | -0.48 |
| **139** | -0.99 | -0.12 | 1.54 | 0.84 | 1.00 | -3.02 |
| **140** | 0.28 | 1.52 | 0.86 | 1.06 | 1.54 | 1.39 |
| **141** | 0.68 | 1.21 | 1.49 | 1.24 | 1.39 | 1.06 |
| **142** | -0.66 | 2.03 | 1.05 | 0.67 | 2.13 | 1.89 |
| **143** | 0.52 | 1.69 | 1.37 | 1.30 | 1.77 | 1.27 |
| **144** | 0.57 | -0.91 | 1.21 | 1.20 | 1.07 | -1.02 |

APPENDIX II Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. *R* and *φ* are corresponding to the position of songs on the VA plane in polar coordinates

| No. | Valence | Arousal | St. Dev. Valence | St. Dev. Arousal | R | φ |
|---|---|---|---|---|---|---|
| **145** | 0.46 | -0.44 | 1.41 | 1.41 | 0.63 | -0.76 |
| **146** | 0.06 | 0.00 | 1.32 | 1.58 | 0.06 | 0.00 |
| **147** | -0.25 | -0.40 | 1.42 | 1.47 | 0.47 | -2.14 |
| **148** | -0.15 | -0.83 | 1.35 | 0.88 | 0.84 | -1.75 |
| **149** | 0.24 | -1.49 | 0.92 | 1.20 | 1.51 | -1.41 |
| **150** | 1.08 | 0.47 | 1.39 | 1.11 | 1.18 | 0.41 |

| No. | Valence | Arousal | St. Dev. Valence | St. Dev. Arousal | R | φ |
|---|---|---|---|---|---|---|
| **A1** | -0.12 | -1.00 | 0.75 | 0.91 | 1.01 | -1.69 |
| **A2** | -0.30 | 2.21 | 0.85 | 0.51 | 2.23 | 1.71 |
| **A3** | 0.91 | 0.94 | 1.08 | 1.26 | 1.31 | 0.80 |
| **A4** | -0.20 | 1.10 | 0.94 | 1.09 | 1.12 | 1.75 |

APPENDIX III Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. Rows indicate songs and columns mood labels. The value describes the percent of occurrences of each label.

| No. | Aggressive [%] | Depressive [%] | Exciting [%] | Energetic [%] | Neutral [%] | Relaxing [%] | Sad [%] | Calm [%] | Joyful [%] |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 20 | 0 | 7 | 67 | 0 | 0 | 0 | 0 | 7 |
| **2** | 0 | 0 | 5 | 74 | 5 | 5 | 0 | 0 | 11 |
| **3** | 0 | 18 | 0 | 0 | 0 | 12 | 41 | 29 | 0 |
| **4** | 0 | 0 | 0 | 42 | 8 | 17 | 17 | 8 | 8 |
| **5** | 0 | 0 | 14 | 24 | 5 | 5 | 0 | 0 | 52 |
| **6** | 0 | 0 | 0 | 9 | 9 | 5 | 5 | 5 | 68 |
| **7** | 7 | 0 | 0 | 80 | 7 | 0 | 0 | 0 | 7 |
| **8** | 0 | 0 | 6 | 53 | 0 | 0 | 0 | 0 | 41 |
| **9** | 0 | 0 | 0 | 13 | 7 | 13 | 7 | 0 | 60 |
| **10** | 0 | 5 | 21 | 21 | 5 | 37 | 5 | 0 | 5 |
| **11** | 0 | 10 | 0 | 0 | 15 | 25 | 35 | 10 | 5 |
| **12** | 0 | 13 | 7 | 0 | 7 | 27 | 27 | 7 | 13 |
| **13** | 0 | 5 | 16 | 11 | 0 | 37 | 21 | 5 | 5 |
| **14** | 0 | 12 | 0 | 12 | 6 | 24 | 35 | 6 | 6 |
| **15** | 0 | 0 | 18 | 47 | 6 | 6 | 0 | 0 | 24 |
| **16** | 13 | 6 | 38 | 19 | 0 | 19 | 0 | 6 | 0 |
| **17** | 7 | 0 | 53 | 7 | 7 | 27 | 0 | 0 | 0 |
| **18** | 0 | 25 | 0 | 0 | 0 | 19 | 13 | 44 | 0 |
| **19** | 0 | 0 | 0 | 0 | 0 | 40 | 40 | 20 | 0 |
| **20** | 0 | 0 | 24 | 18 | 0 | 6 | 6 | 0 | 47 |
| **21** | 5 | 0 | 65 | 30 | 0 | 0 | 0 | 0 | 0 |
| **22** | 0 | 0 | 28 | 50 | 0 | 0 | 0 | 0 | 22 |
| **23** | 0 | 0 | 21 | 0 | 0 | 36 | 36 | 0 | 7 |
| **24** | 0 | 0 | 19 | 5 | 5 | 43 | 19 | 5 | 5 |
| **25** | 0 | 13 | 0 | 0 | 6 | 25 | 6 | 50 | 0 |
| **26** | 0 | 0 | 11 | 0 | 0 | 21 | 37 | 32 | 0 |
| **27** | 0 | 0 | 0 | 0 | 13 | 38 | 44 | 6 | 0 |
| **28** | 0 | 12 | 6 | 0 | 0 | 41 | 24 | 12 | 6 |
| **29** | 6 | 0 | 0 | 6 | 0 | 29 | 12 | 0 | 47 |
| **30** | 0 | 0 | 35 | 10 | 0 | 15 | 5 | 0 | 35 |
| **31** | 13 | 0 | 13 | 44 | 13 | 0 | 6 | 0 | 13 |
| **32** | 11 | 0 | 11 | 72 | 0 | 0 | 0 | 0 | 6 |
| **33** | 0 | 10 | 10 | 0 | 5 | 20 | 25 | 30 | 0 |
| **34** | 0 | 0 | 5 | 58 | 11 | 5 | 0 | 0 | 21 |
| **35** | 0 | 0 | 0 | 20 | 20 | 0 | 10 | 0 | 50 |
| **36** | 0 | 0 | 7 | 60 | 7 | 7 | 0 | 0 | 20 |

APPENDIX III Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. Rows indicate songs and columns mood labels. The value describes the percent of occurrences of each label.

| No. | Aggressive [%] | Depressive [%] | Exciting [%] | Energetic [%] | Neutral [%] | Relaxing [%] | Sad [%] | Calm [%] | Joyful [%] |
|---|---|---|---|---|---|---|---|---|---|
| **37** | 6 | 0 | 13 | 63 | 6 | 0 | 0 | 0 | 13 |
| **38** | 0 | 4 | 17 | 8 | 4 | 13 | 17 | 25 | 13 |
| **39** | 0 | 0 | 0 | 0 | 13 | 27 | 40 | 0 | 20 |
| **40** | 0 | 26 | 5 | 11 | 5 | 5 | 21 | 26 | 0 |
| **41** | 0 | 25 | 0 | 0 | 0 | 15 | 40 | 20 | 0 |
| **42** | 0 | 0 | 0 | 0 | 0 | 50 | 39 | 6 | 6 |
| **43** | 0 | 0 | 0 | 0 | 5 | 26 | 58 | 11 | 0 |
| **44** | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 85 |
| **45** | 0 | 0 | 19 | 38 | 6 | 6 | 0 | 0 | 31 |
| **46** | 41 | 6 | 0 | 41 | 6 | 0 | 0 | 0 | 6 |
| **47** | 6 | 17 | 6 | 0 | 11 | 28 | 6 | 28 | 0 |
| **48** | 5 | 10 | 29 | 29 | 0 | 14 | 14 | 0 | 0 |
| **49** | 5 | 0 | 0 | 80 | 5 | 0 | 0 | 0 | 10 |
| **50** | 0 | 0 | 28 | 50 | 17 | 0 | 0 | 6 | 0 |
| **51** | 20 | 13 | 7 | 47 | 0 | 0 | 0 | 0 | 13 |
| **52** | 0 | 6 | 13 | 50 | 31 | 0 | 0 | 0 | 0 |
| **53** | 0 | 11 | 0 | 0 | 5 | 47 | 26 | 5 | 5 |
| **54** | 0 | 6 | 22 | 0 | 11 | 33 | 17 | 11 | 0 |
| **55** | 0 | 19 | 6 | 0 | 0 | 44 | 25 | 6 | 0 |
| **56** | 0 | 14 | 0 | 0 | 7 | 14 | 43 | 21 | 0 |
| **57** | 0 | 7 | 7 | 7 | 0 | 43 | 29 | 0 | 7 |
| **58** | 13 | 44 | 6 | 0 | 19 | 6 | 6 | 6 | 0 |
| **59** | 6 | 0 | 6 | 28 | 0 | 0 | 0 | 0 | 61 |
| **60** | 0 | 0 | 18 | 59 | 0 | 0 | 0 | 0 | 24 |
| **61** | 74 | 5 | 5 | 11 | 0 | 0 | 0 | 5 | 0 |
| **62** | 93 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| **63** | 16 | 11 | 16 | 11 | 16 | 0 | 21 | 11 | 0 |
| **64** | 30 | 0 | 5 | 60 | 0 | 0 | 0 | 5 | 0 |
| **65** | 24 | 0 | 6 | 65 | 6 | 0 | 0 | 0 | 0 |
| **66** | 25 | 0 | 13 | 56 | 0 | 0 | 0 | 0 | 6 |
| **67** | 25 | 0 | 0 | 69 | 0 | 0 | 0 | 6 | 0 |
| **68** | 13 | 7 | 7 | 47 | 7 | 13 | 0 | 0 | 7 |
| **69** | 6 | 24 | 0 | 12 | 18 | 0 | 18 | 24 | 0 |
| **70** | 29 | 21 | 21 | 21 | 0 | 0 | 0 | 7 | 0 |
| **71** | 0 | 6 | 13 | 0 | 13 | 13 | 19 | 38 | 0 |
| **72** | 29 | 6 | 6 | 47 | 6 | 0 | 0 | 6 | 0 |

APPENDIX III Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. Rows indicate songs and columns mood labels. The value describes the percent of occurrences of each label.

| No. | Aggressive [%] | Depressive [%] | Exciting [%] | Energetic [%] | Neutral [%] | Relaxing [%] | Sad [%] | Calm [%] | Joyful [%] |
|---|---|---|---|---|---|---|---|---|---|
| **73** | 0 | 25 | 0 | 0 | 0 | 0 | 50 | 25 | 0 |
| **74** | 65 | 0 | 0 | 29 | 6 | 0 | 0 | 0 | 0 |
| **75** | 37 | 5 | 0 | 26 | 11 | 0 | 0 | 0 | 21 |
| **76** | 8 | 0 | 8 | 46 | 8 | 0 | 0 | 0 | 31 |
| **77** | 6 | 0 | 22 | 28 | 6 | 28 | 0 | 6 | 6 |
| **78** | 5 | 68 | 5 | 0 | 0 | 0 | 11 | 11 | 0 |
| **79** | 0 | 43 | 7 | 0 | 7 | 7 | 21 | 14 | 0 |
| **80** | 0 | 0 | 47 | 26 | 0 | 16 | 11 | 0 | 0 |
| **81** | 0 | 0 | 0 | 13 | 31 | 25 | 6 | 0 | 25 |
| **82** | 5 | 0 | 10 | 57 | 0 | 10 | 0 | 0 | 19 |
| **83** | 0 | 0 | 21 | 50 | 0 | 17 | 0 | 0 | 13 |
| **84** | 0 | 0 | 23 | 0 | 8 | 54 | 15 | 0 | 0 |
| **85** | 0 | 10 | 24 | 5 | 0 | 29 | 19 | 0 | 14 |
| **86** | 0 | 11 | 0 | 0 | 6 | 39 | 11 | 33 | 0 |
| **87** | 0 | 5 | 0 | 0 | 0 | 58 | 16 | 21 | 0 |
| **88** | 0 | 6 | 0 | 0 | 13 | 50 | 31 | 0 | 0 |
| **89** | 0 | 0 | 17 | 17 | 0 | 6 | 0 | 0 | 61 |
| **90** | 5 | 0 | 25 | 30 | 0 | 20 | 10 | 0 | 10 |
| **91** | 11 | 11 | 0 | 56 | 6 | 0 | 0 | 17 | 0 |
| **92** | 32 | 0 | 0 | 53 | 5 | 0 | 0 | 0 | 11 |
| **93** | 6 | 0 | 29 | 6 | 6 | 18 | 29 | 6 | 0 |
| **94** | 0 | 0 | 10 | 0 | 15 | 30 | 10 | 25 | 10 |
| **95** | 6 | 0 | 6 | 29 | 6 | 0 | 0 | 0 | 53 |
| **96** | 10 | 10 | 5 | 35 | 10 | 5 | 5 | 15 | 5 |
| **97** | 27 | 9 | 0 | 41 | 9 | 0 | 0 | 9 | 5 |
| **98** | 17 | 0 | 6 | 61 | 11 | 6 | 0 | 0 | 0 |
| **99** | 0 | 5 | 5 | 0 | 5 | 11 | 53 | 21 | 0 |
| **100** | 0 | 6 | 6 | 0 | 6 | 18 | 47 | 12 | 6 |
| **101** | 6 | 24 | 6 | 0 | 6 | 0 | 41 | 18 | 0 |
| **102** | 0 | 6 | 0 | 0 | 17 | 17 | 22 | 39 | 0 |
| **103** | 0 | 6 | 0 | 33 | 22 | 33 | 6 | 0 | 0 |
| **104** | 0 | 0 | 5 | 60 | 0 | 0 | 0 | 0 | 35 |
| **105** | 0 | 0 | 0 | 11 | 0 | 32 | 37 | 5 | 16 |
| **106** | 11 | 0 | 11 | 56 | 6 | 6 | 0 | 11 | 0 |
| **107** | 13 | 13 | 38 | 13 | 13 | 6 | 0 | 6 | 0 |
| **108** | 0 | 6 | 11 | 6 | 6 | 11 | 44 | 17 | 0 |

APPENDIX III Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. Rows indicate songs and columns mood labels. The value describes the percent of occurrences of each label.

| No. | Aggressive [%] | Depressive [%] | Exciting [%] | Energetic [%] | Neutral [%] | Relaxing [%] | Sad [%] | Calm [%] | Joyful [%] |
|---|---|---|---|---|---|---|---|---|---|
| **109** | 0 | 5 | 0 | 0 | 5 | 62 | 24 | 5 | 0 |
| **110** | 0 | 0 | 5 | 42 | 0 | 0 | 0 | 0 | 53 |
| **111** | 0 | 6 | 0 | 31 | 25 | 6 | 6 | 0 | 25 |
| **112** | 0 | 0 | 0 | 63 | 6 | 0 | 0 | 6 | 25 |
| **113** | 0 | 0 | 6 | 19 | 0 | 44 | 31 | 0 | 0 |
| **114** | 0 | 0 | 0 | 7 | 0 | 60 | 27 | 7 | 0 |
| **115** | 0 | 6 | 0 | 0 | 0 | 33 | 56 | 6 | 0 |
| **116** | 6 | 6 | 11 | 11 | 6 | 11 | 33 | 11 | 6 |
| **117** | 0 | 12 | 6 | 0 | 0 | 47 | 35 | 0 | 0 |
| **118** | 0 | 0 | 0 | 0 | 5 | 32 | 58 | 5 | 0 |
| **119** | 6 | 0 | 22 | 28 | 11 | 17 | 11 | 0 | 6 |
| **120** | 0 | 5 | 0 | 11 | 26 | 21 | 21 | 0 | 16 |
| **121** | 71 | 0 | 0 | 24 | 6 | 0 | 0 | 0 | 0 |
| **122** | 50 | 0 | 11 | 33 | 6 | 0 | 0 | 0 | 0 |
| **123** | 4 | 17 | 0 | 9 | 26 | 9 | 13 | 17 | 4 |
| **124** | 0 | 25 | 13 | 0 | 19 | 0 | 13 | 31 | 0 |
| **125** | 47 | 12 | 0 | 29 | 12 | 0 | 0 | 0 | 0 |
| **126** | 41 | 0 | 6 | 41 | 0 | 0 | 0 | 0 | 12 |
| **127** | 18 | 24 | 18 | 24 | 0 | 0 | 0 | 0 | 18 |
| **128** | 6 | 6 | 6 | 50 | 11 | 0 | 0 | 0 | 22 |
| **129** | 7 | 0 | 7 | 20 | 7 | 13 | 0 | 0 | 47 |
| **130** | 26 | 26 | 0 | 16 | 5 | 5 | 5 | 16 | 0 |
| **131** | 29 | 6 | 6 | 6 | 12 | 18 | 12 | 6 | 6 |
| **132** | 6 | 0 | 11 | 28 | 22 | 11 | 0 | 0 | 22 |
| **133** | 22 | 17 | 6 | 6 | 0 | 17 | 11 | 17 | 6 |
| **134** | 0 | 0 | 6 | 35 | 12 | 0 | 6 | 0 | 41 |
| **135** | 35 | 0 | 6 | 35 | 12 | 0 | 6 | 0 | 6 |
| **136** | 76 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 |
| **137** | 33 | 0 | 7 | 47 | 0 | 0 | 0 | 13 | 0 |
| **138** | 0 | 15 | 23 | 0 | 15 | 15 | 0 | 31 | 0 |
| **139** | 6 | 12 | 6 | 0 | 12 | 12 | 0 | 53 | 0 |
| **140** | 10 | 0 | 15 | 60 | 0 | 5 | 5 | 0 | 5 |
| **141** | 19 | 0 | 6 | 31 | 0 | 0 | 0 | 0 | 44 |
| **142** | 44 | 0 | 6 | 50 | 0 | 0 | 0 | 0 | 0 |
| **143** | 10 | 5 | 19 | 52 | 0 | 0 | 0 | 0 | 14 |
| **144** | 0 | 0 | 10 | 0 | 5 | 25 | 45 | 10 | 5 |

APPENDIX III Results of key experiment described in Chapter 7, tracks are indexed according to Appendix I. Rows indicate songs and columns mood labels. The value describes the percent of occurrences of each label.

| No. | Aggressive [%] | Depressive [%] | Exciting [%] | Energetic [%] | Neutral [%] | Relaxing [%] | Sad [%] | Calm [%] | Joyful [%] |
|---|---|---|---|---|---|---|---|---|---|
| **145** | 8 | 0 | 8 | 0 | 15 | 23 | 23 | 8 | 15 |
| **146** | 0 | 6 | 18 | 18 | 12 | 12 | 12 | 24 | 0 |
| **147** | 6 | 35 | 12 | 6 | 0 | 18 | 6 | 18 | 0 |
| **148** | 0 | 11 | 0 | 0 | 6 | 11 | 33 | 22 | 17 |
| **149** | 0 | 6 | 0 | 0 | 22 | 28 | 39 | 6 | 0 |
| **150** | 0 | 0 | 6 | 28 | 6 | 0 | 6 | 6 | 50 |

| No. | Aggresive | Depressive | Exciting | Energetic | Neutral | Relaxing | Sad | Calm | Joyful |
|---|---|---|---|---|---|---|---|---|---|
| **A1** | 0 | 10 | 0 | 0 | 0 | 28 | 50 | 12 | 0 |
| **A2** | 71 | 0 | 0 | 20 | 10 | 0 | 0 | 0 | 0 |
| **A3** | 6 | 0 | 29 | 6 | 6 | 18 | 29 | 6 | 0 |
| **A4** | 10 | 0 | 15 | 60 | 0 | 5 | 5 | 0 | 5 |