University of Coimbra - Portugal

Faculty of Science and Technology

Department of Informatics Engineering

A MUSICAL SYSTEM FOR EMOTIONAL EXPRESSION

PH.D. THESIS

DOCTORAL PROGRAM IN INFORMATION SCIENCES AND TECHNOLOGIES

AREA OF ARTIFICIAL INTELLIGENCE

António Pedro Oliveira

Advisor: Dr. Amílcar Cardoso

April 12, 2013

ACKNOWLEDGEMENTS

I would like to express my thanks to:

- My parents and sister, as well as my grandmother and grandfather for all the personal support given day by day.
- All my friends, for all the shared moments of joy, during times of rest of the thesis.
- Dr. Amílcar Cardoso, for support, kind words of advice and continuous encouragement during my Ph.D..
- Armando Oliveira and Alba Grieco from the Faculty of Psychology of the University, for al the support given during the experiments of the calibration/validation of the system.
- Rui Pedro Paiva, of the Centre for Informatics and Systems of the University of Coimbra, for guiding me in the initial stages of the thesis and for reviewing my thesis.
- Penousal Machado, of the Centre for Informatics and Systems of the University of Coimbra, for helping me in analysing experimental data.
- Students from the Department of Informatics Engineering, for answering to the questionnaire given to obtain data for the calibration/validation of the system.
- Students from the Faculdade de Psicologia e de Ciências da Educação, for participating in the behavioral and physiological experiment in the calibration/validation of the system.
- Students from the Department of Informatics Engineering, for participating in the application of the EDME system as an emotion-driven music engine.
- All the people from different areas of all around the world for answering to the web-based questionnaires, that helped me in obtaining experimental data.
- Diana Taborda, Jorge Ávila, Ricardo Ruivo, Carlos Figueiredo and Isabel Lourenço for all the academic support given during the development of the thesis.

RESUMO

O controlo automático da expressão emocional na música (tonal) é um desafio que está longe de ser resolvido. Esta tese apresenta o EDME - um sistema que pode ser usado para a geração de novas peças musicais que exprimem uma determinado emoção especificada pelo utilizador. O sistema funciona com ficheiros MIDI standard e está dividido em duas etapas: a primeira off-line, a segunda on-line. Na primeira etapa, os ficheiros MIDI são divididos em segmentos com conteúdo emocional uniforme. Estes são submetidos a um processo de extracção de características, sendo posteriormente classificados de acordo com os valores emocionais de valência e activação e armazenados numa base de músicas. Na segunda etapa, os segmentos são seleccionados e transformados de acordo com a emoção especificada pelo utilizador e, em seguida, arranjadas de acordo com uma forma musical. A modularidade, adaptabilidade e flexibilidade da arquitectura do nosso sistema torna-o aplicável em contextos diversos, como vídeo-jogos, teatro, filmes e contextos de saúde.

O sistema está a usar uma base de conhecimento, baseada em resultados empíricos de obras de Psicologia da Música, tendo sido aperfeiçoado com dados experimentais obtidos com questionários. Para as experimentais, preparamos questionários com segmentos musicais de conteúdo emocional diferente. Após ouvir cada segmento, cada indivíduo classificou-o com valores de valência e excitação. Inferimos que as experiências conduzidas via web tinham um elevado grau de fiabilidade, apesar de terem sido feitas num contexto não-controlado.

Nós também calibramos/validamos o sistema EDME em duas experiências destinadas a verificar a precisão do sistema na classificação de valência e excitação usando dados experimentais obtidos num ambiente controlado. A primeira experiência obteve dados através de questionários com base no Self-Assessment Manikin. Na segunda experiência obtivemos dados fisiológicos e comportamentais. Os dados mostraram que a actividade do músculo corrugador aumenta com a excitação; os batimentos por minuto da frequência cardíaca aumentam com a excitação, a resposta galvânica da pele aumenta com a valência e excitação. Apenas na atividade do músculo zigomático há um aumento significativo em ambos, valência e excitação.

ABSTRACT

The automatic control of emotional expression in (tonal) music is a challenge that is far from being solved. This thesis presents EDME - a system with such capabilities used for the generation of novel musical works which express a particular emotion as specified by the user. The system works with standard MIDI files and develops in two stages: the first offline, the second online. In the first stage, MIDI files are partitioned in segments with uniform emotional content. These are subjected to a process of feature extraction, then classified according to emotional values of valence and arousal and stored in a music base. In the second stage, segments are selected and transformed according to user specified emotion and then arranged into song-like structures. The modularity, adaptability and flexibility of our system's architecture make it applicable in various contexts like video-games, theatre, films and healthcare contexts.

The system is using a knowledge base, grounded on empirical results of works of Music Psychology, which was refined with experimental data obtained with questionnaires. For the experimental setups, we prepared questionnaires with musical segments of different emotional content. Each subject classified each segment after listening to it, with values for valence and arousal. We inferred that the experiments conducted via online had a high degree of reliability, despite the fact of being done in a non-controlled context.

We also calibrated/validated EDME in two experiments where we intended to verify the accuracy of EDME in classifying valence and arousal by using experimental data obtained in a controlled environment. The first experiment collected data with questionnaires based on Self-Assessment Manikin. The second experiment collected behavioral and physiological data. The data show that corrugator muscle activity increase with arousal; heart rate measure in beats per minute increase with arousal, and galvanic skin response increase with both valence and arousal. Only for zigomatic muscle activity there is a significant increase with both, valence and arousal.

KEYWORDS

Knowledge-based system, automatic music production, expression of emotions, music and emotions, real-time system, tonal music.

Contents

List of Figures

List of Tables

Part I.

Introduction

1. Motivation

"Music can change the world because it can change people."
– *Bono (U2)*

Emotions are widely accepted as being an important factor in the society. Their multidimensional nature is the main reason why there is still so much to discover in order to understand them. Throughout history, many scientists have dedicated most of the time of their lives to study emotions (Damásio and Sutherland, 1996; Ekman, 1999; Frijda, 2000; Lazarus, 1999; Ortony and Collins, 1988); however, there is not yet a consensus in an important aspect as is their definition (Scherer, 2005). They belong to an extended area which is the area of affects. Scherer (2000) suggest that emotions are among five types of affects: emotions, moods, interpersonal stances, preferences and affect dispositions. There are two main dimensions that usually help to distinguish between emotions from the other types of affect, they are the duration and intensity. Emotions are characterized by having the highest intensity and the lowest duration. We accept emotions as corresponding to the manifestation of our psychophysiological state (Larsen et al., 2008).

Music is another area with many repercussions in society. Like emotions, they also have a multidimensional nature with also so many to discover in order to understand the processes involved in our mind while listening to the music. Nowadays, music is almost everywhere, and the most interesting fact is that it is a powerful stimulus capable of influencing our emotions. This is evidenced by research findings on Music Psychology (Deutsch, 1982; Lerdahl and Jackendoff, 1983; Narmour, 1990; Temperley, 2004; Widmer and Goebl, 2004; Gabrielsson and Lindstrom, 2001). There are established relations between musical and emotional areas. For instance, tempo is widely accepted as having direct influence on the pleasantness of emotions.

The scientific challenge of automatically producing music with an appropriate emotional content has involved a lot of research in emotional and musical domains. Many research areas have been working to reduce the semantic gap that exists between these two domains (Serra et al., 2007). We are focused in the areas of Music Psychology, Music Computing and Affective Computing. Computational systems with the capability of producing music with an appropriate emotional content have an enormous application potential, which makes them usable in every context where there is a need to create environments capable of inducing certain emotional experiences. The production of soundtracks for video-games, films and theatre are examples. They can also be applied in hospitals, shopping centres, gymnasiums and houses of worship places. This motivated the development of Emotion-Driven Music Engine (EDME), a system with the mentioned capabilities.

2. Aim

The central goal of this thesis is to find a computational system for the control of the emotional content of produced music, so that it expresses a given emotional specification. This system shall be flexible, independent from musical styles and also scalable. The flexibility is grounded on the possibility of controlling emotional content in different levels, like the segmentation, classification, selection and transformation. The scalability of the system allows not only the easy integration of other levels of control like the composition, but it also allows the production of music that, originally, was not part of the system. Produced music is solely instrumental, which is known to be sufficient to express desired emotions (Kimura, 2002). This thesis is focused on tonal music, a type of music characterized by having a note (the tonic) that all other notes gravitated toward.

It is important to mention that due to the multidimensional nature of both emotions and music, many dimensions of these areas are not going to be controlled. This thesis is focused on the music content. For instance, concerning emotions, social variables like context and human listener experience are not controlled; where it concerns to music, editorial, cultural metadata and song lyrics are not analysed.

3. Contributions

There are already some proposed approaches to solve the problem addressed by this thesis. However, none of these approaches gives an entirely satisfactory response to our requirements. We have found especially promising a particular hybrid approach that consists in combining classification/selection with transformation. In fact, the transformation can improve the classification/selection result when there is not a solution in the music base (database of music) close to the emotional specification. On the other hand, as the selection tends to produce an output with characteristics close to the desired ones, the transformation assumes less risks of degrading music quality, because the adjustments needed to get the music characteristics fit the emotional specification are limited.

The solution proposed in this thesis has the advantage of being able to produce outputs of acceptable quality quite independently from the music base: it is able to find the best possible match and then transform it in order to increase the match even further. It is also quite flexible: the music base can be completely redefined to adapt to the specific needs of a given use scenario. The system uses mechanisms (modules) that are independent from the music it is working with, i.e., the musical output corresponds to the emotional specification independently of the original music base. The system is also reliable, thanks to the experimental calibration using different subjects.

We have found other opportunities to contribute to the advance of the state of the art: adopt both the discrete and dimensional representation of emotions; systematization of the relations between emotions and musical features in the knowledge base by studying the musical features with an emotional impact; development of modules to control the emotional content of music; use of techniques of human emotional recognition for validation and calibration of the system. We also tested the usability of a version of EDME system ready to be used in real-time and with an interface that can be used in application domains like entertainment and healthcare.

4. Publications Relevant to this Thesis

This section is devoted to the presentation of all the publications relevant to this thesis. For each publication we enumerated other works where it was cited.

4.1. Journals

1. Oliveira, A., Cardoso, A., 2010. A Musical System for Emotional Expression. In: Knowledge-Based Systems, Elsevier, 23, 901-913.
    a) Wang, H., Lee, Y., Yen, B., Wang, C., Huang, S., Tang, K., 2011. A Physiological Valence/arousal Model from Musical Rhythm to Heart Rhythm. In: IEEE International Symposium on Circuits and Systems, 1013-1016.
    b) Liu, Y. and Liu, M. and Lu, Z., Song, M., 2012. Extracting Knowledge from On-Line Forums for Non-Obstructive Psychological Counseling Q&A System. In: International Journal of Intelligence Science, Scientific Research Publishing, 2(2):40-48.

4.2. Conference Papers

1. Oliveira, A., Cardoso, A., 2007. Towards Affective-Psychophysiological Foundations for Music Production. In: Lecture Notes in Computer Science, Affective Computing and Intelligent Interaction, Springer, 4738, 511-522.
    a) Monteith K., Martinez T., Ventura D., 2010. Computational Modeling of Emotional Content in Music. In: Cognitive Science.
    b) Perry K., Martinez T., Ventura D., 2010. Automatic Generation of Music for Inducing Emotive Response. In: International Conference on Computational Creativity.
    c) Caporusso, N., 2011. The Body and the Mind “through the Lens” of Music: Exploiting Brain-Computer Interfaces and Embodied Music Cognition to Assess Sensorimotor Synchronization and Developmental Disorders. In: Cognitive Practices.

d) Monteith, K., Martinez, T., Ventura, D., 2012. Automatic Generation of Melodic Accompaniments for Lyrics. In: International Conference on Computational Creativity, 87.

2. Oliveira, A., Cardoso, A., 2007. Control of Affective Content in Music Production. In: International Symposium on Performance Science.

   a) Rad, R., Firoozabadi, M., Rezazadeh, I., 2011. Discriminating Affective States in Music Induction Environment Using Forehead Bioelectric Signals. In: 1st Middle East Conference on Biomedical Engineering, 343 - 346.

3. Oliveira, A., Cardoso, A., 2007. A Computer System to Control Affective Content in Music Production. In: Portuguese Conference on Artificial Intelligence.

4. Oliveira, A., Cardoso, A., 2008. Controlling Music Affective Content: A Symbolic Approach. In: Conference on Interdisciplinary Musicology.

5. Oliveira, A., Cardoso, A., 2008. Towards Bi-dimensional Classification of Symbolic Music by Affective Content. In: International Computer Music Conference.

   a) Baldan, S. and Barate, A., Ludovico, L.A., 2012. Automatic Performance of Black and White n. 2: The Influence of Emotions Over Aleatoric Music. In: International Symposium on Computer Music Modeling and Retrieval (CMMR).

6. Oliveira, A., Cardoso, A., 2008. Modeling Affective Content of Music: A Knowledge Base Approach. In: Sound and Music Computing Conference.

   a) Knautz, K., Neal, D., Schmidt, S., Siebenlist, T., Stock, W.G., 2011. Finding Emotional-Laden Resources on the World Wide Web. In: Information, 2(1):217-246.

   b) Wallis, I., Ingalls, T., Campana, E., Goodman, J., 2011. A Rule-Based Generative Music System Controlled by Desired Valence and Arousal. In: Sound and Music Computing.

7. Oliveira, A., Cardoso, A., 2008. Affective-driven Music Production: Selection and Transformation of Music. In: International Conference on Digital Arts - ARTECH.

8. Oliveira, A., Cardoso, A., 2008. Emotionally-controlled Music Synthesis. In: Encontro de Engenharia de Áudio da AES Portugal.

   a) Psarras B., Floros A., Strapatsakis M., 2009. Elevator: Emotional Tracking using Audio/visual Interaction. In: 126th Convention of Audio Engineering Society.

b) Psarras, V., Floros, A., Drosos, K., Strapatsakis, M., 2011. Emotional control and visual representation using advanced audiovisual interaction. In: International Journal of Arts and Technology.

9. Oliveira, A., Cardoso, A., 2009. Automatic Manipulation of Music to Express Desired Emotions. In: Sound and Music Computing Conference.

   a) Livingstone, S., Muhlberger, R., Brown, A., Thompson, W. 2010. Changing musical emotion: A computational rule system for modifying score and performance. In: Computer Music Journal, 34(1):41-64.

   b) Kirke, A. 2011. Application of Intermediate Multi-Agent Systems to Integrated Algorithmic Composition and Expressive Performance of Music. University of Plymouth.

10. Ventura, F., Oliveira, A., Cardoso, A., 2009. An emotion-driven Interactive System. In: Portuguese Conference on Artificial Intelligence.

11. Lopez, A., Oliveira, A., Cardoso, A., 2010. Real-time Emotion-driven Music Engine. In: International Conference on Computational Creativity.

   a) Indurkhya, B., 2012. Whence is Creativity? In: International Conference on Computational Creativity, 62.

5. Thesis Organization

Part I presents the motivation and the aim of the thesis, as well as the contributions and publications that resulted from it.

Part II presents a state of the art of areas related to the work done by reviewing works of Music Psychology, Music Computing and Affective Computing. In the end of each section, we include a summary where we highlight the more relevant works for this thesis.

Part III presents our computational system in seven sections. The first section presents the approach. The second section presents the details of the architecture. The third one describes the experiments conducted in order to improve the quality of the output of the system. The fourth section describes the systematization of the knowledge base. The fifth one presents the evaluation of the classifiers. The sixth one one describes all the stages of the calibration and validation of the system. The seventh and last section of this part presents the application of the system.

Part IV presents a section of discussion where we highlight the main things approached in this thesis. There is also a section describing

future applications of the system.

Part II.

Background

6. Music Psychology

> ”There is geometry in the humming of the strings, there is music in the spacing of the spheres.”
>
> – *Pythagoras.*

Music is used to communicate values, attitudes and self-views (Rentfrow and Gosling, 2003). It is a powerful stimulus capable of influencing our emotions. This has been proved by research findings on music perception and expression (Deutsch, 1982; Lerdahl and Jackendoff, 1983; Narmour, 1990; Temperley, 2004; Widmer and Goebl, 2004), and more recently by studies that have found relations between musical features and emotions (Gabrielsson and Lindstrom, 2001; Juslin, 2001). For instance, tempo is widely accepted as having direct influence on the pleasantness of emotions.

Music Psychology is a field of Psychology that helps us to understand the emotional processes involved in our mind with the help of music (Deutsch, 1982). The communication of emotional content by music can be studied at three different levels: considering the composer's message, the emotional intentions of the performer, and the listener's perceptual experience (Livingstone et al., 2007). There are several research areas contributing to this study. Music Perception and Music Cognition are focused on the listener's perceptual experience, Music Performance is focused of the emotional intentions of the performer and Music Theory is focused on the composer's message. In this chapter, we present a systematic overview of works in Music Psychology. Bearing in mind the focus of this thesis, we highlight in particular those that provide an insight on the relations between emotions and music. We present four sections that explain Music Psychology from four perspectives: perceptive, cognitive, performative and theoretical.

6.1. Music Perception

The major findings on music perception (and music cognition) can be found in (Justus and Bharucha, 2002). Justus and Bharucha divided these findings into five domains from which we highlight three: pitch, time and musical performance. In the pitch domain they reviewed pitch height, pitch class, pitch categorization, relative pitch, absolute pitch, consonance, dissonance, scales and tonal hierarchies of stability, chords and harmonic hierarchies of stability, harmonic perception, harmonic representation, harmonic expectation, melodic perception, melodic representation and melodic expectation. In the time domain they reviewed tempo, rhythmic pattern, grouping, meter, event hierarchies and reduction, and the relationship between time and pitch. In musical performance area they evaluated the interpretation and planning, communication of structure, and musical expertise and skill acquisition. This section (and the following ones) are not going to explore all these areas in detail, instead we will focus on those that we have found more relevant to this thesis. In the next subsection, we are going to put emphasis on four categories of features intervening in music perception: melody, harmony, rhythm and timbre.

6.1.1. Melodic Expectation

"*Affect . . . is aroused when an expectation activated by the musical stimulus, is temporarily inhibited or permanently blocked*" as was said by Meyer (1956). Melody expectation is correlated to feelings of surprise, disappointment, fear and closure. Cross-cultural comparisons suggest that certain psychological principles of expectation are quite general (Krumhansl, 2002). This section gets some insight on this by reviewing important works on this area. Schellenberg et al. (2002) compared the Implication-Realization (I-R) (Narmour, 1990) and 2-factor (Schellenberg, 1997) models of melodic expectation using 3 features: simplicity, scope and selectivity. They tried to examine the change of melodic expectation along the time. The implication-realization model analyses registral direction, intervallic difference, registral return, proximity and closure. On the other hand, 2-factor model analyses pitch proximity and pitch reversal. Narmour's theory has been extended to mathematical models of melodic tension (Margulis (2005) and Larson (2004)).

Larson (2004) developed a theory of musical forces for melodic expectation. He describes two computational models founded on musical forces of gravity, magnetism and inertia. Computer-generated and participant-generated expectations were compared and results showed a positive correlation between them. The Larson's theory of musical forces states that "*we tend to hear music as purposeful action within a dynamic field of musical forces*", making an analogy between physical motion through space and the perceived "*motion*" of a melodic line. The musical forces act continuously on musical lines in a dynamically shifting musical context.

Margulis (2005) designed a hierarchical model to evaluate melodic expectation with four factors: stability, proximity, direction and mobility. This model links expectancy rating to listeners experience of musical tension, as well as theorized expectations and dynamics, affective contours of musical experience. Margulis' models include elements of both Narmour's (1990) and Lerdahl's (1983) models. Tonal pitch space and innate bottom-up processing are given significant status in the model. It describes how expectation connects to the experience of affect and tension. This is done with the help of three tension types: the experience of intensity (surprise-tension); the highlighting of a melody's apparent intentionality (denial-tension); and the impression of desire or forward-directedness in melody (expectancy-tension). For instance, people experience a more positive affect in relation to small deviations from expectedness than they did in relation to large deviations or no deviations.

Melodic expectancy can be understood with cross-cultural and statistical approaches (Eerola, 2003). Eerola studied processes used in structuring, interpreting, remembering and performing music. This work supports the idea that cultural background shapes the influence of these processes during music perception. Melodic expectancies can be of two types: pitch-related or temporal. Short-term auditory priming, auditory stream segregation, sensitivity to frequency of occurrences and rule-based heuristics of melodic continuations are pitch-related processes which are related to musical events stored in sensory memory. On the other hand, there is pitch-related stylistic knowledge that is also important for melody expectation: tonal hierarchy, western schematic expectations, harmony, melody anchoring and melodic archetypes.

6.1.2. Harmonic Tension

Musical tension allows us to gain insight on how music structure translates into emotions (Farbood, 2006). Increasing tension induces a feeling of building excitement or impending climax, or an increase in uncertainty, while decreasing tension induces a feeling of relaxation, resolution, or fulfillment. Tension is central to Western music theory and has been studied by several music theorists and cognitive psychologists.

Farbood (2006) made a quantitative and parametric model of musical tension. This model used six musical parameters: harmony, melodic expectation, pitch height, tempo, onset frequency and dynamics. Melodic expectation, harmony and dynamics were calculated with the help of the models made by, respectively, Margulis (2005), Lerdahl (1983) and Jehan (2005). The validation of the system was done in two experiments to analyse how these features affect subjects' overall perception of tension. Linear and polynomial regression were used in the second experiment. All the features tested alone had an effect on the perception of tension. On the one hand pitch height had the clearest effect, on the other hand onset frequency had the weakest effect. Unlike non-musicians, harmony was more important than pitch height for musicians. Also, changes in onset frequency and tempo have a great influence on musicians.

Steinbeis et al. (2006) studied the role of harmonic expectation in emotional experience. Harmonic expectations were based on relations of harmonic distance. They argued that music tension is related to the experienced emotion and that the expectation of an harmonic event is inversely proportional to the expected tension, overall subjective emotionality and electrodermal activity. This work supports Meyer's (1956) idea that musical emotions arise through the suspension and fulfillment of expectations; harmony expectancy violations were related to the increase of the listener's arousal.

Tonality induction is the process through which the sense of key arises and changes over time. The dynamics of this process was studied by Toiviainen and Krumhansl using two self-organized models (Toiviainen and Krumhansl, 2003). One model is based on pitch class distributions, the other on tone-transition distributions. Principles of auditory scene analysis were used to design a dynamics matrix for the tone-transition model. The dynamic process of tonality induction was associated with musical tension. Tension was measured using key distance and dissonance. The computer model and subjects' responses are available on the web[1].

There is also the schenkerian analysis, which intends to interpret the underlying structure of a tonal work. This is done by studying how harmonic progressions are arranged to accomplish a goal (Schenker, 1973). It influenced recent theoretical developments including the Generative Theory of Tonal Music (Lerdahl and Jackendoff, 1983) (section 6.4).

6.1.3. Rhythmic Perception

Rhythm recognition involves three stages: finding the beat, discovering the rhythmic structure and mapping the note onsets to musical timings (Dixon, 1997). Beat induction is only part of the first of the stages. It is the process in which a regular pattern (the beat) is activated while listening to music. The induced beat carries the perception of tempo and is the basis of temporal coding of temporal patterns (Desain et al., 1999). Dixon described a rhythm recognition process which analyses acoustic data, detecting a sequence of note onsets, and then discovers patterns in the intervals between the onsets.

Desain and Honing worked on the categorization of rhythmic patterns (Desain and Honing, 2003). Continuous time intervals were transformed into rhythmic categories that can be seen in categorization maps (Figure 6.1.1). This was done by partitioning the space of musical performances into a small set of connected rhythmic regions (categories). In Figure 6.1.1 different colors represent different rhythmic categories. Their music notation and integer representation is shown in the legend, which lists them in order of response proportion. Grey lines are category boundaries. Darker shades of color indicate a larger proportion of participants who choose this identification.

[1] http://www.perceptionweb.com/misc/p3312/

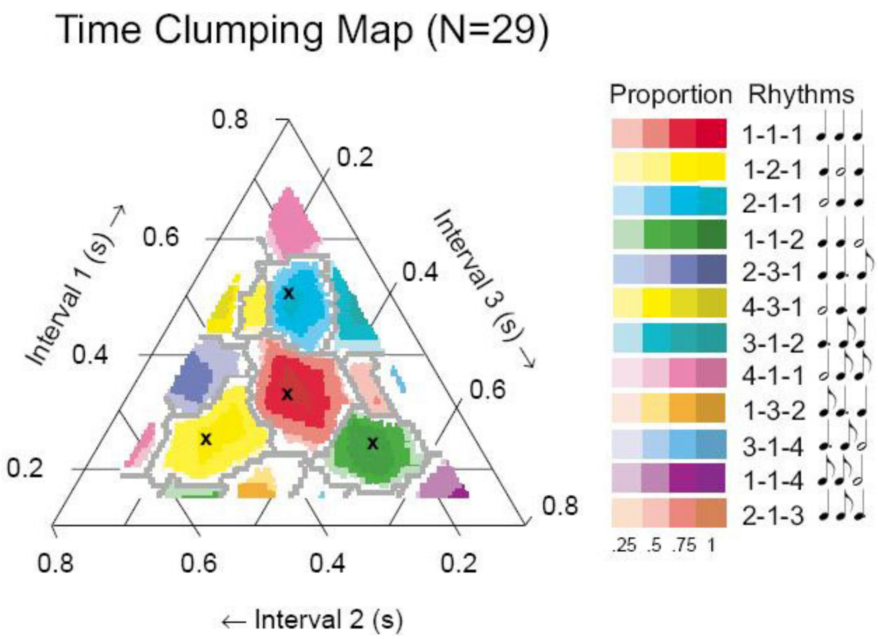


Figure 6.1.1.: Representation of rhythmic categories in a map (Desain and Honing, 2003)

6.1.4. Timbre Perception

Eronen and Klapuri (2000) found a wide set of features to model the temporal and spectral characteristics of instruments. Padova et al. analysed the emotional responses to variations of spectral energy, spectral structure and spectral density (Padova et al., 2005). They found that changes of harmonic dynamic and harmonic ratios induce negative emotions and that spectral energy variations induce high levels of happiness. The repetition of stimuli induces a decrease of intensity in positive emotions and an increase of intensity in negative emotions, fear and surprise. They support that different timbre is associated with different emotions. Piano and hybrid sounds induce negative emotions; flute sound induces other pattern of emotions. The study of timbre made by Lucassen (2006) led him to the conclusions that the piano is emotionally neutral; marimba is joyful; cello invokes sad emotions; and alt saxophone provokes negative and positive emotions.

Deutsch (1982) reviewed timbre perception from three perspectives: identification, fusion and sequencing. Table 6.1 presents a summary of her psychological inferences.

Timbre perception tasks	Psychological findings
Identification of timbre	Differences in timbre of complex tones are related to the strengths of their various harmonics; Simple tones are pleasant, but dull at low frequencies; Tones with strong upper harmonics sound rough and sharp; Complex tones with only odd harmonics sound hollow; Critical band, attack segment and steady state segment (if timbre varies with time) play an important role on timbre perception; Geometric models with at least two dimensions were developed to represent the timbral space, being the first dimension related to the spectral and distribution of sound and the second to temporal features, such as details of the attack.
Spectral fusion and separation	Musical tones of the same source are usually fused together and musical tones of different sources are usually separated to perceive usually distinct sound images; Spectral fusion can be promoted by: onset synchronicity of spectral components; coordinated modulation in a steady state; and harmonicity of the components of a complex spectrum.
Perception of sequences of timbres	The Warren effect says that sounds are organized into separate streams, according to sound type. Due to this it is hard to form temporal relationships across streams.

Table 6.1.: Psychological inferences about timbre perception

6.2. Music Cognition

The process by which the human auditory system organizes sound into perceptually meaningful elements is called Auditory Scene Analysis (ASA). Roughly speaking, we can generalize this process into a set of steps following presented. Cochlea does a spectral analysis, which decomposes the perceived signal into different frequency components. This decomposition is useful for pitch perception of complex tones, sound segregation and sound identification. Temporal patterns of vibration are encoded on the basilar membrane, more properly in auditory nerves. Interaural time differences are used to localize sounds. Most of the works in this area recur to psychoacoustics concepts (Plack, 2004). The following subsections, will be focused on music cognition systems, the study of the personality, models of emotions in music and on emotionally-relevant musical features.

6.2.1. Systems

Computer Auditory Scene Analysis (CASA) systems are machine listening systems that aim to separate mixtures of sound sources in the same way that human listeners do. Scheirer (2000) developed a CASA framework that embeds the most important music perception theories. Psychoacoustic theories of human listening were tested with computer-modeling approaches. Signal-processing techniques were used to extract important musical features from audio music. This model extracts 16 musical features, which are based on loudness, tempo, pitch and ASA. Martin et al. (1998) present the advantages of using a research framework based on a music listening approach, by taking into account the limitations of music content analysis based both on musical signal processing and music theory. They studied various case studies on the extraction of rhythm, timbre and harmony from audio signals.

Jehan (2005) developed a music cognition framework that can also belong to the group of CASA frameworks. It creates music by using audio examples and by applying machine listening and machine learning techniques. He tried to automate the process of listening, composing and performing using a song database. Sounds and structures of music were analysed and musical parameters extracted. These parameters were used in synthesis of musical structures. This thesis contributed to the fields of music analysis and synthesis with a practical implementation grounded on music cognition. In the realm of music synthesis/transformation several algorithms were implemented (see subsection 7.5).

Whitman (2005) presented ways to represent information from the musical signal and context. Whitman's framework represented contextual and extra-signal data in the form of community metadata. He worked with two kinds of musical data to obtain musical meaning. Cepstral modulation extracted musical meaning from audio signal and Natural Language Processing, and Statistics were used to extract meaning from community metadata. The framework gave the following semantics of music information: funky, cool, loud, romantic, etc.

Temperley (2004) explored cognitive processes involved in perception of six kinds of musical structures: metrical, melodic phrase, contrapuntal, tonal-pitch-class, harmonic and key. Metrical structure is a framework of levels of beats. Melodic phrase structure is a segmentation of the input into phrases. Contrapuntal structure is a segmentation of a polyphonic texture into melodic lines. Harmonic structure is a segmentation of a piece into harmonic segments labelled with roots. Pitch spelling involves a labelling of pitch events in a piece with spellings. Key structure is a segmentation of a piece into larger sections labeled with keys. For each of these structures, Temperley developed preference rules. Lerdahl and Jackendoff (1983) were the first to use these types of rules (see section 6.4 for more details). There are some similarities between these two works. Metrical structure is related to the meter model of Lerdahl and Jackendoff; and phrase structure is related to the grouping model of Lerdahl and Jackendoff. The metrical structure uses nine rules (Table 6.2). The phrase structure uses three rules (Table 6.3). The contrapuntal structure uses four rules (Table 6.4). The pitch spelling model uses three rules (Table 6.5). The harmonic model uses four rules (Table 6.6). The key model uses two rules (Table 6.7). Temperley and Sleator (1999) implemented preference rules to generate the harmonic and metrical structures (subsection 7.2.1).

Meter rule	**Description**
Event	prefers a structure that aligns beats with event onsets
Length	prefers a structure that aligns strong beats with onsets of longer events
Regularity	prefers beats at each level to be maximally evenly spaced
Grouping	prefers to locate strong beats near the beginning of groups
Duple bias	prefers duple over triple relationships between levels
Harmony	prefers to align strong beats with changes in harmony
Stress	prefers to align strong beats with onsets of louder events
Linguistic stress	prefers to align strong beats with stressed syllables of text
Parallelism	prefers to assign parallel metrical structures to parallel segments

Table 6.2.: Rules of the meter model

Melodic phrase rule	**Description**
Gap	prefers to locate phrase boundaries at (a) large interonset intervals and (b) large offset-to-onset intervals
Phrase length	prefers phrases to have roughly 8 notes
Metrical parallelism	prefers to begin successive groups at parallel points in the metrical structure

Table 6.3.: Rules of the phrase structure model

Contrapuntal rule	**Description**
Pitch proximity	prefers to avoid large leaps within streams
New stream	prefers to minimize the number of streams
White square	prefers to minimize the number of white squares in streams
Collision	prefers to avoid cases where a single square is included in more than one stream

Table 6.4.: Rules of the contrapuntal model

Pitch spelling rule	Description
Pitch variance	prefers to label nearby events so that they are close together on the line of fifths
Voice-leading	Given two events that are adjacent in time and a half-step apart in pitch height: if the first event is remote from the current center of gravity, it should be spelled so that it is five steps away from the second on the line of fifths
Harmonic feedback	prefers TPC representations which result in good harmonic representations

Table 6.5.: Rules of the pitch spelling model

Harmony rule	Description
Compatibility	prefers roots that result in certain pitch-root relationships
Ornamental dissonance	prefers events that are closely followed by another event a half-step or whole-step away and metrically weak, when labelling events as ornamental
Harmonic variance	prefer roots such that roots of nearby chords spans are close together on the line of fifths
Strong-beat	prefers to start chord spans on strong beats

Table 6.6.: Rules of the harmonic model

Key rule	Description
Key-profile	For each segment, prefer a key which is compatible with the pitches in the segment, according to the (modified) key-profile formula
Modulation	prefers to minimize the number of key changes from one segment to the next

Table 6.7.: Rules of the key model

6.2.2. Personality

Music cognition benefits from the analysis of personality. Music selection/recommendation systems (subsections 7.4 and 8.4.2) also benefit from this analysis as they are commonly grounded on music preferences (Kuo et al., 2005). The influence of the personality on music preferences is now going to be analysed with some detail (Rentfrow and Gosling, 2003). Studies of music preferences were made with over 3500 individuals. Data from these studies reveals a correlation between music genres and four dimensions of music preferences: reflective and complex; intense and rebellious; upbeat and conventional; energetic and rhythmic. Heavy metal fans tend to experience higher resting arousal and arousal levels than country music fans. Preference for highly arousing music (e.g. heavy metal, rock, alternative, rap and dance) appears to be positively related to resting arousal, sensation seeking, and antisocial personality. The attributes of music vary across a wide range of moods, energy levels, complexities and lyrical contents. For example, some genres emphasize negative emotions (e.g., heavy metal), whereas others emphasize positive emotions (e.g., religious); some genres are technically complex (e.g., classical), although others tend to be basic (e.g., country); some genres have relatively few songs with vocals (e.g., jazz), while others only have songs with vocals (e.g., pop).

Music is listened to most often while driving, alone at home, exercising, and hanging out with friends. Even in social gatherings where music is not the primary focus, it is an essential component - imagine, for instance, a party or wedding without music. Individuals may seek out particular styles of music to regulate their emotional states; for example, depressed individuals may choose styles of music that sustain their melancholic mood.

Individuals enjoy listening to changes on a day-to-day basis, perhaps depending on the mood the person is in. Blues, jazz, classical and folk music facilitate introspection and are structurally complex. Rock, alternative and heavy metal are full of energy and emphasize themes of rebellion. Country, soundtrack, religious and pop emphasize positive emotions and are structurally simple. Rap/hip-hop, soul/funk and electronica/dance are lively and emphasize the rhythm.

6.2.3. Emotions Modeling in Music

Several works have been devoted to modeling emotional perception in music (Schubert, 1999; Korhonen, 2004; Mosst, 2006). Some use time series analysis (Schubert, 1999), others use system identification techniques (Korhonen, 2004). Korhonen selected, estimated and validated ARX (Auto-Regression with eXtra inputs) and State-Space models. These models tested the emotional output using 20 subsets of musical features as input. He distinguished between global features and local features. He used dynamics, mean pitch, pitch variation, timbre, harmony, tempo and texture. He used two tools (Marsyas (Tzanetakis and Cook, 2000b) and PsySound (Cabrera, 1999)) to extract features related to the mentioned properties. Mosst (2006) used quantitative techniques. Several individuals made time-varying emotion annotations. He extracted loudness, spectral centroid, onset density, articulation and mode features. Multiple linear regression method was used to relate these features with emotional annotations.

6.2.4. Emotionally-Relevant Musical Features

Musical tension and relaxation are very significant to the expectations of the sounds played (Krumhansl, 2002). Listeners' tension ratings coincide with the phrase structure. The work of Krumhansl helped us to establish various types of relations between emotions and musical features (Table 6.8); emotions and psychophysiological responses (Table 6.9); and concerns related to music and emotions (Table 6.10).

Emotion	Tempo	Harmony	Ranges of Pitch	Ranges of Dynamics	Rhythms
Sadness	Slow	Minor	Constant	Constant	-
Fear	Rapid	Dissonant	Large	Large	-
Happiness	Rapid	Major	Constant	Constant	Dancelike

Table 6.8.: Relations between emotions and musical features

Emotion	Heart rate	Blood pressure	Skin conductance	Temperature	Respiration	Rate of blood flow	Amplitude of blood flow
Sadness	Change	Change	Change	Change	Normal	Normal	Normal
Fear	Normal	Normal	Normal	Normal	Normal	Change	Change
Happiness	Normal	Normal	Normal	Normal	Change	Normal	Normal

Table 6.9.: Relations between emotions and psychophysiological responses

Musical concerns	Other concerns
Global aspects of musical structure	Overall mood of the music
Tension	Mostly fear, but also happiness and sadness
Tension	Heart rate, blood pressure, pitch height of the melody, density of notes, dissonance, dynamics and key changes
Tension	Musical form (Lerdahl's tree model chromatic tones, interruption of harmonic processes, denial of stylistic expectations)
Emotional expression in music	Emotional expression in dance and speech
Pattern of temporal organization in music	Patterns of intonational units in discourse

Table 6.10.: Relations between concerns related to music and emotions

There have been various studies about the relations between emotional states and musical features (Gabrielsson and Lindstrom, 2001; Berg and Wingstedt, 2005; Webster and Weir, 2005; Collier and Hubbard, 2001; Ilie and Thompson, 2006). A summary of these relations is presented in the Table 6.11. Tempo is more important than mode to make emotional judgments in music (Dalla Bella et al., 2001; Gagnon and Peretz, 2003)[2].

For an extensive review of works that studied emotionally-relevant musical features we recommend Schubert's work (Schubert, 1999). He divided his review by using seven

[2]http://www.brams.umontreal.ca/plab/research/Stimuli/Dalla%20Bella%20et%20al%20(2001)/dallabella_2001_stimulis.html

Emotion	**Articulation**	**Harmony**	**Loudness**	**Melodic range**	**Melodic direction**	**Mode**	**Pitch level**	**Rhythm**	**Tempo**	**Timbre**	**Notes duration**	**Melodic texture**
Sadness	Legato	Complex and dissonant	Low	Narrow	Falling	Minor	Low	Firm	Slow	Few harmonics, soft, dark	Long	Thick harmonized
Happiness	Staccato	Simple and consonant	High	Wide	Rising	Major	High	Regular / Smooth	Fast	Few harmonics,bright	Short	Simple
Grace	-	-	-	-	-	Major	High	-	-	-	-	-
Serenity	-	-	-	-	-	Major	High	-	-	-	-	-
Solemnity	Legato	-	Low	-	-	Major	Low	-	-	-	-	-
Tension	-	-	High	-	-	Minor	-	-	Fast	-	-	-
Disgust	-	-	-	-	-	Minor	-	-	-	-	-	-
Anger	Staccato	-	High	-	-	Minor	High	-	-	-	-	-
Fear	Staccato	-	Low	-	-	-	High	-	-	-	-	-
Tenderness	Legato	-	Low	-	-	-	-	-	-	-	-	-
Surprise	-	-	-	-	-	-	High	-	-	-	-	-
Excitement	-	-	-	-	-	-	High	-	-	-	-	-
Boredom	-	-	-	-	-	-	Low	-	-	-	-	-
Pleasantness	-	-	-	-	-	-	Low	-	-	-	-	-

Table 6.11.: Relations between emotional states and musical features

types of musical stimuli: isolated non-musical sounds, isolated musical sounds, especially composed melodies, pre-existing melodies, especially composed pieces, pre-existing pieces with modification and pre-existing pieces. We summarize his findings in Table 6.12.

Musical feature	High valence	Low valence	High arousal	Low arousal
Loudness	-	-	High	Low
Average Pitch	High	Low	High	Low
Pitch range	-	-	High	Low
Pitch variation	High	Low	High	Low
Melodic contour variation	Rising	Falling	Rising	Falling
Register	High	Low	-	-
Mode	Major	Minor	-	-
Timbre	Piano, strings, few harmonics, bright, soft	Brass, low register instruments, timpani, harsh, violin, woodwind, voice	Brass, low register instruments, timpani, harsh, violin, bright, strings	Woodwind, voice, few harmonics, soft
Harmony	Consonant	Dissonant, Melodic or harmonic sequence, melodic appoggiatura	Complex, dissonant, diminished seventh chord	-
Tempo	-	-	fast	slow
Articulation	staccato	legato	non-legato with sharp contrasts between long and short notes, staccato	legato
Note onset	-	-	rapid onset	slow onset
Vibrato	intense	deep	fast	deep and intense
Rhythm	rhythmic activity, smooth, flowing motion	rough	sophisticated, rough, rhythmic activity, smooth, flowing motion	-
Meter	-	-	triple	duple

Table 6.12.: Relations between emotional dimensions and musical features

6.3. Music Performance

The contribution of the performer to expression communication has two facets: to clarify the composer's message by enlightening the musical structure and to add his personal interpretation of the piece. A mechanical performance of a score is perceived as lacking of musical meaning and considered dull and inexpressive as a text read without prosodic inflexion. Indeed, human performers never respect tempo, timing and loudness notation in a mechanical way when they play a score: some deviations are always introduced, even if the performer explicitly wants to play mechanically. Thus, in general, expressiveness refers both to the means used by the performer to convey the composer's message and to his own contribution to enrich the musical message. Next paragraphs are dedicated to the presentation of models and theories used for expressive musical performances.

There are several models of music performance. These models specify the physical parameters defining a performance. Widmer and Goebl (2004) reviewed four of these models: the KTH rule-based model; the structure-level models of timing and dynamics made by Todd; the mathematical model of musical structure and expression by Mazzola; and a model, induced with machine learning methods, which combines note-level rules with structure-level expressive patterns. They studied the role, principles and assumptions used to change emotional expression and concluded that the models are complementary. This work also presents empirical evaluations of the models. Friberg et al. (2006) presented in more detail the KTH rule system. This system has rules that relate musical performance features and emotional expression (Figure 6.3.1). These rules transform features like sound level, notes duration and phrasing level. Bresin and Friberg (2000) used a program based on the KTH rule system. This program models performance parameters like phrasing, micro-level timing, metrical patterns and grooves, articulation, tonal tension, intonation, ensemble timing and performance noise.

	Happy	Sad	Angry	Tender
Overall changes				
Tempo	somewhat fast	slow	fast	slow
Sound level	medium	low	high	low
Articulation	staccato	legato	somewhat staccato	legato
Rules				
Phrase arch	small	large	negative	small
Final ritardando	small	-	-	small
Punctuation	large	small	medium	small
Duration contrast	large	negative	large	-

Figure 6.3.1.: KTH rules used to relate emotions with performance features (figure taken from (Friberg et al., 2006))

Taylor et al. (2005) designed a virtual character to respond in real-time to the musical input. Appropriate behaviours were defined in the character to reflect the perception of the musical input. These characters were developed through a 3-layer framework. The first layer (perception) was responsible for the extraction of musical features (e.g., pitch, amplitude, tone and chord) from musical input. The second layer (cognition) used the major findings of music perception and cognition (e.g., harmonic structural rules), and Gestalt theory to organize these features. The third layer (expression) was responsible for the character animation using musical data obtained from the previous layers.

The understanding of the communication of emotions in music performances can be done through the application of several algorithms and theories. Most works established rules for controlling musical expressivity by changing musical features (e.g., pitch, intensity, articulation and tempo). These features are usually mapped to emotions (e.g., anger, sadness and happiness). Friberg (2004) used fuzzy sets to make this mapping. Cognitive and cultural factors help in the analysis of expressive intentions in musical improvisation (Baraldi, 2003). We highlight the action-perception theory (Vickhoff and Malmgren, 2004). This theory is based on three constructs: present moment perception, implicit knowledge and imitation. It also considers that there is a 2-way connection between emotions and movements. Empathy is important to understand feelings of other people. These feelings can be categorized into three groups: categorical (happiness, fear, etc.), vitality (crescendo, pulsing and other kinetics terms) and relational (being loved, esteemed, etc.). There are three empathy catalysts: similarity, familiarity and cue salience. Entrainment is another important concept to emotional contagion.

Kimura (2002) used instrumental pieces of music to induce seven emotions: fear, sadness, anger, tenderness, happiness, frustration and surprise. Violinists' expression of sadness, tenderness and happiness were perceived by the listeners with more than 70% of success rate. This work is grounded on Juslin's (2001) study (Figure 6.3.2).

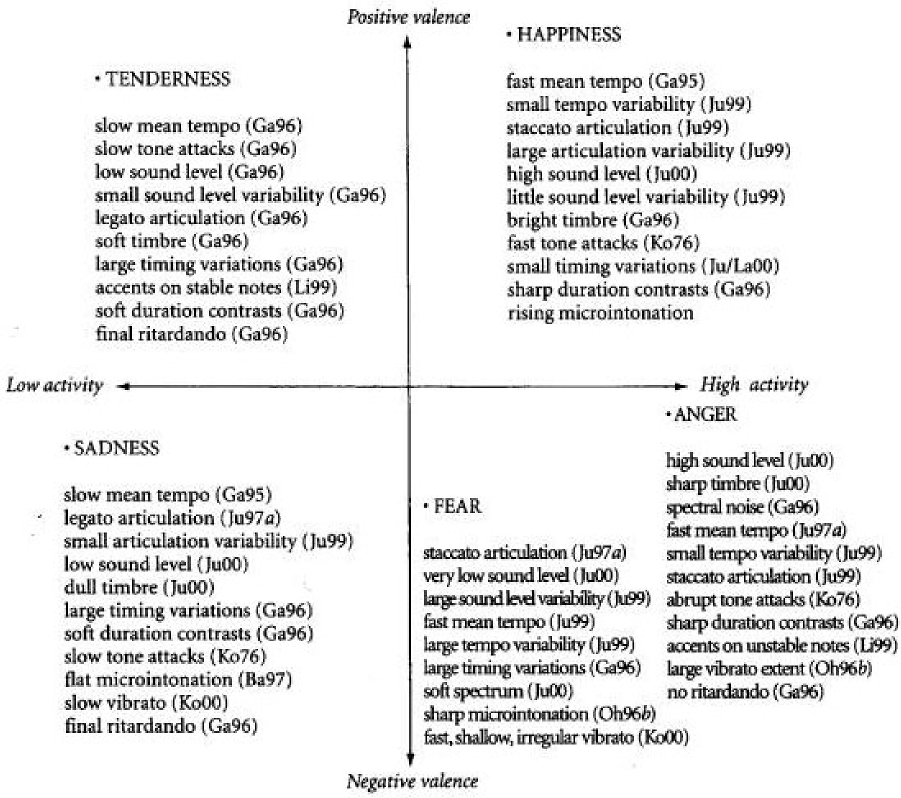


Figure 6.3.2.: Representation of musical features on a 2 dimensional emotion space (Juslin, 2001)

Deutsch (1982) studied the importance of the physical space in the performance of music and drawn several conclusions. The perception of melody is more influenced by the pitch, timbre and loudness than by the localization of instruments. Temporal relationships between tones, from different spatial locations, are also a source of influence. The recognition of individuals tones in a sequence is affected by the pitch proximity. The connectedness of a sequence of tones is affected by factors like pitch relationship, tempo, attentional set and the sequence length. Melodic progression should be by steps instead of skips, according to the law of stepwise progression. . Similar sounds in the frequency spectrum are likely to emanate from the same source and dissimilar sounds in the frequency spectrum from different sources. Transposed melodies retain their essential form.

Finally, it is worth mentioning that there is an annual international competition where it is possible to present the computer systems developed for generating expressive musical performances (Hashida et al., 2008).

6.4. Music Theory

Music Theory and Psychology are two interconnected areas (Deutsch, 1982). The Gestalt theory was the outcome of concrete investigations in psychology, logic and epistemology that lead to establish four key principles: emergence, reification, multistability and invariance (Ellis, 1999). These principles were used by Meyer (1956) in the study of the meaning of emotions in music. He studied the characteristics that affect the continuity (Table 6.13), completeness and closure (Table 6.14) in melody, rhythm, meter and harmony. Finally, the role of the music structure and shape were also objects of analysis from which a few conclusions were drawn. A weak/bad shape can be characterized by their excessive similarity and this leads to tension. Pitch uniformity is characterized by equidistant series of tones. Harmony uniformity is characterized by equal vertical intervals or unchanging harmony or repetitive progressions. Expectation (subsection 6.1.1), expressive variations in pitch, tempo, rhythm, ornamentation and tonality were also identified as important characteristics for understanding the emotional meaning of music.

Musical Law	**Music characteristics**
Melodic continuity	delay; acceleration; contrast of parts; ornaments; shape expectation of harmonic structures
Rhythmic continuity	pulse; meter; rhythm; accent; hierarchical organization (iamb, anapest, trochee, dactyl, amphibrach); rhythmic reversals
Metric continuity	hierarchical organization; time signature; metric changes (hemiola); polymeters

Table 6.13.: Meyer's laws of music continuity

Musical Law	Music characteristics
Melodic completeness and closure	tonality; instrument tessitura; higher-level analysis (Schenker analysis); relaxation of closure linked to lower pitches
Rhythmic completeness and closure	string of accented/unaccented
Harmonic completeness and closure	tonic/key

Table 6.14.: Meyer's laws of music completeness and closure

Deutsch also concluded that interval class can be perceived in a successive context, as an example of top-down shape analysis by the listener. This conclusion is grounded on concepts from musical shape analysis and from the theory of twelve-tone composition. She argued that music is stored in a hierarchical structure. This principle was applied to Schenker's (1973) 3-level system in which notes at one level are prolonged by a sequence of notes at the next-lower level. This system is explained with the tree-based approach of the Generative Theory of Tonal Music (Lerdahl and Jackendoff, 1983). Lerdahl and Jackendoff found that the fundamental relationship expressed in the tree is the elaboration of a single pitch event by a sequence of pitch events. This theory proposes a generative grammar for homophonic tonal music. It characterizes the way listeners perceive hierarchical structures in tonal music. This grammar models musical intuition and takes the form of rules that assign structures that listeners perceive while listening to music:

- grouping structure - segmentation of music into motives, phrases and sections;
- metrical structure - hierarchy of alternating strong and weak beats;
- time-span reduction - hierarchy of structural importance of pitches with respect to their position in the grouping and metrical structures;
- prolongation reduction - hierarchy that expresses harmonic and melodic tension and relaxation.

For each of these structures, they developed a series of well-formedness rules. They analysed the syntax of music using hierarchical trees of relaxation/tension. The form of the resulting trees (right-branching and left-branching) give us indications about the tension/relaxation character of the analysed music.

Tillmann et al. (2000) proposed a self-organizing neural network to embed the knowledge of western musical grammar, e.g., pitch dimension regularities. The process of learning used in this system intended to internalize the correlational structure of tonal music. This work gave rise to empirical findings on processing of tone, chord, key relationship, relatedness judgments, memory judgments and expectancies.

Cambouropoulos (1998) also proposed a computational theory of musical structure. Several modules grounded on cognitive and logical principles like similarity and categorization were developed. They contribute to form a structural description of a musical surface. A comparison with other theories and computational models of music is presented in (Cambouropoulos, 1998).

Models of semiotics and pragmatics can be used to analyse film music (Chattah, 2006). Chattah took into consideration formal design, melodic contour, pitch content, harmonic gestures, cadential formulas and other structural aspects of music. Aspects of film were metaphorically related to aspects of music: motion in vertical space, weight and size with fluctuation in pitch frequency; speed of physical movement with speed of musical events; psychological tension with volume; psychological state with instrumental timbre; and psychological/physical state with harmonic consonance. On the one hand, leitmotifs and topics (symbols), and music and sound parameters (icons) were studied using semiotic constructs; on the other hand, qualitative and structural aspects of music, as well as similarities and dissimilarities between film narrative and music were studied from a pragmatic perspective.

6.5. Summary

This chapter reviewed aspects of music from different psychological perspectives. We presented works of music perception and cognition that explained some processing mechanisms of the listener. The section of music perception was dedicated to the presentation of several models and theories about the perception of different musical features: melody (Eerola, 2003), harmony (Toiviainen and Krumhansl, 2003), rhythm (Desain and Honing, 2003) and timbre (Padova et al., 2003). The section of music cognition presented music cognition systems (Temperley, 2004; Jehan, 2005; Whitman, 2005) and one study about the influence of the personality (Rentfrow and Gosling, 2003). The section ended with the description of models of emotions in music (Schubert, 1999; Mosst, 2006; Korhonen, 2004) and with the presentation of emotionally-relevant musical features (Krumhansl, 2002; Gabrielsson and Lindstrom, 2001; Dalla Bella et al., 2001; Ilie and Thompson, 2006).

Then, we studied music expression with the presentation of some models (Widmer and Goebl, 2004) and theories (Vickhoff and Malmgren, 2004) for music performance. Finally, we entered into the theoretical domain and presented models for tonal music (Deutsch, 1982; Lerdahl and Jackendoff, 1983; Cambouropoulos, 1998; Tillmann et al., 2000).

7. Music Computing

”A good composer does not imitate, he steals.”
– *Igor Stravinsky*.

”Music Computing research can be traced back to the 1950’s, when a handful of composers, together with engineers and scientists, began exploring the use of the new digital technologies for the creation of new music and multimedia content. (...) Today, Music Computing is Europe’s most advanced multidisciplinary approach to music and multimedia. By combining scientific, technological and artistic methodologies it aims at understanding, modeling and producing music using computational approaches .” (Serra et al., 2007)

The research in Music Computing can be classified according to two imaginary axes: music representation and type of the problem. Music can be represented in audio or MIDI format (Moog, 1986), see section 3.1 of (McKay, 2004) for more details about this last format. In the audio domain one uses techniques of signal processing, whereas in the MIDI domain techniques of symbolic processing are the most appropriate. Roughly, there are three main types of problems: analysis (decomposition into simpler elements), synthesis (composition of complex elements by using simple elements) and transformation (recomposition of simple elements). Now, we present examples for each of these problems. For analysis, in the audio domain we can extract features (tonality, tempo, etc.) and also identify parts (melody and rhythm); in the MIDI domain we can analyse harmony and rhythm. For synthesis, in the audio domain we can synthesize and sequence audio; in the MIDI domain, we can do automatic composition and arranging, and symbolic sequencing. For transformation, in the audio domain we can change pitch, change loudness and apply effects; in the MIDI domain we can change pitch, rhythm and tonality, for example.

In this thesis, we work in the MIDI domain, exception made to the synthesis, where the timbre of instruments is relevant, which takes us to perform analysis also at the audio level. For MIDI analysis, we make segmentation, extraction of features, as well as classification and selection. For MIDI transformation we change features like the rhythm and harmony. For audio analysis, we make feature extraction. We also work on the problem of synthesis, particularly in MIDI sequencing and sound synthesis. The choice of using MIDI in most of the tasks is because it is much more adequate than audio if one wants to extract high-level features. This is a very important advantage, as it is easier to bridge the semantic gap between music and emotions when we are using high-level features obtained from MIDI, instead of low-level features obtained from audio recurring to techniques like signal processing. See section 1.4 of (McKay, 2004) for more details about the reasons behind using MIDI instead of audio.

The state of the art in this chapter focuses on the above areas of music computing, i.e., those that are approached in this thesis. Each contribute in some way to support some of the modules of our system presented in the next part. The state of the art focuses mostly on the reviewing of techniques and tools available in these areas. In Chapter 11, when describing the architecture of our system, we will clarify which of these tools and techniques are being used, and in which concrete context.

7.1. MIDI Segmentation

The segmentation of the auditory stream into smaller units, melodic phrases, motifs, i.e., repeated patterns that are structures easily perceived by listeners, is a fundamental process in music perception, music cognition and music theory as was presented in previous chapter. The phrase structure of Temperley (2004), implemented in Melisma Music analyser[3], and the grouping structure of Lerdahl and Jackendoff (1983) are just two of the models most important to the process of segmentation.

There are different approaches available to find repeated patterns in MIDI representations. Lartillot (2005) identified structures based solely on pattern repetitions. He used global selective mechanisms, based on pattern frequency and length to filter combinatorial redundancy. Grilo (2002) used two evolutionary algorithms: genetic programming and genetic algorithms. The objective of this work was to find a segmentation of a musical piece that allowed the identification of the most meaningful patterns that existed in that piece. Paulus and Klapuri (2006) developed a system for finding structural descriptions. The structure of a musical piece was depicted with segments having a description. This system used an algorithm to find the optimal description with regard to a cost function.

There is software developed for the segmentation of MIDI music. MIDI toolbox (Eerola and Toiviainen, 2004) do this with two different approaches: probabilistic and gestaltic. The probabilistic approach analyses melodies. This analysis consists in defining probabilities of phrase boundaries derived from specific distribution of features at the segment boundaries of music collections. The gestaltic approach finds plausible points of

[3] http://www.link.cs.cmu.edu/music-analysis/

segmentation that depend on large changes of pitch intervals, inter-onset intervals and silence.

7.2. Feature Extraction

Feature extraction consists in transforming the input data into a set of features, in order to reduce the dimensionality of the data we work with. We dedicate this section to the presentation of tools and algorithms useful in the process of extracting features from audio and MIDI music. Before entering into details, we highlight the importance of developing taxonomies for the musical features, in order to systematize the features being used in the extraction process. Lesaffre et al. (2003) present a user-dependent taxonomy with five categories for audio music: melody, harmony, rhythm, timbre and dynamics. These categories were analysed in two levels: structural and conceptual. Typke et al. (2004) present an overview of Music Information Retrieval systems by comparing, among other things, the features extracted from MIDI and audio music: pitch, note duration, timbre, rhythm, contour, intervals and others.

7.2.1. MIDI

There are systems that work with MIDI data and that provide features that can be used, for instance, to classify music. JSymbolic (McKay and Fujinaga, 2006) is a free software package that extracts features of instrumentation, musical texture, rhythm, dynamics, pitch statistics and melody. Table 7.1 presents some of the available features. A detailed description of all the features is provided in (McKay, 2004).

Instrumentation	Musical texture	Rhythm	Dynamics	Pitch Statistics	Melody
Pitched instruments	Maximum number of independent voices	Strongest rhythmic pulse	Overall Dynamic Range	Most Common Pitch Prevalence	Melodic Interval Histogram
Unpitched instruments	Average number of independent voices	Rhythmic Variability	Variation of Dynamics	Most Common Pitch Class Prevalence	Average Melodic Interval
Note prevalence of pitched instruments	Variability of number of independent voices	Harmonicity of two strongest rhythmic pulses	Variation of Dynamics In Each Voice	Relative Strength of Top Pitches	Most Common Melodic Interval
Note prevalence of unpitched instruments	Voice equality - number of notes	Strength of Strongest Rhythmic Pulse	Average Note To Note Dynamics Change		
Time prevalence of pitched instruments	Voice overlap	Polyrhythms			
Percussion prevalence	Voice equality - dynamics	Note Density			

Table 7.1.: Summary of McKay's (2004) features

Eerola and Toiviainen (2004) developed a toolbox that finds the following features: melodic contour, similarity, key, meter and segments. Besides these, it calculates twelve melodic features: melodic accent, melodic attraction, melodiousness, melodic range, expectancy-based model, implication-realization principles (Narmour, 1990), melodic tessitura, melodic distance, melodic mobility, melodic measure, accent synchrony and melodic contour; nine rhythmic features: concurrent onsets, duration accents of events, tempo, meter, metrical hierarchy, note density, variability of events, onset autocorrelation and onset distribution; and four harmonic features: key mode, pitch distribution visualization, correlation of the pitch distribution with K&K profiles and tonality (major/minor).

Temperley and Sleator (1999) presented a computational rule-based system to model meter and harmony. This system uses a list of notes with pitch, on-time and off-time as input. Melisma Music analyser[4], the name of the system, covers several aspects of music structure (as presented in section 6.2).

[4]http://www.link.cs.cmu.edu/music-analysis/

There is also JMusic (Sorensen and Brown, 2000) which consists of a music data structure adequate for the extraction of several features. Climax position, rhythmic variety, rhythmic range, note density, pitch variety and pitch range are just some of the available features.

7.2.2. Audio

There are some systems that meet the needs of researchers by providing a library of analysis algorithms on the audio domain that are suitable for a wide array of tasks. JAudio (McEnnis et al., 2005) is one of these systems which allow the extraction of features like spectral centroid, RMS, power spectrum, zero crossings, strongest beat, MFCC, LPC, moments, peak finder and harmonic spectral centroid. Marsyas (Tzanetakis and Cook, 2000b) is another system used for prototyping and experimentation with computer audition applications. It uses four features extractors: Short Time Fourier Transform, Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Crest Factor and Spectral Flatness Measure of MPEG-7 (Allamanche et al., 2001). It is composed by many audio information retrieval tools (Tzanetakis and Cook, 2000a). Pitch, harmonicity, MFCCs, LPC, and the centroid, flux and moments of spectrum are some of the features that can be extracted. MIR Toolbox (Lartillot and Toiviainen, 2007) is a framework that includes most of the features available in both JAudio and Marsyas, plus lower and higher-level features related to timbre, tonality, rhythm and form. It also allows statistical analysis, segmentation and clustering of music.

There are also tools that are focused on the extraction of perception features. PsySound (Cabrera, 1999) extracts psychoacoustic features. It comes with several models that obtain psychoacoustic measures: level, spectrum, cross-channel, loudness, dissonance and pitch. IPEM Matlab toolbox (Leman et al., 2001) models the human auditory system. It allows the analysis of music in three different levels: sensorial, perceptual and cognitive. Each of these levels has its own modules. The sensorial level has roughness and onset modules. The perceptual level has pitch completion, rhythm and echoic memory modules. The cognitive level has a contextuality module.

7.3. Classification

Music genre classification is the most common task in music classification[5] (mood classification will be presented in detail in subsection 8.4.2). Scaringella et al. (2006) made a survey of systems used in music classification by genre and identified the most common features: melody, harmony, rhythm and timbre. There are three different

[5] http://www.music-ir.org/mirex/wiki/2011:MIREX_Home

approaches to classify music: expert systems, unsupervised classification, and supervised classification. Next subsections present works about music genre classification on the MIDI and audio domains.

7.3.1. MIDI

McKay (2004) made a system of music genre classification of MIDI data. They used a library of features available in the JSymbolic, which is described in subsection 7.2.1. He made use of hierarchical classification, flat leaf category classification and round robin classification.

7.3.2. Audio

Tzanetakis and Cook (2002) and McKinney and Breebaart (2003) worked on the music genre classification on the audio domain. Tzanetakis and Cook used three feature sets to do music classification by genre: timbral texture, rhythmic content and pitch content. The importance of these features was analysed using audio collections to train statistical pattern recognition classifiers. To represent timbral texture the following features were used: spectral centroid, spectral rolloff point, spectral flux, time domain zero crossings, MFCCs, analysis and texture window and low-energy feature. To represent rhythm content a Wavelet transform was used to extract the following features (from the beat histogram): strength of the main (and second) beat, regularity of the rhythm, relation of the main beat to the subbeats, relative strength of the subbeats to the main beat, period of the first and second peak in beats per minute and overall sum of the histogram. To represent pitch content the signal was decomposed into two frequency bands (below and above 1000Hz) to build a pitch histogram. From this histogram the following features were calculated: most dominant pitch class, its octave range, main pitch class, main tonal interval relation, overall sum of the histogram.

McKinney and Breebaart used four feature sets for audio classification: low-level signal parameters, 13 MFCCs, psychoacoustic features and an auditory filterbank temporal envelope. The low-level signal parameters are based on the subsequent properties: root-mean-square level, spectral centroid, bandwidth, zero-crossing rate, spectral roll-off frequency, band energy ratio, delta spectrum magnitude, pitch and pitch strength. Three psychoacoustic features were analysed: roughness (musical dissonance), loudness (signal strength) and sharpness (spectral density and strength of high-frequency energy). Temporal modulations of features were the most important for the classification of audio and music.

7.4. Audio Selection/Recommendation

Music selection can be divided into two categories: query systems and recommendation systems (Pachet et al., 2000). The works presented in this section belong to the second category. In this category there is also the distinction between content-based and collaborative filtering (Kuo et al., 2005). The works of this section analyse the content of music that users liked in the past and recommends the music with relevant content. Corthaut et al.(2006) developed a music player that selects appropriate musical content to specific musical contexts. Musical characteristics are manually annotated by music experts. This system extracts music metadata. There are similar systems: MusicLens[6] is a music recommendation system based on genre, volume, tempo, voice, orchestra/solo, listening purpose, gender, mood, color and composition year; MoodLogic[7] is a music recommendation system based on genre, type recording, voice, sound quality, similar artists, energy, energy level, heat, mood, tempo, danceable, melody memorability, lyrics language, lyrics topic, instruments and composition year; Sony StreamMan[8] is a mobile streaming music service based on genre, mood, atmosphere, decade and rating; MusicIP mixer[9] is a tool that does acoustic fingerprinting on music libraries and generates playlists for specific moods. It is also relevant to mention two approaches for music selection. Weiss's (2000) approach combines popularity, catalogue coverage, style continuity and multi-user dimensions. Pachet (2000) uses a combinatorial approach based on constraint satisfaction programming. It was based on the desire of repetition, desire of surprise and exploitation of catalogues. See subsection 8.4.2 for a couple of studies on emotion-driven selection.

7.5. Transformation

This section is devoted to the presentation of works about transformation on the MIDI and audio domains.

7.5.1. MIDI

For the MIDI domain we have JMusic (Sorensen and Brown, 2000), a tool adequate for non real time music composition, but also for music transformation. It has many algorithms that can be used to modify MIDI music at different levels (notes, phrases, parts or score). The first beat of each bar/measure can be changed by increasing the dynamic of notes; notes durations and rhythm can be changed; it is possible to append notes, phrases, parts and scores; notes pan value can be alternated; crescendos,

[6] http://www.musiclens.de/contest/
[7] http://www.moodlogic.com/
[8] http://www.streamman.net/evo/web/stream/257_EN
http://tvnomics.typepad.com/Rodriguezfinal.pdf
[9] http://www.musicip.com/mixer

decrescendos and diminuendo can be applied to phrases; phrases and parts can be looped.

7.5.2. Audio

The audio domain is fruitful in works that transformed different musical aspects. It is possible to make harmonic transformations such as modulation, reduction and harmonization; melodic transformation such as transposition (or pitch shifting), various symmetries and ornamentation/reduction; rhythmic transformations such as time compression and dilatation (time stretching), various symmetries and accent and silence changes; and dynamics and timbre transformations (Amatriain et al., 2003). Pitch shifting is an effect that aims at transposing the original pitch of a sound, time-scaling consists in changing the length of the sound. However, this is not all. Jehan (2005) applied several transformation algorithms on his system based on the model analysis/resynthesis. The beat matching, music morphing, music cross-synthesis, music texture and music restoration are just some of the transformations. Beat matching technique intended to select songs with similar tempos and align their beat over the course of a transition while cross-fading their volumes. Music cross-synthesis/mosaicing was a technique used for sound production, whereby one parameter of a synthesis model is applied in conjunction with a different parameter of another synthesis model. Music texture and music restoration were two additional types of techniques that could be used to transform music. Music texture (Figure 7.5.1) sequenced different segments of the original piece of music to produce a longer piece of music. Music restoration (Figure 7.5.2) used segments of different parts of the original song to recover the part of the music that was corrupted.

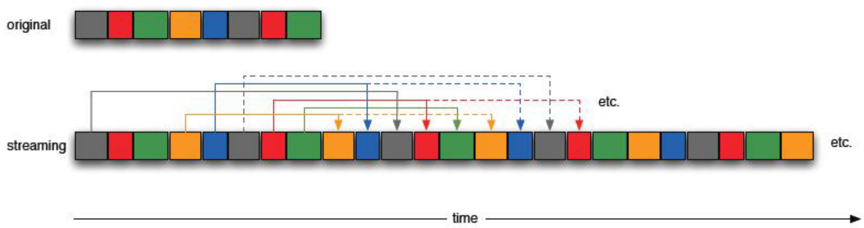


Figure 7.5.1.: Music textures (Jehan, 2005)

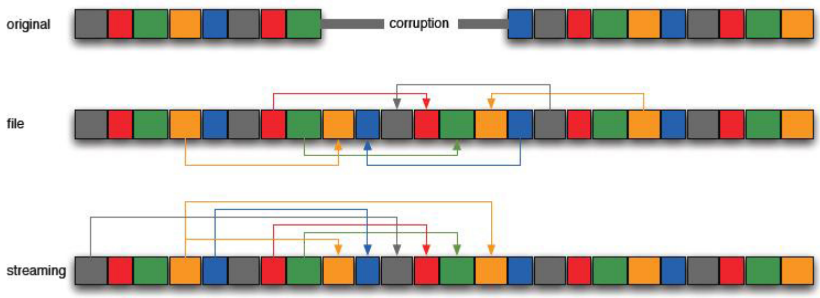


Figure 7.5.2.: Music restoration (Jehan, 2005)

Grachten (2006) worked on tempo transformations of monophonic audio. These transformations preserved the quality of the musical performance and obtained audio more naturally than the one obtained by uniform time stretching. Grachten used a case based reasoning system that did the audio analysis/synthesis and the manipulation of the input audio recording. The manipulation part first received a MIDI with the melody, a melodic description and a target tempo. The problem was defined from these data. The CBR part was used to select and reuse the case more appropriate to the problem. Fabiani and Friberg (2007) extend this work by allowing the transformation of sound level and tone duration besides the tempo transformation.

Gomez et al. (2003) developed a system for melodic transformation. This was done with the help of high-level melodic descriptions.

7.6. Audio Sequencing

Sequencing music includes the ordering of tracks by musical features, namely tempo (Cliff, 2000). The need of crossfading involves the synchronization in the pitch, tempo, and phase of the two sequenced tracks. Figure 7.6.1 illustrates this process between an outgoing track A and an incoming track B. As track B has a faster tempo it is being time-stretched to match tempos of both tracks. The sequence of tracks can also be specified with the help of trajectories of musical features.

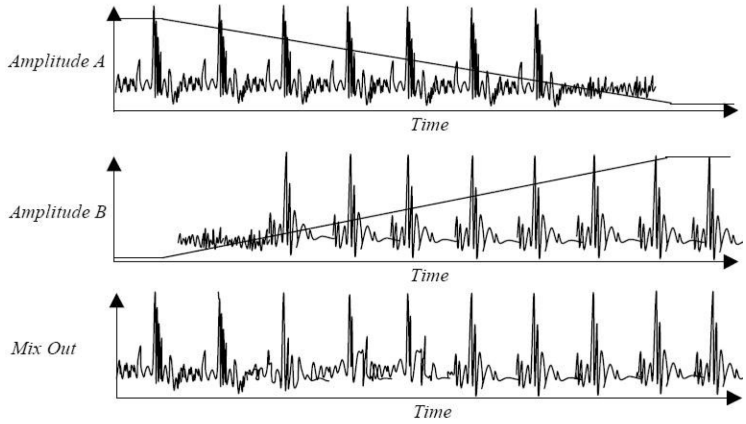


Figure 7.6.1.: Cross fading in a sequencing process(Cliff, 2000)

DJs use some techniques in the process of music remixing to manipulate sound. Time warping algorithms use time stretching and shrinking techniques to change audio duration and tempo matching. Pitch shifting is used for samples matching. There are also other techniques to concatenate music samples based on loudness (amplitude envelope, tremolo), spectral shape (spectral motion) and pitch (vibrato and pitch bends) similarity. From psychological studies (Meyer, 1956) we also know that music continuity is influenced by some musical features: melody, rhythm and metric.

Beat-matching is a widely used technique during music sequencing. In this technique there is a period of cross-fading where we need to adjust the tempo of a song with the tempo of other song. It is possible to do this with different approaches: linear, exponential, Stype, and constant-power (Jehan, 2005). Some techniques used in concatenative synthesis (Schwarz, 2004) can be helpful to automatic sequencing. Musical mosaicing (Zils and Pachet, 2001) was presented as a sequence generation mechanism. With this technique we can generate sequences of sound samples by specifying high-level properties of the music sequence we want to obtain.

7.7. Audio Synthesis

"Musical sound synthesis allows the creation of new sounds, either from scratch, or by changing an existing sound (this is usually called resynthesis). In both cases, the parameters of the synthesis model used have to be specified. In synthesis from scratch they are completely given by the user. In resynthesis, the parameters obtained by analyzing an existing sound are modified." (Schwarz, 2004)

There are two approaches of music synthesis: parametric and concatenative (Schwarz, 2004). Parametric synthesis is based on physical or signal models. The signal model can be subtractive based on oscillators and filters, or additive, which is based on the harmonics plus noise model. Concatenative synthesis is based on fixed inventory or unit selection. Musical mosaicing (Zils and Pachet, 2001) can be seen as a type of concatenative synthesis based on unit selection. Concerning the synthesis of singing voice the approaches are almost the same as in music and speech synthesis.

There is a widely used technology for concatenative sound synthesis which is called Virtual Studio Technology (VST). There are several manufacturers (Vir2[10], Garritan[11], Vienna Symphonic Library[12] and others) that have developed models for this technology. These models are called VST instruments.

7.8. Summary

This section reviewed disciplines that study on different aspects of music with the help of the computer. Firstly, we presented works focused in the analysis of music. We presented some works focused on the extraction of information from music. There was a particular emphasis on music segmentation (Lartillot, 2005; Grilo, 2002; Eerola and Toiviainen, 2004) and extraction of features (Eerola and Toiviainen, 2004; McKay and Fujinaga, 2006). Obtained features were useful for the classification task (van de Laar, 2006; Wu and Jeng, 2006).

Secondly, we presented studies focused in the production/creation of music. We presented techniques used to select (Corthaut et al., 2006; Weiss, 2000; Pachet et al., 2000), sequence (Jehan, 2005), transform (Jehan, 2005; Sorensen and Brown, 2000) and synthesize (Schwarz, 2004) music.

[10] http://www.vir2.com/
[11] http://www.garritan.com/
[12] http://vsl.co.at/en/65/71/84/1349.vsl

8. Affective Computing

”Let’s not forget that the little emotions are the great captains of our lives and we obey them without realizing it.”
– *Vincent Van Gogh.*

Throughout history, many scientists have studied emotions (Damásio and Sutherland, 1996; Ekman, 1999; Frijda, 2000; Lazarus, 1999; Ortony and Collins, 1988); however, there is no consensus in their definition (Scherer, 2005). We accept emotions as corresponding to the manifestation of our psychophysiological state (Larsen et al., 2008). In this area it is important to understand emotion and their role in human behaviour and cognition (Vesterinen, 2001). They interfere with our decisions and learning processes. The outcome guides our reason. Memory works in a similar fashion. Positive events are stored with good emotions, negative events are stored with negative emotions. This background is used to build devices used to express, recognize and have emotions (Picard, 1997).

This chapter presents an overview of relevant theories and possible representations of emotions; techniques used to recognize emotions; and systems that intend to drive emotionally their musical output.

8.1. Emotion Theories

We distinguish emotions from moods and them from other types of affect. Scherer (2000) suggests five types of affect: emotions, moods, interpersonal stances, preferences and affect dispositions (Figure 8.1.1). The main differences between these are their duration and intensity. On the one hand, emotions have the highest intensity and lowest duration, affective dispositions have the lowest intensity and the highest duration. Preferences generate unspecific positive or negative feelings, with low behavioural impact except tendencies toward approach or avoidance. Attitudes are relatively enduring beliefs and predispositions toward specific objects or persons. Moods are characterized by a predominance of feelings that affect the experience and behaviour of a person. Affect dispositions describe the tendency of a person to experience certain moods more frequently or to be prone to react to certain types of emotions. Interpersonal stances are characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of people.

Design Features / ***Types of Affect***	Intensity	Duration	Synchro-nization	Event focus	Appraisal elicitation	Rapidity of change	Behavior impact
Emotions: *angry, sad, joyful, fearful, ashamed, proud, elated, desperate*	●	•	●	●	●	●	●
Moods: *cheerful, gloomy, irritable, listless, depressed, buoyant*	●	●	•	•	•	●	•
Interpersonal stances: *distant, cold, warm, supportive, contemptuous*	●	●	•	●	•	●	●
Preferences/Attitudes: *liking, loving, hating, valuing, desiring*	●	●	•	•	•	•	●
Affect dispositions: *nervous, anxious, reckless, morose, hostile*	•	●	•	•	•	•	●

Figure 8.1.1.: Scherer's types of affect (Scherer, 2000)

Ortony and Turner (1990) presented a summary of emotions theories, their basic emotions and approaches used to infer emotions (Table 8.1). Basic emotions have eleven characteristics in common: distinctive universal signals, distinctive physiology, automatic appraisal, distinctive universals in antecedent events, distinctive appearance developmentally, presence in other primates, quick onset, brief duration, unbidden occurrence, distinctive thoughts and distinctive subjective experience (Ekman, 1999). There are divergences in Ortony's summary. For instance, Weiner & Graham proposed only two basic emotions, happiness and sadness, while Arnold proposed 11 basic emotions. By analysing basic emotions from each emotions theory, we can testify the occurrence of 7 central emotions, common to most of them: anger, happiness, fear, sadness, surprise, disgust and love.

Theorist	Basic emotions	Basis for inclusion
Arnold	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	Relation to action tendencies
Ekman, Friesen and Ellsworth	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Frijda	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray	Rage and terror, anxiety, joy	Hardwired
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James	Fear, grief, love, rage	Bodily involvement
McDougall	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mowrer	Pain, pleasure	Unlearned emotional states
Oatley and Johnson-Laird	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Panksepp	Expectancy, fear, rage, panic	Hardwired
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson	Fear, love, rage	Hardwired
Weiner and Graham	Happiness, sadness	Attribution independent

Table 8.1.: Theories of emotions (Ortony and Turner, 1990)

8.2. Emotion Representation

Concerning the representation of emotions, the prevailing alternative is between discrete and dimensional systems with two or three dimensions (Daly et al., 1983). The most common interpretation for dimensions interprets them as: arousal (activation/relaxation), valence (pleasantness/unpleasantness) and dominance (degree of control over the emotional state). The first two dimensions capture most of the empirical variance, which explains that the third one is often ignored.

In the discrete representation each word describes an emotion with specific values of valence and arousal. Several authors have attempted to classify human emotions based on different criteria and coming from different fields of study (Gabrielsson and Lindstrom, 2001; Juslin and Laukka, 2004; Russell, 1989; Schubert, 1999). Although there is no consensus about considering emotions as discrete categories or as points in a multidimensional space, it is reasonable to assume that each category can be loosely mapped to a point in the valence-arousal plane. There is usually high agreement among listeners about the broad emotional category expressed by music, but less agreement concerning the nuances within this category (Juslin and Laukka, 2004). Ekman (1999) has a list of generally accepted basic emotions. Russell (1989) and Mehrabian (1980) both have lists which map specific emotions to dimensional values (using 2 or 3 dimensions). Juslin and Laukka (2004) propose a specific list for emotions expressed by music. Plutchik proposed a three-dimensional circumplex model (Plutchik, 1980). It describes the relations among emotion concepts, which are analogous to the colours on a colour wheel. In this model, the cone's vertical dimension represents intensity, and the circle represents degrees of similarity within the emotion (Figure 8.2.1).

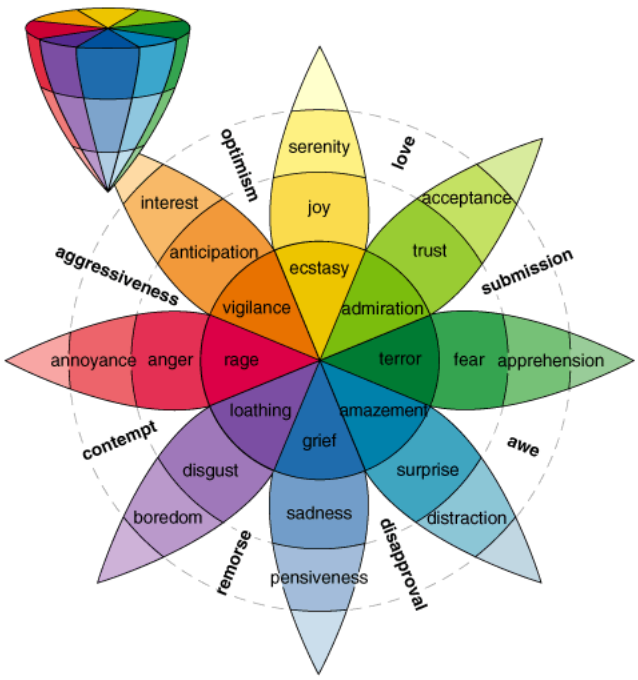


Figure 8.2.1.: Plutchik's emotions categorization

Russell (1989) proposed a two Dimensional Emotion Space (valence and arousal) to categorize 28 emotions (Figure 8.2.2). In the horizontal axis it represents valence, in the vertical axis it represents arousal. He proposed a mapping between the 28 emotions and points in the bi-dimensional space using multidimensional scaling methods. These points are an approximation of the centre of spaces representative of each emotion. This mapping is very useful because it allows to establish relations between works done in the discrete domain and the ones done in the bi-dimensional domain. It also allows to have a better perception of semantic proximity between emotions. From observing the dimensional space we can conclude that emotions are far from the centre of this space.

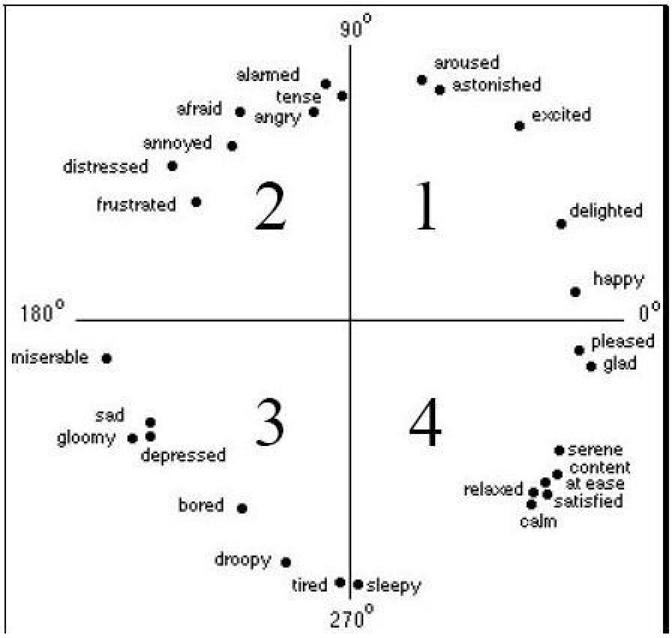


Figure 8.2.2.: Russell's emotions categorization (Russell, 1989)

Ritossa and Rickard (2004) studied the role of pleasantness and liking to predict emotions expressed by music. Songs were rated by 121 subjects, according to pleasantness, liking, arousal and familiarity. They formed positive correlations between pleasantness and liking, and familiarity and liking. They found that pleasantness is a better predictor. This study confirms the usefulness of the use of valence and arousal as dimensions used to classify emotions.

In the remaining of this document we will adopt these two dimensions to represent emotions. When doing it graphically, we will represent them as points in a bi-dimensional space with the horizontal axis representing valence and the vertical axis representing arousal.

8.3. Emotion Recognition

The interaction between humans and computers is more natural if computers are able to perceive and respond to human emotions (Busso et al., 2004). Emotions can be recognized in a number of ways. This may be via speech, facial expression, gesture, body language (Bartneck, 2001) and with a variety of other physical and physiological cues.

The research of emotion recognition gains with the development of standard tools. The Self-Assessment Manikin (SAM), for instance, is a widely used pictographic emotional rating system (Bradley and Lang, 1994). It measures the pleasure, arousal and dominance associated with a person's emotional reaction to a wide variety of stimuli. Also, the Facial Action Coding System (FACS) (Ekman and Rosenberg, 2005) is used to categorize facial expressions according to 47 muscles. EMFACS (Friesen and Ekman, 1983) is a counterpart of this system which extends this categorization to the emotional dimension.

The role of psychophysiological signals in emotion recognition is part of numerous studies, some of which we will refer to. For instance, to understand emotions we can use analytical techniques on psychophysiological signals (Etzel, 2006). Etzel studied cardiovascular and respiratory responses to identify the moods induced by music. Haag et al. (2004) showed a method of recognizing emotions using electromyography (EMG), electrodermal activity, skin temperature, blood volume pulse (BVP), electrocardiogram and respiration. Pattern recognition techniques can also be used to recognize emotional states from physiological data (Vyzas, 1999). Like Haag et al. (2004), Vyzas used BVP, EMG, electrodermal activity and respiration but also heart rate. Each emotion was characterized by 24 features extracted from the psychophysiological signals. Lisetti and Nasoz (2004) started by establishing a systematization of the results of studies that related these signals with emotions. Then, they used mappings between physiological signals and emotions to train three different supervised learning algorithms. These were used to categorize physiological signals in terms of emotions.

By knowing the important role of psychophysiological signals in the recognition of emotions, several studies have used these signals to recognize the emotions induced with music. For instance, Sloboda (1991) related physiological reactions like shivers, laughter, tears and lump in the throat with musical content. Tears were induced by music with sequences and melodic appogiaturas; shivers were evoked by new or unexpected harmonies and crescendos. Tears and shivers were also associated with syncopation and enharmonic changes. The induction of emotional peaks with the denial of musical expectation (Meyer, 1956; Eerola, 2003) is supported by the results of this work. Facial electromyography, heart rate and skin conductance were proved to be relevant signals in the detection of both discrete and continuous representations of emotions (Bradley and Lang, 2000; Klein, 2003; Khalfa et al., 2002). Klein (2003) found that corrugator EMG is negatively correlated with pleasantness and skin conductance is positively correlated with activation.

Speech is another relevant cue in the recognition of emotions. Vayrynen et al. used statistical classification to recognize emotional states from speech corpus (Vayrynen et al., 2003). They classified emotional speech stored in a database into four emotional states: neutral, sad, angry and happy. Classifiers used 43 prosodic features extracted from the speech corpus, e.g., F0 frequency, segment energy, voiced/unvoiced temporal and spectral derivatives.

8.4. Emotionally-Driven Musical Approaches

Scientific advances in Music Psychology (section 6) have been the key source of inspiration to four main approaches being used to tackle the scientific challenge of this thesis. The first approach consists in composing/arranging music, e.g., by generating music from scratch according to emotional cues (Sugimoto et al., 2008; Wassermann et al., 2003). Current automatic music composition approaches are not flexible enough to allow the adaptation of the output to different styles, which sets this approach outside our options. The second approach consists in selecting pre-composed music. It requires the extraction of musical features – statistical and perceptual – which are subsequently used to make recommendation/classification models (Baum, 2006; Trohidis et al., 2008; Yang et al., 2008; Healey et al., 1998; Wu and Jeng, 2006). The third approach involves transforming/adapting pre-composed music - currently, this approach works better at a MIDI representation level. This can be done through a knowledge-based control of structural factors of pre-composed musical scores (Livingstone et al., 2007; Wingstedt et al., 2005; Winter, 2005).

These two last approaches produce solutions with low quality when the emotional content of the source music is far from the required one. The sequential use of classification stage before the transformation overcomes the limitation of both approaches. This drives us to the fourth approach that consists in combining some of the above-mentioned alternatives (Chung and Vercoe, 2006). Chung and Vercoe, for example, used mixed techniques, but this work is grounded on an approach that seems quite ad-hoc and no technical details are available. In the remaining of this section we will present an overview of works of each of the four approaches described above.

8.4.1. Music Composition/Arranging

Automatic music composition is a challenge for the scientific world today. The challenge of composition guided by emotional cues is even bigger. We present some studies on this area. The methods being used in music composition are various: genetic algorithms (Birchfield, 2003), rule-based models (Wallis et al., 2011; Ka-Hing et al., 2006; Robertson et al., 1998; Eladhari et al., 2006; Wassermann et al., 2003; Casella and Paiva, 2001; Numao et al., 1997, 2002; Legaspi et al., 2007; Winter, 2005), n-gram models, Hidden Markov Models and other statistical models (Monteith et al., 2010, 2012). For instance, n-gram models that represent pitch intervals can generate melodies and Hidden Markov Models can produce harmonies. The number of emotions tackled in each work varies and the results are in overall satisfactory. Love, joy, surprise, anger, sadness, serenity and fear are just some examples of emotions. The areas of applications are also various, from which we highlight virtual environments

(Robertson et al., 1998; Wassermann et al., 2003; Casella and Paiva, 2001) and videogames (Eladhari et al., 2006).

Some works control the emotional content of composed music with psychophysiological data. Heart rate (McCaig and Fels, 2002), facial expressions (Funk et al., 2005), galvanic skin response, electromyography (Kim and André, 2004), muscle tension, breathing, temperature and gestures Nakra (1999) are some examples. Several mappings can be established to help the composition of music that expresses appropriate emotions. McCaig and Fels mapped heart rate to musical parameters (tempo, timbre, pitch, repetitiveness of musical structure) that reflect musical tension. Funk et al. mapped musical features to specific zones of the face areas. Nakra mapped performers' gestures and breathing signals to real-time expressive effects by defining musical features (beats, tempo, articulation, dynamics and note length) in a musical score.

User behaviors and context are another type of data which can guide the process of music composition (Gaye et al., 2003). The system developed by Gaye et al. extracted variables from the body and environment. Discrete factors (e.g., user action change) and continuous factors (e.g., physiological state and continuous actions) were used to change musical characteristics. Sound layers, temporal structure, timbre and envelope were some of these characteristics. Discrete factors triggered short events (e.g., doubling the tempo), continuous factors were used to define the timbre of the composition.

Aesthetics principles can also be the source of inspiration for music composition systems like the one proposed by Goga and Goga (2003). This system produced melodies based on particular music structure patterns or musical rules. This work aimed to induce feelings like restlessness, peace, consolation, innocence, delicacy, sadness, trust, love, joviality and joy. For instance, love is characterized by this pattern: "Gradually ascending movement followed by gradually descendant movement combined with jumps of fifths and followed by the repetition of the same sound"; sadness is characterized by "Gradually descending movement followed by descending movement in jumps combined with ascending movement in jumps (large values for the times of the notes)".

8.4.2. Classification/Selection of Pre-composed Music

Music is said to be one of the languages of emotions (Pratt, 1948). This section focuses on the classification of emotional content in music and on the emotionally-driven selection that uses the analysis and selection of specific features. A good overview about existing research in music emotion recognition is (Kim et al., 2010). This task involves disciplines like signal processing, machine learning, auditory perception, psychology and music theory.

Classification of emotional sounds can be done through matching specific patterns of energy dynamics (Moncrieff et al., 2001). Moncrieff et al. used four patterns of sound energy to induce four emotions during horror films: surprise or alarm; apprehension or event emphasis; surprise followed by alarm; apprehension up to a climax. They analysed six dynamic features to classify sounds with certain patterns: step edge attack; step edge decay; slope attack; slope decay; low sound energy; sustained energy.

Prominence, roughness, loudness, articulation, brightness, onset and tempo are some features that can be used to study expressiveness in audio music (Leman et al., 2003). Leman et al. mapped these features to a three dimension emotional space.

Emotions detection can be seen as a classification problem, therefore the selection of the classifier model and the feature set are crucial to obtain good results (Carvalho and Chao, 2005). Van de Laar (2006) made a comparison between six emotion detection methods in music based on acoustical feature analysis (Table 8.2). He used four central criteria in this comparison: precision, granularity, diversity and selection. The referred methods consider eight fundamental features: timbral texture features, spectral flatness measure, spectral crest factor, mel frequency cepstral coefficients, Daubechies wavelet coefficient histogram, beat and tempo detection, genre information and lyrics.

Criteria	**Carvalho and Chao**	**Li and Ogihara (2003)**	**Li and Ogihara (2004)**	**Feng, Zhuang and Pan**	**Liu, Lu and Zang**	**Yang and Lee**
Precision	good	moderate	good	excellent	excellent	excellent
Granularity	bad	excellent	moderate	bad	bad	good
Diversity	moderate	excellent	moderate	moderate	moderate	very bad
Selection	bad	bad	bad	bad	excellent	bad

Table 8.2.: Comparison of emotion detection methods (van de Laar, 2006)

We are now going to present details about several detection methods. Different types of classifiers have been used: sequential stack classifier (Carvalho and Chao, 2005), support vector machines (Li and Ogihara, 2003; Muyuan et al., 2004; Baum, 2006), backpropagation neural network (Feng et al., 2003), gaussian mixture models (Liu et al., 2003), psychological models (Yang and Lee, 2004), fuzzy approaches (Yang et al., 2006), regression models (Yang et al., 2008), self-organizing maps, naive bayes and random forests (Baum, 2006). The feature set for the classifier also varies: timbral texture features (Carvalho and Chao, 2005; Li and Ogihara, 2003; Liu et al., 2003; Yang and Lee, 2004; Yang et al., 2008; Trohidis et al., 2008), rhythmic content (Li and Ogihara, 2003; Liu et al., 2003; Yang and Lee, 2004; Yang et al., 2008; Trohidis et al., 2008), pitch content (Li and Ogihara, 2003; Yang et al., 2008), relative tempo, the mean and standard deviation of average silence ratio (Feng et al., 2003), intensity, features from MPEG-7 audio standard (Allamanche et al., 2001) and features using the Sony Extractor Discovery System (Yang and Lee, 2004), statistical and perceptual (Muyuan et al., 2004; Yang et al., 2008), frequency centroid, spectral dissonance (Liu et al., 2006;

Yang et al., 2008), pure tonalness (Liu et al., 2006) and loudness (Yang et al., 2008). Some of these features were extracted with the help of Psysound (Cabrera, 1999) and Marsyas (Tzanetakis and Cook, 2000b). The number of emotions is also another parameter that varies across these works: two (Carvalho and Chao, 2005), four (Feng et al., 2003), five (Carvalho and Chao, 2005), six (Muyuan et al., 2004) and thirteen (Li and Ogihara, 2003). The number of songs also varies: 499 (Li and Ogihara, 2003), 593 (Trohidis et al., 2008), 1000 (Baum, 2006) and others not mentioned by the authors.

The sequential stack classifier used by Carvalho and Chao (2005) outperformed classifiers like decision trees, logistic regression and conditional random fields. Two-label classification obtained a success of 86%, five-label classification achieved a success of 36%. Li and Ogihara (2003) obtained an average precision of 0,32 and an average recall of 0,54.

Another approach consists in extracting emotional expression from music (Wu and Jeng, 2006). The method designed by Wu and Jeng has three steps: subject responses, data processing and segments extraction. The use of the results of this method allows the association of emotional content to musical fragments, according to features like pitch, tempo and mode. Similarly, Friberg et al. (2002) designed a model to predict the expressive intention during music performance. Average and variability values of sound level, tempo, articulation, attack velocity and spectral content were extracted. Listening experiments served to build linear regression models to predict intended emotion based on features.

There are also models to recommend music based on emotions (Kuo et al., 2005). The model of Kuo et al., based on association discovery from film music, proposed prominent musical features according to a queried emotion description. These features were compared with features extracted from a music database (chord, rhythm and tempo). Then, music was ranked and a music list was recommended. This system used MIDI files and 15 groups of emotions (e.g., love, distress, sadness and pity).

Affective and psychophysiological data is very important to adapt music to our needs. With this in mind, the following paragraphs describe some works that recognize this data at some intervals to control the selection of music. Physiological data like Galvanic Skin Response (GSR), skin temperature, heat flow, body temperature and heart rate is used to guide the selection of music (Oliver and Flores-Mangas, 2006; Dornbush et al., 2005; Wijnalda et al., 2005; Janssen et al., 2009). These emotion aware systems automate the process of selecting music by learning the user's preferences, emotions and activity. Neural networks, regression and kernel density estimation are just some of possible models that can be used in the learning process. These systems are used to improve exercise performance by personalizing music to exercises.

Healey et al. (1998) developed an interface of a wearable computer that perceives and responds to the user's affective state. It recognizes and responds to signals with emotional information. They used an algorithm in music selection to change from current affective state to the intended state. This algorithm compares GSR of the last 30 seconds of previous song with the first 30 seconds of the current song. Current affective state is predicted based on user preferences and physiological variables. These variables are measured based on electromyogram, photoplethysmograph (heart rate and vasoconstriction) and galvanic skin response.

Vavrille (2006) developed an interactive web radio. In this system it is possible to select music by mood (and genre) and also to see the relationship between music pieces. Music classification by mood is based on a two Dimensional Mood Space (Thayer mood model). Users can select and listen to music by using a mood matrix and then navigate through artists that evoke similar moods.

Meyers (2007) developed a system to generate music playlists based on the emotion or mood of the user. Chunks of texts were associated with a set of songs and ConceptNet (Liu and Singh, 2004) was used to extract emotional content from these texts. This extraction process was aided by All Music Guide[13] mood classification, used to link songs and artists to moods.

8.4.3. Transformation of Pre-composed Music

Music emotional content can be transformed in both audio and MIDI domains. The following paragraphs present some works in these areas.

8.4.3.1. MIDI

Livingstone and Brown (2005a) established relations between music features and emotions using results of previous work (Schubert, 1999; Gabrielsson and Lindstrom, 2001). Both emotions and a set of music-emotion structural rules are represented in a two dimensional emotion space with an octal form (Figure 8.4.1). They designed a rule-based architecture to affect the perceived emotions of music by modifying the musical structure (Livingstone and Brown, 2005b). They used a music performance engine (Livingstone et al., 2005) to adapt the symbolic score's reproduction to the audience emotions. This engine is composed by three modules: the engine that contains the rule system and emotive algorithms; the score (MIDI); and data of the audience and application. Later, Livingstone et al. Livingstone et al. (2006, 2007) made a list of performative and structural features and their emotional effect. Tempo, mode, loudness, articulation, pitch and harmony are the structural parameters. Expressive contour, tempo variation,

[13] http://allmusic.com/

tone attacks, stable note accent, phrase arch, pedal accent, originality, stochastic fluctuations, chord asynchrony, melody accent, note accent and slurs are the performative parameters.

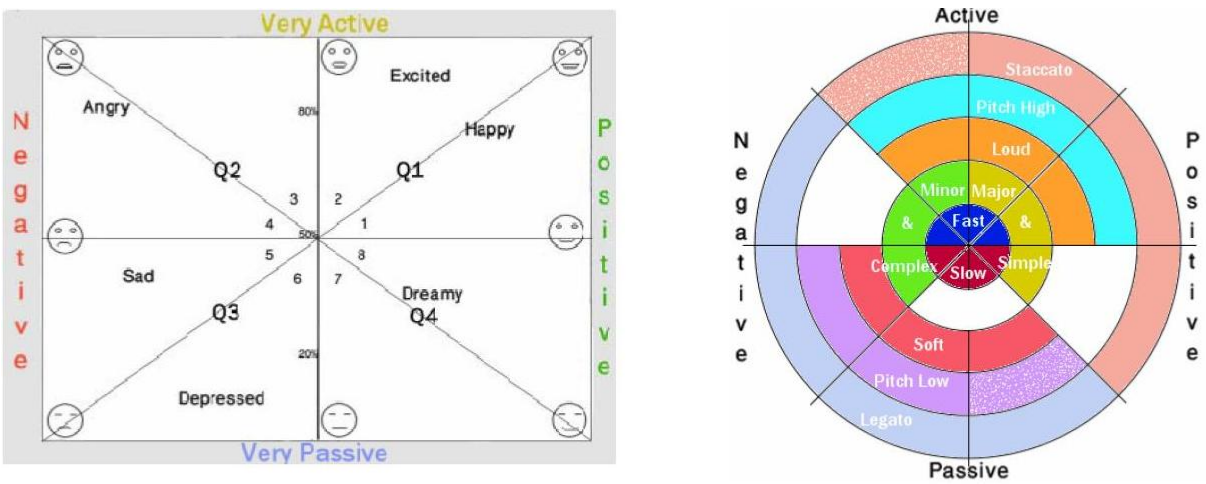


Figure 8.4.1.: Livingstone's space of emotions and space of music-emotion structural rules (Livingstone and Brown, 2005a)

The MIDI-based software named REMUPP was designed to study aspects of musical experience (Wingstedt et al., 2005). This system allows the real-time manipulation of musical parameters like tonality, mode, tempo, harmonic and rhythmic complexity, register, instrumentation and articulation. For instance, articulation is changed by altering the length of notes and register is changed by altering the pitch of notes. This system has a music player that receives music examples and musical parameters. Music examples are composed by a standard MIDI file (SMF) and a set of properties. Musical parameters can be used to control the sequencer and synthesizers or to employ filters and effects on MIDI stream. The music player loads the SMF into the sequencer. Musical parameters are both used to manipulate MIDI data and the way this data is rendered by synthesizers. Figure 8.4.2 details aspects of the music player of this system.

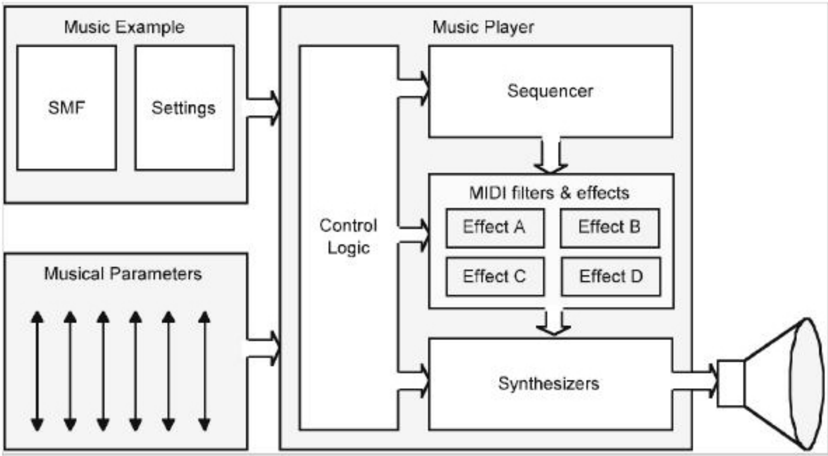


Figure 8.4.2.: Architecture of the REMUPP music player (Wingstedt et al., 2005)

Music can be generated based on examples of human performances (Arcos and de Mantaras, 2000). The cited work combines Case-Based Reasoning (CBR) and fuzzy techniques. Musical knowledge is stored in cases that represent the score (melody with a sequence of notes and harmony with a sequence of chords), the musical analysis of the score (a tree describing metrical, tensing and relaxing relations among notes) and information about expressive performances of the score (affective expressivity of sequences of notes). The system uses fuzzy techniques in the reuse step of CBR.

Winter (2005) created a system expanding on pDM (Friberg, 2006) which also manipulates harmonic features of the music. He built a real-time application to control structural factors of a composition. This application is grounded on models of musical communication of emotions. These models showed the emotional relevance of some musical features (Figure 8.4.3). In this figure, we can see weights of emotional importance (between -1 and 1). Weights closer to -1 or 1 are the most important: mode (-0.73) and tempo (0.55) stand out. These values were obtained through regression analysis. Pre-composed music scores are manipulated through the application of rules that control values of features: mode, instrumentation, rhythm and harmony. Winter uses an emotional control space (valence and arousal) to define these values. A MIDI file is produced to give emotional feedback to the user.

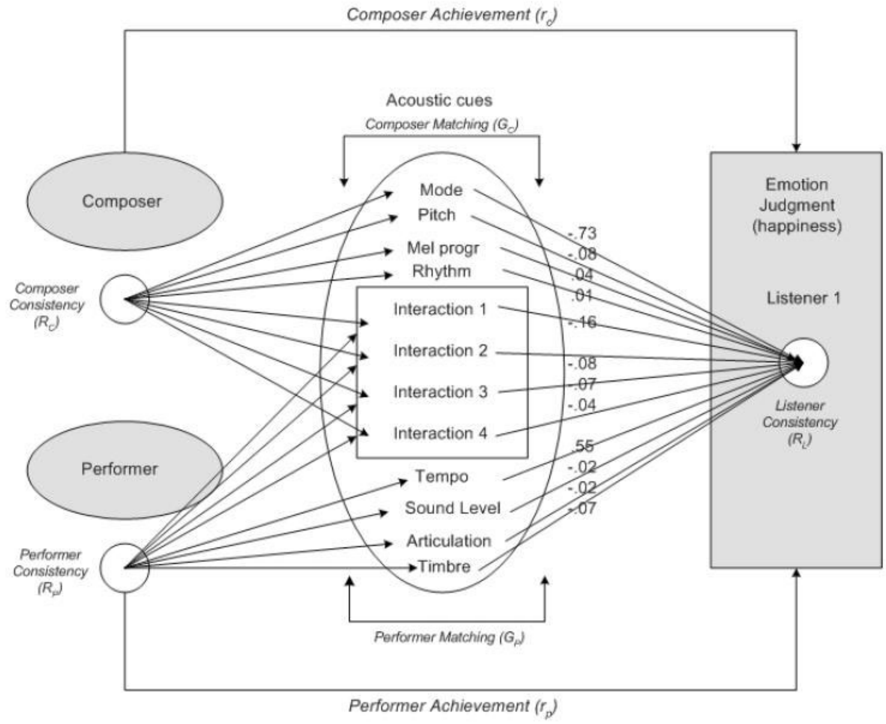


Figure 8.4.3.: Expanded model of musical communication (Juslin and Laukka, 2004)

8.4.3.2. Audio

Musical effects can be used to communicate affects on ambient music Barrington et al. (2006). In this work, 14 musical effects were tested: no effect; low-pass filter; skip backwards by two beats; skip forward by two beats; high-pass filter; repeat eight beat passage delayed eight beats, with reverb; filter sweep from 100 to 10000 Hz, twelve beat period; play backwards for 1 beat, then play forward; dub with low-pass filter; dub with high-pass filter; modulate tempo linearly +- 10%, period = one beat; add flange effect intermittently on the beat; modulate amplitude linearly +- 100%, twelve beat period; flange effect intermittently on the beat with high-pass filter. Ten subjects evaluated 45 music samples, lasting around 20 seconds, using three criteria: activity level (relaxed, normal or agitated), awareness (no effect, just noticeable, detectable, obvious or dominant) and enjoyment (very pleasant, pleasant, neutral, unpleasant or very unpleasant). Experiments suggested that relaxed states can be induced with the application of the low-pass filter, the dub effect plus low pass filter or filter sweep. On the other hand, agitated states can be induced by using high-pass filtered dub and rewind effects.

Expressive audio synthesis is used to change a set of musical parameters to synthesize music with different affective content DInca and Mion (2006). This work had the goal to derive mappings between an expressive control space and parameters of a rendering model, through a synthesis by analysis framework. The analysis stage consisted in establishing associations between affective and sensorial categories, and in finding relevant features from music performances. The synthesis stage consisted in the development of an expressive tone generator. Tempo parameters (attack, duration, notes per second), intensity parameters (peak sound level, sound level range) and perception parameters (roughness and centroid) were manipulated. Tempo and intensity parameters were controlled with ADSR envelope values; perception parameters were controlled by changing frequency and amplitude of harmonics. Listening tests carried out in this work, by using real and synthetic sounds, confirmed the possibility to communicate different intentions with simple sounds.

8.4.4. Hybrid Approaches

Chung and Vercoe (2006) developed a system to generate music in real-time based on intended listener's affective cues. This system correlates musical parameters with changes in affective state. Personal expression is analysed while listening to music, like head nodding, hand tapping, foot tapping, hand clapping, mock performing, mock conducting, dancing and other gestures. Both affective states and musical parameters are represented in a two dimensional emotion space. Music is composed using a multi-track audio environment and is listened to by eight subjects. Music files are generated in real-time by music composition/production, segmentation and re-assembly of music. The analysis of listeners' affective state is based on physiological data, physical data and a questionnaire. Listener data is used to develop a probabilistic state transition model to infer the probability of changing from one affective state to another. This work supports the ideas that: engaged and annoyed listeners tend to stay in the same affective state, soothed listeners tend to stay soothed but can become easily bored and/or engaged, and annoyed listeners tend to become engaged if induced to boredom. Foot-tapping is a useful indicator of subjects' valence.

8.5. Summary

This chapter reviewed some theories (Ortony and Turner, 1990; Scherer, 2000) and representations (Russell, 1989) for emotions. We presented techniques used to recognize emotions (Etzel, 2006; Vyzas, 1999) and studied four principal approaches used in generation of music with appropriate emotional content. These approaches consist in the transformation/adaptation of pre-composed music (Livingstone, 2008; Wingstedt et al., 2005), music composition/arranging (Winter, 2005; Kim and André, 2004) and music selection/classification (Healey et al., 1998; Trohidis et al., 2008). There are also hybrid approaches (Chung and Vercoe, 2006).

9. Reflexion on the State Of The Art

As we have discussed in section 8.4, there are four approaches to solve the problem addressed by this thesis. Automatic composition mechanisms are generally conceived for a bounded range of musical styles, and sometimes do not tackle the whole composition process (e.g., only deal with melody or with rhythm). We would like to have the flexibility of producing complete music pieces in a wide range of styles, so this approach is not very suitable. Studies grounded on classification of pre-composed music and subsequent selection are scalable, but the quality of their answers is very dependent of the original music base. This one is, actually, a finite database, and thus cannot cover entirely the whole emotional spectrum. Therefore, one has to expect to select pieces that don't match exactly the intended emotion (see Figure 9.0.1). The approach based on transformation has the disadvantage of producing outputs with low quality when the original music has characteristics very different from the desired ones. None of these three approaches, alone, gives an entirely satisfactory response to our requirements. The fourth approach consists in the hybrid combination of the former ones in order to overcome some of their weaknesses.

For the purpose of our work, we found especially promising a particular hybrid approach that consists in combining classification/selection with transformation. In fact, the transformation can improve the classification/selection result when there is not a solution in the music base close to the emotional specification (Figure 9.0.2). On the other hand, as the selection tends to produce an output with characteristics close to the desired ones, the transformation assumes less risks of degrading music quality, because the adjustments needed to get the music characteristics to fit the emotional specification are limited.

The solution proposed in this thesis has the advantage of being able to produce outputs of acceptable quality quite independently of the music base: it will be able to find the best possible match and then transform it in order to increase the match even further. It is also quite flexible: the music base can be completely redefined to adapt to the specific needs of a given use scenario. The system uses mechanisms (modules) that are independent from the music it is working with, i.e., the musical output corresponds to the emotional specification independently of the original music base. The system is also reliable, thanks to the experimental calibration using different subjects (section 8.3 is particularly relevant for this task), as described in Chapter 15.

Grounded on the state of the art, we found other opportunities to contribute to its advance: to adopt both the discrete and dimensional representation of emotions; systematize the relations between emotions and musical features in the knowledge base (subsection 11.6.2) by studying the musical features with an emotional impact (as we have seen in chapter 6); develop modules to control the emotional content of music; use techniques of human emotional recognition for validation and calibration of the system. This thesis explored these directions of research to achieve its goal. It also tested the usability of a version of EDME system ready to be used in real-time and with an interface that can be used in application domains like entertainment and healthcare.

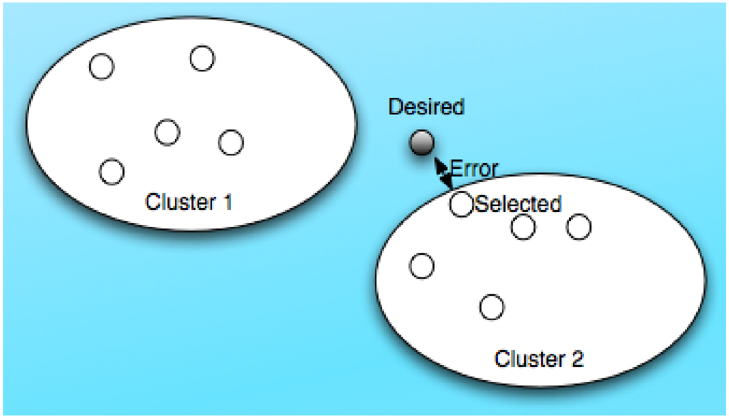


Figure 9.0.1.: Error resulting from Selection when no music exists with an exact match

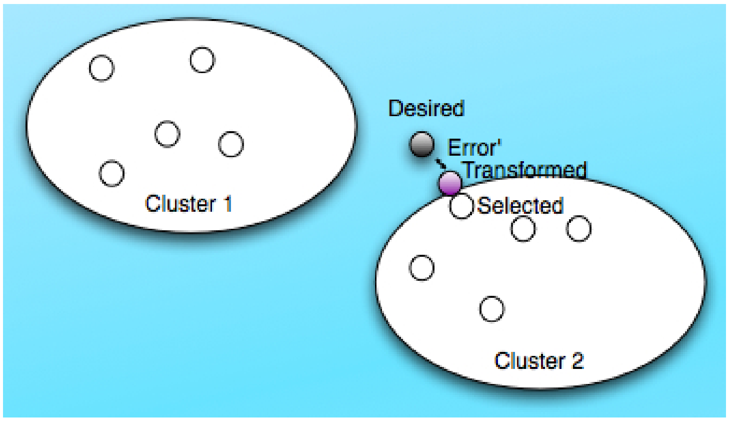


Figure 9.0.2.: The effect of Transformation after Selection on the error

Part III.

Emotion-Driven Music Engine

10. Approach

In the last part, we found some approaches that focus on musical composition (Legaspi et al., 2007; Kim and André, 2004), others in musical selection/classification (Yang et al., 2008; Kuo et al., 2005) and others in transformation of pre-composed music (Winter, 2005; Wingstedt et al., 2005; Livingstone, 2008; Friberg et al., 2006). Those based on a combination of these approaches are rare (Chung and Vercoe, 2006). We intend to face the problem of controlling the emotional content of produced music by using a combination of these approaches that uses four modules: segmentation, classification, selection and transformation. We propose to take the best from the control opportunities in each module to achieve better results. Segmentation module is meant to obtain musical segments that can express a single emotion. We analysed the influence of the variation of rhythmic, melodic, harmonic, instrumental and dynamic features to obtain favorable points of segmentation. Classification module tags each segment with emotional values (Oliveira and Cardoso, 2008c). Selection module uses the euclidean distance metric to calculate the emotional distance between the music and the desired emotion. Transformation module is intended to bring the emotional content of the selected music closer to the desired emotional expression.

11. Architecture

Our computational system, called Emotion-Driven Music Engine (EDME), produces music expressing a desired emotion (Oliveira and Cardoso, 2010). EDME consists of four main modules (segmentation, classification, selection and transformation) used to control the emotional content of music and three auxiliary modules (feature extraction, sequencing and synthesis) responsible for doing work necessary for the main modules. Some of the modules recur to four auxiliary structures (music base, knowledge base, pattern base and library of sounds) to store content. The system interacts with the listener through a user interface and interacts with the administrator through an administrator interface.

The system works in two stages, one offline and another online (Lopez et al., 2010). In the offline stage (Figure 11.0.1), the segmentation module uses pre-composed music, in order to generate musical segments that express only one emotion. These segments are given to the module of features extraction to obtain values of musical features that will be used by the classification module. This last module uses the knowledge base to label the segments with emotional values of valence and arousal. Segments emotionally classified are stored in the music base. The administrator interface of the system allows the administrator to segment and classify the segments.

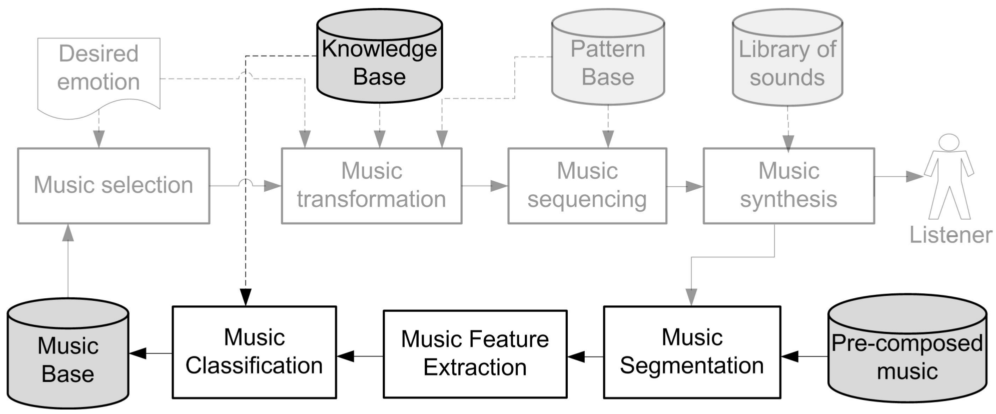


Figure 11.0.1.: EDME architecture: offline stage. Modules of this stage are marked in bold, modules of online stage are greyed out.

In the online stage (Figure 11.0.2), the selection module calculates the distance between the desired emotion and the emotional values of each segment. The segments with the minimum distances are selected from the music base. The transformation module brings the emotional content of the selected segments closer to the desired emotion by changing features emotionally relevant. The sequencer module packs the transformed segments using musical patterns (available in the pattern base) in order to form songs. The synthesis module selects sounds (from the library of sounds) to convert the MIDI output into audio. The user interface of the system allows the listener to define the desired emotion.

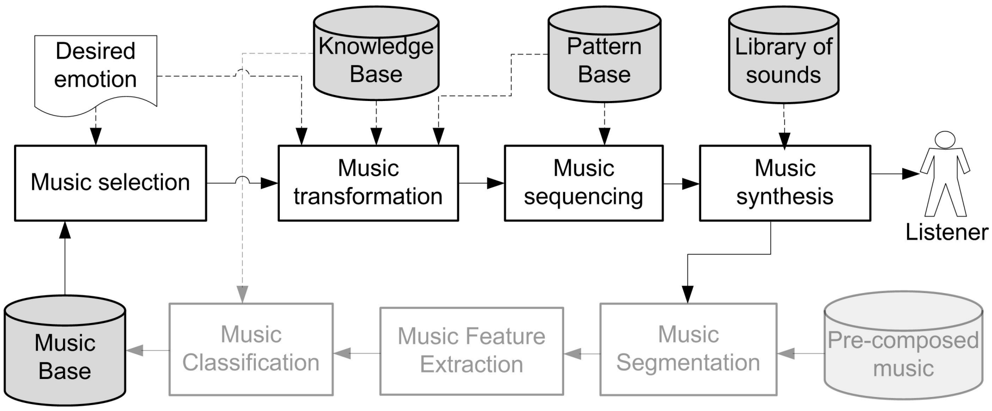


Figure 11.0.2.: EDME architecture: online stage. Modules of this stage are marked in bold, modules of offline stage are greyed out.

11.1. Segmentation

The system is using pre-composed music that consists of standard MIDI files compiled from websites, although they could come from other sources, or possibly composed on purpose. These files are polyphonic and can be of any musical style. The segmentation module uses each of these files to produce segments as much as possible with a musical sense of its own and expressing a single emotion (Figure 11.1.1). By obtaining smaller music pieces, we decrease the probability of finding more than one emotion in the segment. We made some a subjective perceptual assessment of the three segmentation algorithms available on the MIDI Toolbox (Eerola and Toiviainen, 2004). Two of the algorithms are rule-based, the other one is statistical (or memory-based). The statistical algorithm uses probabilities derived from the analysis of melodies (Bod, 2002). The rule-based algorithm of Tenney and Polansky (1980) finds locations where there are large pitch intervals and large inter-onset-intervals. The other rule-based algorithm, which is called the Local Boundary Detection Model (Cambouropoulos, 1997), finds large variations of pitch, rhythm and silence. Both rule-based algorithms are grounded on gestalt principles. The Local Boundary Detection Model (LBDM) was the algorithm that reavealed the best results on this subjective assessment.

The segmentation module works in two stages. In the first stage it attributes weights to each note onset by using an adaptation of LBDM. These weights are attributed according to the musical importance, degree of proximity and degree of variation of five features: pitch, rhythm, silence, loudness and instrumentation. The degree of proximity and the degree of variation are calculated according to the LBDM; musical importance is a parameter that was defined after making some perception tests aiming to find the best points of segmentation.

In the second stage, the module searches for plausible points of segmentation according to the weights attributed at each note onset. There is a threshold defined to reduce the weights' search space: note onsets with weights below this threshold are not considered. The length of obtained segments is defined by a minimum (MIN) and maximum (MAX) number of bars. The module searches for a plausible point of segmentation that corresponds to the maximum weight obtained between the first bar of music file + MIN and the first bar + MAX. This process is then iterated, starting from the bar of the last point of segmentation, till the end of the file.

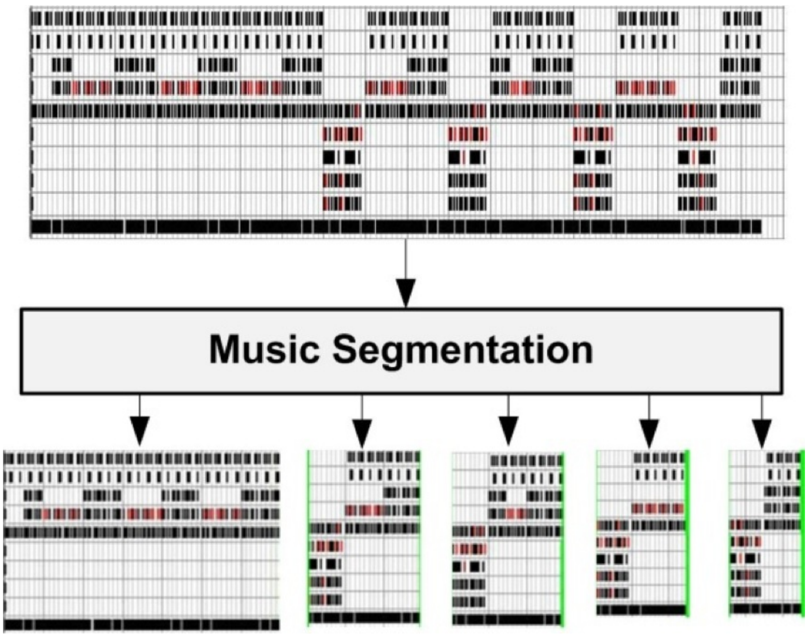


Figure 11.1.1.: Input and output to the segmentation module

11.2. Classification

The classification module uses the knowledge base (subsection 11.6.2) to determine the emotional content of the segments coming from the module of feature extraction (subsection 11.5.1). This last module also gives the values of the features obtained for each emotional dimension. The knowledge base is used to compute the following weighed sums:

$$Valence = \sum_{i=0}^{n} valenceFeatureWeight_i * valenceFeatureValue_i$$

$$Arousal = \sum_{i=0}^{n} arousalFeatureWeight_i * arousalFeatureValue_i$$

The computed values are stored as tags with the segments in the music base (subsection 11.6.1). This module is using the thirteen features identified in chapter 13, i.e., average note duration, average time between attacks, importance of bass register, tempo, note density, percussion prevalence, repeated notes, variation of dynamics, key mode, spectral loudness, spectral dissonance (Sethares), spectral sharpness (Ambres) and spectral similarity.

11.3. Selection

The selection module obtains a list of segments from the music base (subsection 11.6.1) that are closer to the desired emotion. It calculates the Euclidean distance[14] between the desired emotion and the emotional content of each segment. The results are used to put the segments in a list ordered by the degree of similarity to the desired emotion. This module retrieves the segments that are on the top of the list. The number of segments that are retrieved is customizable.

11.4. Transformation

The transformation module was designed to use the two regression models of the knowledge base (subsection 11.6.2) in order to approximate the emotional content of selected segments to the desired emotion. This module should calculate two Euclidean

[14] http://en.wikipedia.org/wiki/Euclidean_distance

distances[15]: the distance between the valence of each selected segment and the valence of the desired emotion; and the distance between the arousal of each selected segment and the arousal of the desired emotion. Both distances should be minimized after transforming musical features by a specific quantity. This quantity depends on the quotient between each distance and the weight of the feature defined in the regression models (Weisberg, 2005) of each emotional dimension.

The features to consider in this module should be obviously the same that we used in the classification, i.e., those found to be the most relevant according to the experiments conducted (chapter 12). However, because we developed the transformation module in a stage (section 12.7) before the systematization of the knowledge base (chapter 13), it does not have algorithms that transform those thirteen features used in the classification module. It has only five algorithms that transform the following features: tempo, pitch register, musical scale, instruments and articulation. We decided to use these features, because of its importance in the literature and in the three experiments. We present details of each algorithm along the section 12.8. The transformation module to be implemented later on should consider the features involved in the classification.

Let us give an example to see how the transformation should work. Suppose we want a desired emotion of $Valence, Arousal = (0.95, 0.4)$ with $Valence, Arousal \in [-1, 1]$ and the music with the closest emotional content that the system can retrieve has $Valence, Arousal = (0.5, 0.4)$. The dimension of arousal does not need to be changed; however, the system needs to change the dimension of valence from 0.5 to 0.95. If the regression model of valence has an equation of 0.005*tempo+0.005*pitch, the system has to transform the tempo and pitch. Supposing that the retrieved music has a tempo 50 and pitch of 50, the desired valence can be achieved by transforming tempo to 120 and pitch to 70, in order to meet the desired emotion.

11.5. Auxiliary Modules

Auxiliary modules are important to a good functioning of the system. Next subsections present the modules of feature extraction, sequencing and synthesis in detail. Auxiliary modules differ from auxiliary structures (section 11.6) since modules process data and structures only store data.

11.5.1. Feature Extraction

The feature extraction module labels each segment with emotionally relevant features (Figure 11.5.1). This module uses toolboxes that obtain features known to be rele-

[15] http://en.wikipedia.org/wiki/Euclidean_distance

vant to our system according to empirical results obtained both from the literature (e.g., Gabrielsson and Lindstrom, 2001; Livingstone et al., 2007) and from our experiments (see next chapter). We were focused only on global features, local features were not considered. The JSymbolic (McKay and Fujinaga, 2006), MIDI Toolbox (Eerola and Toiviainen, 2004) and JMusic (Sorensen and Brown, 2000) extract MIDI features; MIR Toolbox (Lartillot and Toiviainen, 2007) and Psysound Toolbox (Cabrera, 1999) extract audio features. We developed our own algorithms to extract additional MIDI and audio features (e.g., average loudness and spectral similarity). Average loudness corresponds to the average velocity of all the MIDI notes. Spectral similarity calculates a similarity matrix with the help of MIR Toolbox in order to find the difference between consecutive frames of the frequency spectrum. It reflects the smoothness of the music (the changes of features along the music). Both have a relationship with the arousal of music (Schubert, 1999). It is possible to extract 482 features which belong to six categories: instrumentation, dynamics, rhythm, melody, texture and harmony.

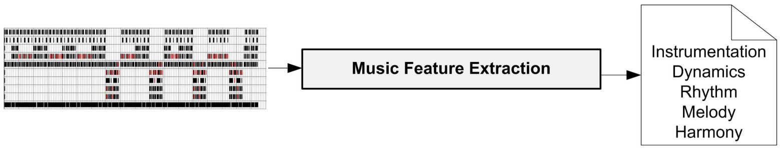


Figure 11.5.1.: Input and output of the module of feature extraction

11.5.2. Sequencing

Music sequencing module has the objective to obtain a smooth sequence of segments with similar emotional content. The sequencing module resorts to the pattern base (subsection 11.6.3) to pack the segments to form a sequence of songs. Segments are arranged in order to match the tempo and pitch of the selected pattern. The tempo of the segments is normalized to their average tempo. The pitch is raised or lowered, by comparing the key of the current pattern with the key of the non-transformed segments. We also applied algorithms of fade in and fade out to smooth the transitions between segments, respectively, by gradually increasing the volume of the starting segment and decreasing the volume of the finishing segment.

We present an example (Figure 11.5.2) where the user wants to hear music expressing a delighted emotion, represented as $Valence, Arousal = (0.8, 0.4)$ with $Valence, Arousal \in [-1, 1]$. The system selects three MIDI segments (the ones closer to the desired emotion) to match the current -ABCA- pattern. The first segment, with C as the tonic and a tempo of 100 bpm, acts as the root of the pattern. The second segment needs transformations to match the tempo (+10 bpm) and the pitch (the IV-subdominant of C is F, so -5 semitones gets B♭ to F). The third segment needs transformations to match the tempo (-20 bpm) and the pitch (the V-dominant of C is G, so +3 semitones gets E to G). Finally the first segment is repeated to end the pattern. The segments are sequenced in order to be perceived as a single part with distinct harmonic relations and equal tempo.

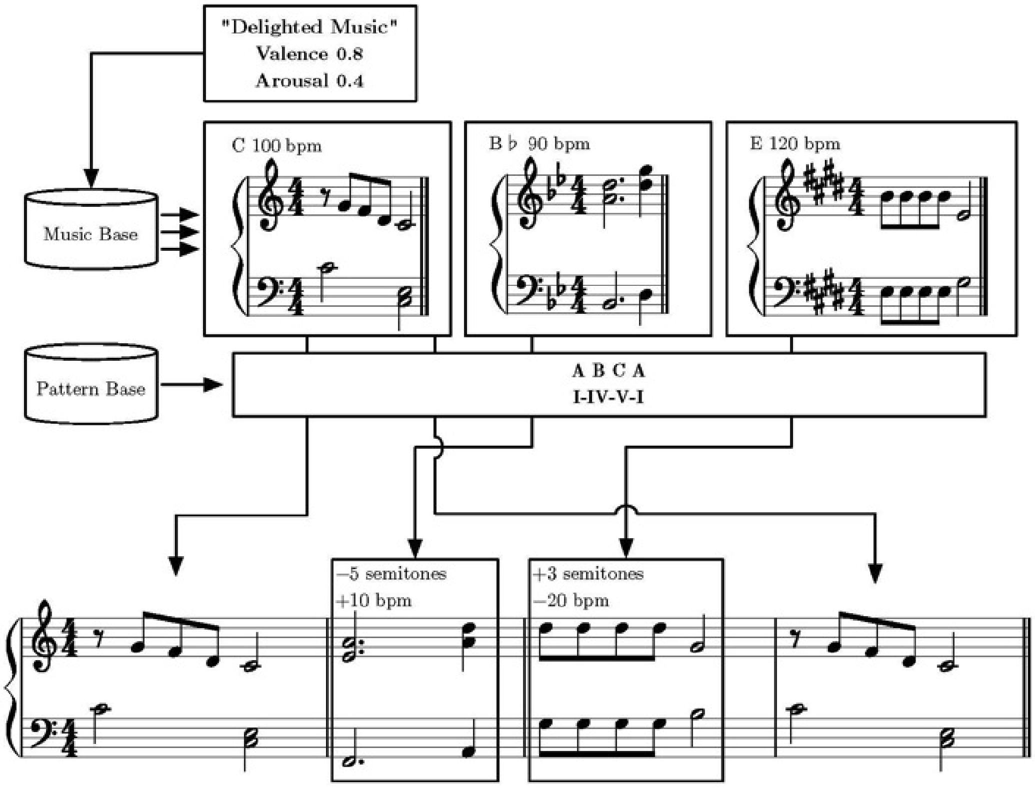


Figure 11.5.2.: Sequencing example

11.5.3. Synthesis

The synthesis module uses the library of sounds (subsection 11.6.4) to analyse and control the emotional content of the instruments being used (Oliveira and Cardoso, 2008b). This module calculates the emotional content of the samples of each instrument according to the spectral dissonance (Figure 11.5.3) and spectral sharpness (Figure 11.5.4). The module is using Psysound toolbox (Cabrera, 1999) to extract these features. Dissonance is used to label arousal and sharpness is used to label valence (Oliveira and Cardoso, 2008b). The emotional content drives the selection of sounds from the library in order to produce an audio output.

11.6. Auxiliary Structures

We defined four structures with the objective of storing content useful for the modules. The music base stores musical material; the knowledge base stores regression mod-

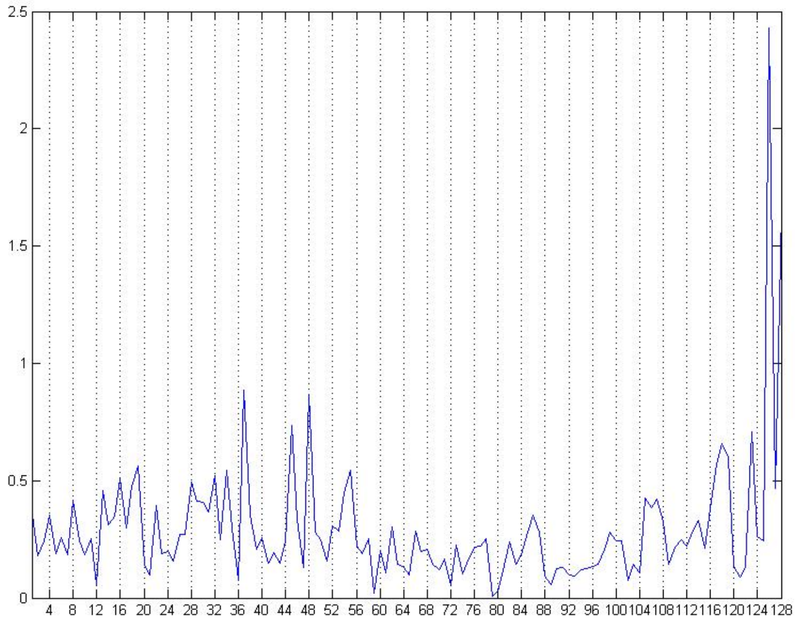


Figure 11.5.3.: Arousal of the instruments

els; the pattern base stores musical patterns; and the library of sounds stores sound material. Next subsections present each structure in detail.

11.6.1. Music Base

The music base stores musical content: standard MIDI files and the corresponding musical features. The system uses music obtained in websites; however, it can be fed by music composed on purpose or obtained from other sources.

Standard MIDI files store structural and performative aspects of music in a binary format. These musical aspects are stored with the help of very simple music information: note onset and note offset, pitch and velocity (loudness). We established quality constraints through the analysis of several parameters: tempo variation, the number of tracks, notes falling on the beats, orchestration, presence of pitch bending, presence of midi control messages. We are using professional MIDI files obtained from websites [16] [17] [18] [19].

11.6.2. Knowledge Base

The knowledge base stores two regression models (Weisberg, 2005) that establish

[16]www.classicalarchives.com
[17]midiworld.com
[18]kssdsd.com
[19]http://www.midi-classics.com/tune1000.htm

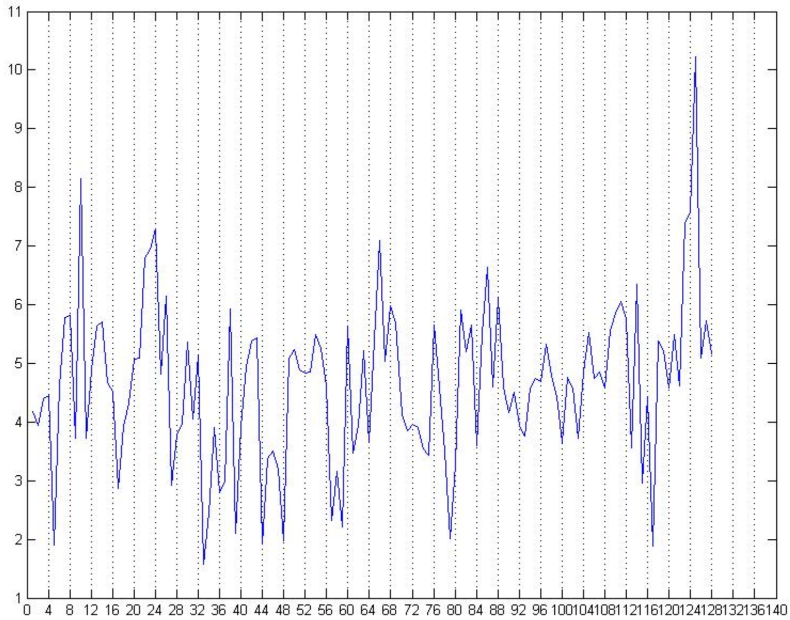


Figure 11.5.4.: Valence of the instruments

relationships between the emotional and musical domains. The knowledge base is using one regression model for each dimension of the emotional domain: valence and arousal. The musical domain is divided into several dimensions, determined by the number of musical features being used. The regression models provide weighted relations between the musical features and the emotional dimension in question. They were built by applying feature selection and regression algorithms (Witten et al., 1999; Weisberg, 2005; Guyon and Elisseeff, 2003) on experimental data obtained with questionnaires (see chapter 12 for more details). The regression models must be independent from social variables like the age of the listeners and musical variables like the musical style.

At the end of the first study of the validation/calibration (subsection 15.1.1), one regression model was using seven features to relate valence and the musical domain: average time between attacks, tempo, repeated notes, variation of dynamics, key mode, spectral dissonance and spectral sharpness. The other regression model was using six features to relate arousal and the musical domain: average note duration, importance of bass register, tempo, note density, spectral loudness and spectral dissonance.

11.6.3. Pattern Base

The pattern base structures musical sequences with the help of musical patterns. Each pattern defines a song structure and the harmonic relations between the segments of the structure (e.g., popular song patterns like AABA).

11.6.4. Library of Sounds

The library of sounds allows the customization of sounds. The library is composed by samples for each instrument of the General MIDI 1 standard - 128 instruments. These samples were obtained from Project SAM Symphobia[20], Garritan Personal Orchestra[21] and a personal library of Soundfont sounds[22].

11.7. Administrator Interface

The system can be controlled offline through an administrator interface (Figure 11.7.1). The administrator can segment, extract features and classify the segments. We can see the values of some of the extracted features (e.g., average note duration, average time between attacks, tempo, note density, percussion prevalence and key mode), as well as the values of valence and arousal. This interface also allows to carry out some tests of transformation, sequencing and synthesis, but it was mainly developed for segmentation, extraction of features and classification.

11.8. User Interface

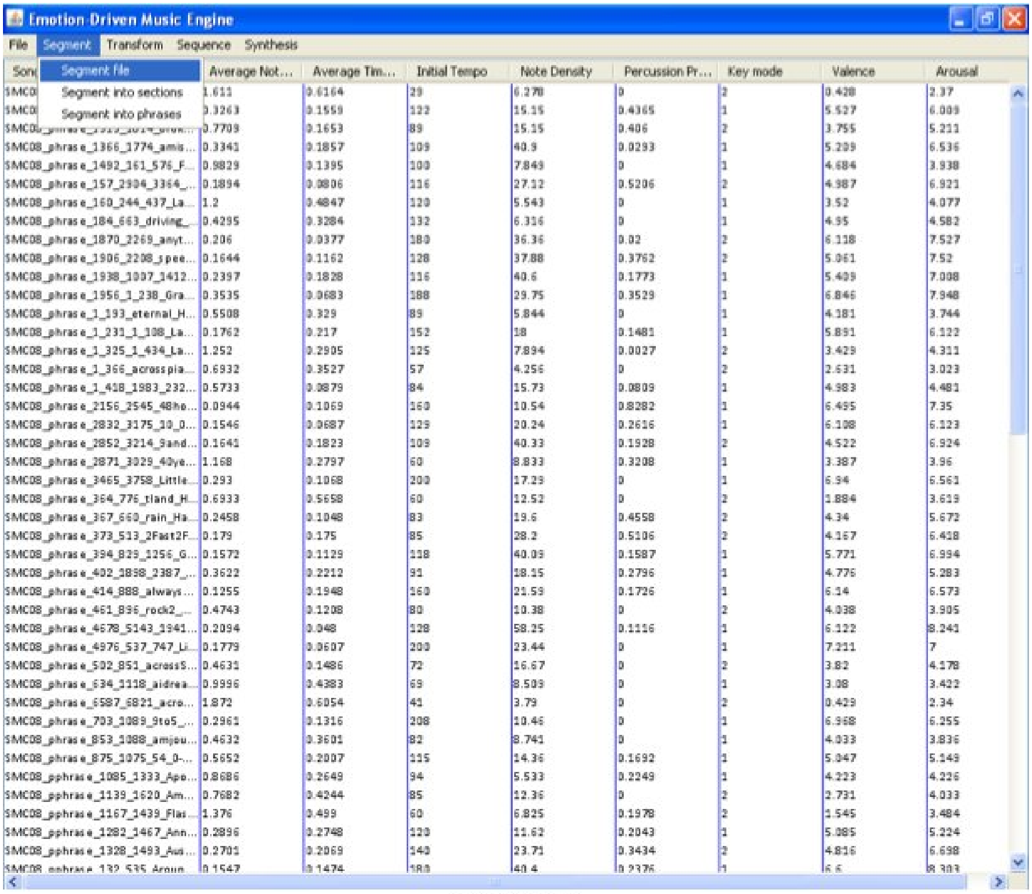
The system can be controlled in real-time through a user interface (Figure 11.8.1) or be driven by an external system providing an emotional specification (Lopez et al., 2010). The input specifies values of valence and arousal. While playing, EDME responds to input changes by quickly adapting the music to a new user-defined emotion.

The user interface serves the purpose of letting the user choose in different ways the desired emotion. The user can type the values of valence and arousal or choose from a list of discrete emotions. It is possible to load several lists of words denoting emotions to fit different uses of the system. For example, Ekman (1999) has a list of generally accepted basic emotions. Russell (1989) and Mehrabian (1980) both have lists which map specific emotions to dimensional values (using 2 or 3 dimensions). Juslin and Laukka (2004) propose a specific list for emotions expressed by music.

[20] http://www.projectsam.com/Products/Symphobia/

[21] http://www.garritan.com/GPO-features.html

[22] http://www.connect.creativelabs.com/developer/SoundFont/Forms/ AllItems.aspx



Off-line user interface

Figure 11.7.1.: Administrator interface of EDME

Another way to choose the emotional state of music is through a graphic representation of the valence-arousal emotional space, based on FeelTrace (Cowie et al., 2000): a circular space with valence dimension in the horizontal axis and the arousal dimension in the vertical axis.

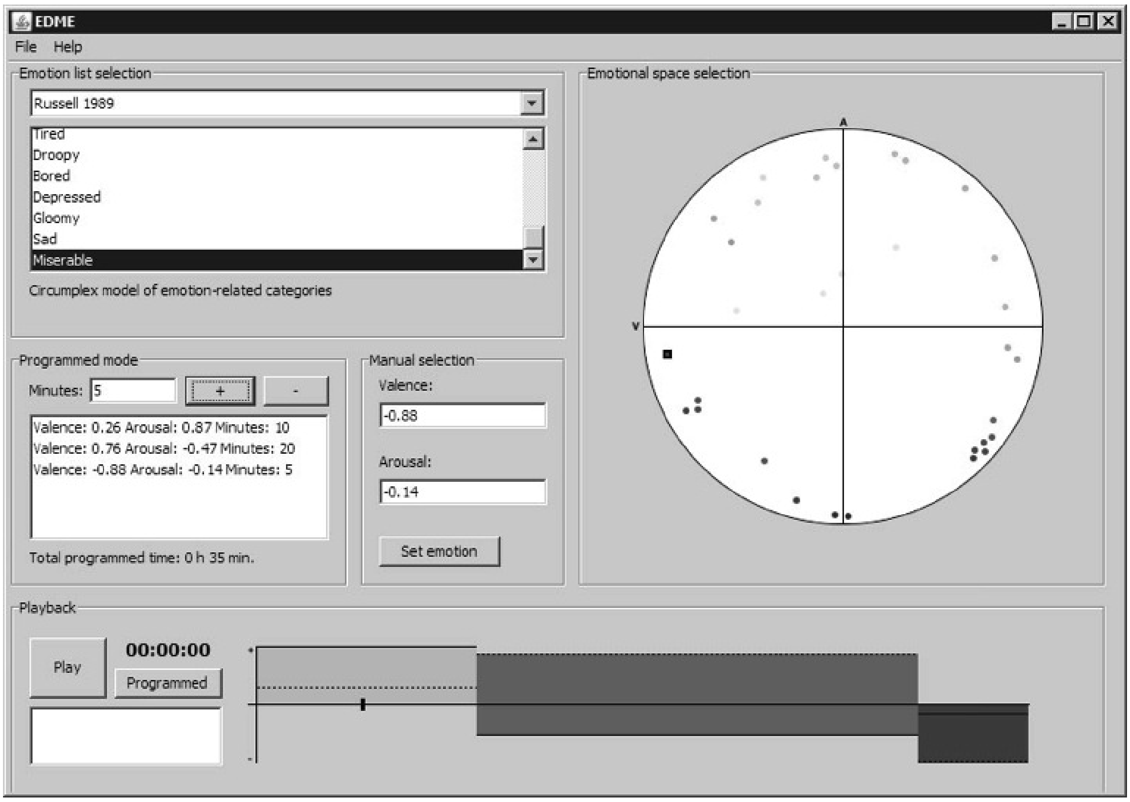


Figure 11.8.1.: User interface of EDME

12. Experiments

The EDME system is composed by four main modules: segmentation, classification, selection and transformation. The segmentation module was tested as was described in section 11.1. Roughly speaking, the selection module has only to calculate Euclidean distances. Because this distance is commonly used, we decided to not waste time in testing other distance metrics, but now come the two principal modules of the system: the classification and the transformation. Both modules are dependent on the auxiliary structure that is the knowledge base. The success of these modules depends on the quality of the knowledge base, more properly in its effectiveness on relating the musical and emotional domains. But we can even highlight the degree of importance of the classification module over the transformation module, and say that this module is the heart of the system. This is particularly true, because it is in the classification module that EDME establishes a bridge between the emotional and the musical dimensions. Therefore, more attention was devoted in this section (and in the experiments) to this module. Further emphasis was put on the identification of the most relevant features to be used by the knowledge base that supports both modules.

12.1. Stages of the experiments

The classification and transformation modules and the knowledge base were refined in three experiments. But before we made the experiments we had an initial phase that consisted in building manually a first version of the knowledge base (Oliveira and Cardoso, 2007) by considering empirical data collected from works of Music Psychology (section 12.3). Figure 12.1.1 presents an overview of the different stages of the initial phase and of the experiments described in this chapter. We carried out three experiments (Oliveira and Cardoso, 2008d,c,b, 2009) conducted via Web to build regression models and to successively refine their set of features and corresponding weights (sections 12.4, 12.6 and 12.7). It is worth noting that the focus of these experiments was in the identification of a small group of features emotionally relevant. We left to the calibration/validation (chapter 15) the identification of the best weights for these features. It is also worth to mention that we were more concerned in finding the set of features, instead of finding the best type of classifier. This approach was followed in other studies (McKinney and Breebaart, 2003), as it seems that in some cases the type of classifier does not influence the classification accuracy, but what seems to influence this accuracy is the feature set being used. This is especially true when the number of features is high as in our case.

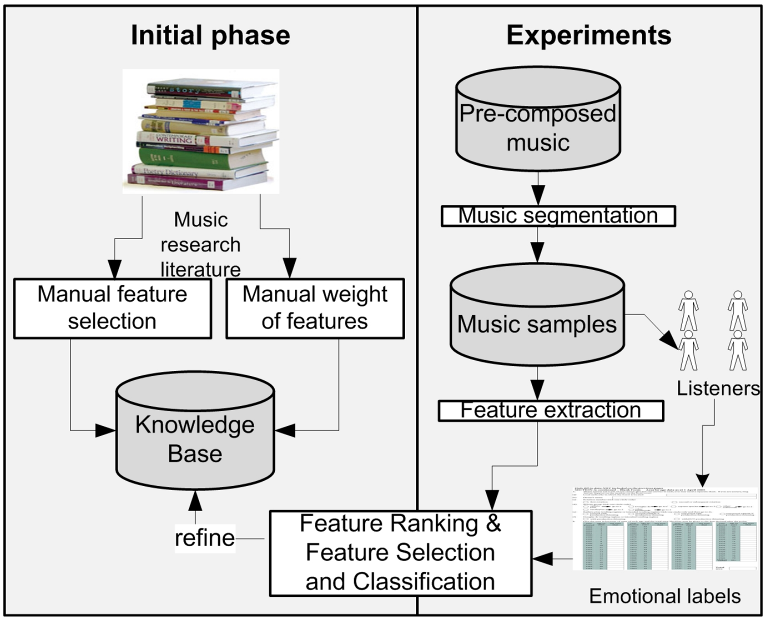


Figure 12.1.1.: Stages of the experiments

In the initial phase we built a first version of a knowledge base by selecting a set of features manually according to what we learned from the literature about their relative importance to emotional expression. We also defined tentative weights for the features in accordance with the literature. Now, we are going to briefly explain the most important steps of each experiment (Figure 12.1.1). As mentioned in section 11.1, pre-composed music consists of standard MIDI files compiled from websites. Each experiment started with the segmentation of the pre-composed music to obtain segments that might express only one kind of emotion. From the large group of obtained segments we selected those that best cover all the bi-dimensional emotional space. This was done by taking into account the classification results obtained with the knowledge base(s) built in the previous experiment(s) (or in the initial phase, in the case of the first experiment). Then, feature extraction algorithms of third party software (McKay and Fujinaga, 2006; Eerola and Toiviainen, 2004; Sorensen and Brown, 2000; Lartillot and Toiviainen, 2007; Cabrera, 1999) were applied to label the segments with music features. Each segment was then made available in web-based questionnaires[23][24][25] (Figure 12.1.2). Each questionnaire was divided into three parts. The first part con-

[23] http://student.dei.uc.pt/%7Eapsimoes/PhD/Music/icmc08/index.html
[24] http://student.dei.uc.pt/%7Eapsimoes/PhD/Music/smc08/index.html
[25] http://student.dei.uc.pt/~apsimoes/PhD/Music/smc09/

sisted in a brief introduction of the work, followed by a description of the content of the questionnaire. It also gives a brief description of the two emotional dimensions to be classified: valence/satisfaction and arousal/activation. The second part consisted of the musical segments and the emotional labels. The third part consisted in personal information as it is the age and gender. So, each segment was classified by human subjects according to two emotional dimensions. Values in the interval [0; 10] were used by the listeners to classify each dimension. Answers from listeners distant more than the mean ± 2*standard deviation were considered as outliers and consequently discarded. The remaining answers were used as emotional labels for the music segments. We obtained, therefore, music features and emotional values for each music segment.

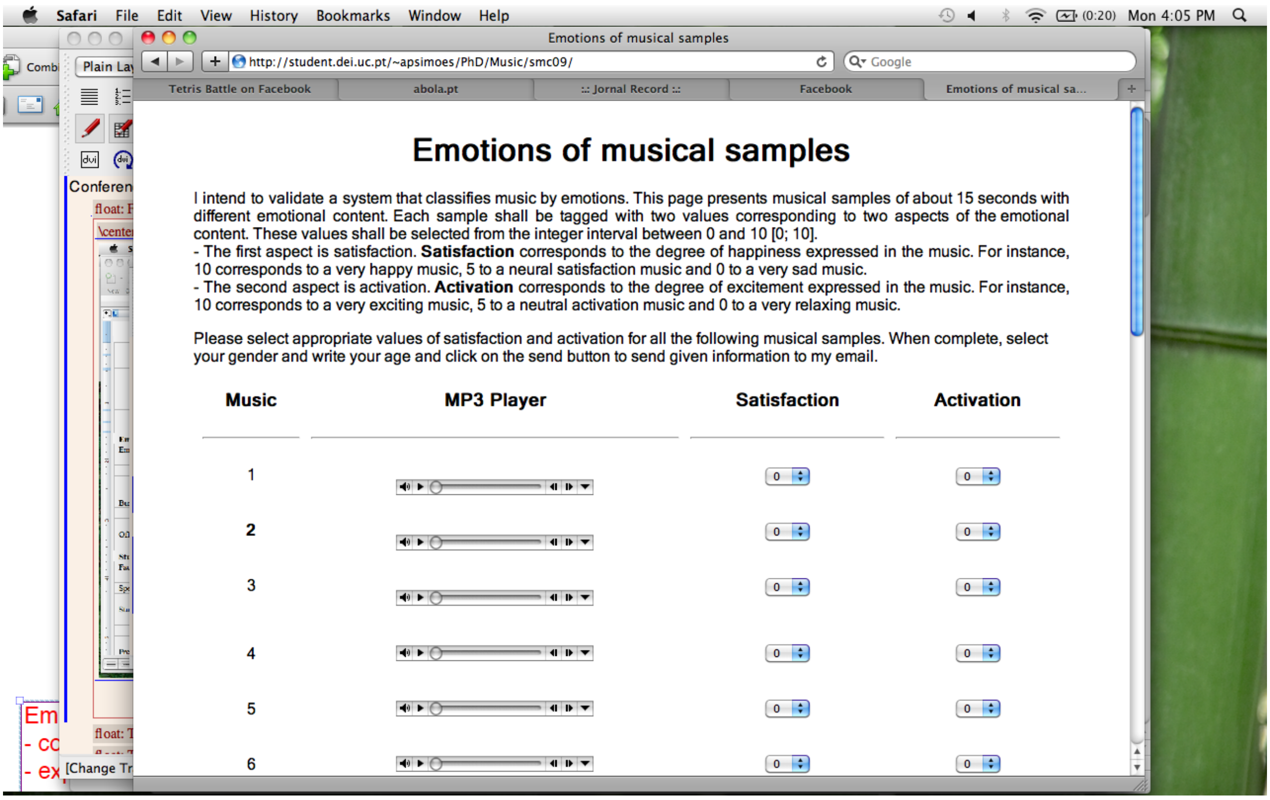


Figure 12.1.2.: Web-based questionnaire for the experiments

The process proceeded in three sequential steps: feature ranking, feature selection and classification. The first step consisted in applying feature ranking to obtain a first group of features that were individually the ones emotionally more relevant. Each feature was ranked individually using the correlation coefficients obtained separately for arousal and valence. The best features in each emotional dimension were the ones with the highest positive/negative coefficients[26]. The second step consisted in applying

[26]In the case of the first experiment, with the highest importance according to the literature review (See Figure 12.3.1 of section 12.3)

feature selection methods to obtain, for each dimension, a smaller group of features that collectively best discriminate the emotional content of music. In general, we applied the best first search method (Witten et al., 1999) on the group of features obtained from the first step. To a better understanding of the relative importance of the categories of the features we decided to group them during the first and second experiment into six categories: instrumental, textural, rhythmic, dynamics, melodic and harmonic. The third step consisted in evaluating the classification performance using this last group of features by applying n-fold cross-validation. As a result, we obtained weights for each feature that contributed to the best performance. We applied 10-fold cross-validation with the best group of features to obtain the classification results, i.e., the correlation coefficients (CC), mean absolute errors (MAE), root mean square errors (RMSE) and the weights of the features. N-fold cross-validation process was made with the help of these measures which were given by the classification models used with WEKA (Witten et al., 1999; Witten and Frank, 2005). Every time we use correlation coefficient (CC) we are referring to the Pearson product-moment correlation coefficient. It is a measure of the linear dependence between two variables X and Y, which gives values between +1 and -1 inclusive. The closer the value is to 1 or -1 the higher is the strength of the dependence between the variables X and Y. The equation for CC is presented:

$$CC = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The mean absolute error (MAE) is used to measure how close predictions are to the results. It is an average of the absolute errors, where f_i is the prediction and y_i is the result. The equation for MAE is presented:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

The root-mean-square error (RMSE) is used to measure the difference between values predicted by a model and the obtained results. It corresponds to the square root of the mean square error, where f_i is the prediction and y_i is the result. The equation for RMSE is presented:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)}{n}}$$

12.2. Overview of the experiments

12.2.1. First experiment

Most of the time dedicated to the three experiments had the objective of working towards automatic bi-dimensional classification of MIDI music by emotional content. The first experiment (entitled ”Preliminary evaluation of the classification”) was totally committed to this purpose. It was the first step to accomplish this objective. In this experiment we obtained the first emotionally-relevant set of features (after proceeding to feature ranking and feature selection), and with this set of features we obtained the first results of classification.

In the beginning of the first experiment, we performed ad-hoc comparisons between a small group of classifiers by using the experimental data (Oliveira and Cardoso, 2008d) and verified that Support Vector Machine regression (Witten et al., 1999) obtained the best results, which took us to use this classifier to calculate all the classification results presented in this chapter.

12.2.2. Second experiment

The second experiment was divided into two parts with different objectives. The first part of the second experiment (entitled ”Extended evaluation of the classification module”) had the same objective of the first one and consisted solely in extending the first experiment by increasing the number of music pieces, the number of extracted features and the number of listeners that answered to the web-based questionnaire. More details about it are left to its respective section.

The second part of the second experiment (entitled ”Analysis of audio features”) introduced a novelty that consisted in analysing audio features. Till then, we had only analysed MIDI features... The main reason why we went from the MIDI to audio domain was because the synthesis module works with the audio domain. It is in this module were we select audio samples for each of the notes of the MIDI file. Knowing the importance of the timbre of the instruments in the emotional domain, we intended to identify audio features emotionally-relevant that could lead to the selection of audio samples guided by the emotional relevance of their features (spectral dissonance, spectral sharpness, etc.).

12.2.3. Third experiment

The third experiment was also divided into three parts with different objectives. The first part of this experiment (entitled ”Improvement of the classification module”) was made with the objective of having more data that could help us in finding the (MIDI and audio) features emotionally more relevant.

The second part of the third experiment (entitled ”Evaluation of the transformation module”) was developed with the objective of verifying the existent correlation between the variation of five features emotionally-relevant features: tempo, pitch register, musical scales, instruments and articulation.

We prepared the third part of the third experiment (entitled ”Melodic analysis”) in order to verify the importance of features of the melody in the discrimination of the emotions. By reducing the amount of data being analysed, and focusing only on the melody, we were expecting to find features with a value of correlation higher than those values of correlation obtained until then.

12.3. Initial Phase - Manually Built Knowledge Base

The contents of this initial phase were published in the proceedings of the 2007 Affective and Intelligent Interaction conference (Oliveira and Cardoso, 2007).

The first version of the knowledge base was built solely with the help of the theories, algorithms, models, frameworks and empirical results found on works of Music Psychology (section 6) (Oliveira and Cardoso, 2007). The objective was to have some rough way of classifying MIDI segments in order to select from a large set of segments those that could cover reasonably the classification space, to be used in further experiments. The features were selected manually, according to what we learned from the literature about their relative importance to emotional expression. We also defined weights for the features in accordance with the literature. Then, we looked for existing software that might extract those features. We could find extractors (McKay and Fujinaga, 2006; Eerola and Toiviainen, 2004; Sorensen and Brown, 2000; Cabrera, 1999; Lartillot and Toiviainen, 2007) for most of the relevant ones (See Figure 12.3.1). A positive or negative tentative weight was defined according to the positive or negative effect and degree of influence of each of the features in each of the dimensions. Here is an example: consider the weight x in $\Re : x \in [-1, 1]$. We know that pitch register has a small direct relationship with the valence of music, so a weight of 0.2 is given; and tempo has a great direct relationship with the arousal of music, so a weight of 1.0 is given.

Musical Feature		Happy music	Sad music	Activating music	Relaxing music
Instruments	timbre	piano, strings instruments, few harmonics, bright, percussion instruments	timpani, violin, woodwind instruments, few harmonics, dull, harsh	brass, low register instruments, timpani, harsh, bright, percussion instruments	woodwind instruments, few harmonics, soft
Dynamics	loudness	high	low	high	low
	articulation	staccato	legato	staccato	legato
	articulation variab.	large	small	-	-
	sound variability	low	-	-	-
Rhythm	tempo	fast	slow	fast	slow
	note density	high	low	high	low
	note duration	small	large	small	large
	tempo variability	small	-	-	-
	duration contrast	sharp	soft	-	-
Melody	pitch register	high	low	-	-
	pitch repetition	high	low	high	low
	stable/ expect notes	accented	-	-	-
	unstab/ unexp notes	accented	-	-	-
Harmony	harmony	consonant	dissonant	complex, dissonant	-
	scale	major, pentatonic	minor, diminished	-	-

Figure 12.3.1.: Features of happy, sad, activating and relaxing music

12.4. First Experiment - Preliminary Evaluation of the Classification Module

The contents of this experiment were published in the proceedings of the 2008 International Computer Music Conference (Oliveira and Cardoso, 2008d).

Features per Category	
Instrumental	Textural
Acoustic Guitar Fraction Brass Fraction Electric Guitar Fraction Electric Instrument Fraction Number of Pitched Instruments Number of Unpitched Instruments Orchestral Strings Fraction Percussion Prevalence Saxophone Fraction String Ensemble Fraction String Keyboard Fraction Var. of Note Prev. of Pitched Instruments Var. Note Prev. of Unpitched Instruments Violin Fraction Woodwinds Fraction Slap Bass Fraction Muted Guitar Fraction Harpsichord Fraction	Average Number of Independent Voices Importance of Loudest Voice Maximum Number of Independent Voices Melodic Intervals in Lowest Line Range of Highest Line Relative Note Density of Highest Line Relative Range of Loudest Voice Var. of Number of Independent Voices Voice Equality - Dynamics Voice Equality - Melodic Leaps Voice Equality - Note Duration Voice Equality - Number of Notes Voice Equality - Range Voice Separation
Rhythmic	Pitch
Average Note Duration Average Time Between Attacks Avg. Time Between Attacks For Each Voice Avg. Var. Time Bet. Attacks Each Voice Changes of Meter Comb. Strength Two Strong. Rhyth. Pulses Compound Or Simple Meter Harmonicity Two Strong. Rhythmic Pulses Initial Tempo Maximum Note Duration Minimum Note Duration Note Density Number of Moderate Pulses Number of Relatively Strong Pulses Number of Strong Pulses Polyrhythms Quintuple Meter Rhythmic Looseness Rhythmic Variability Second Strongest Rhythmic Pulse Staccato Incidence Strength Second Strongest Rhythmic Pulse Strength of Strongest Rhythmic Pulse Strength Ratio Two Strong. Rhyth. Pulses Strongest Rhythmic Pulse Triple Meter Variability of Note Duration Variability of Time Between Attacks	Average Range of Glissandos Dominant Spread Glissando Prevalence Importance of Bass Register Importance of High Register Importance of Middle Register Interval Between Strongest Pitch Classes Interval Between Strongest Pitches Most Common Pitch Class Most Common Pitch Class Prevalence Most Common Pitch Most Common Pitch Prevalence Number of Common Pitches Pitch Class Variety Pitch Variety Primary Register Quality Range Relative Strength of Top Pitch Classes Relative Strength of Top Pitches Strong Tonal Centres Vibrato Prevalence
Melodic	Dynamics
Amount of Arpeggiation Average Melodic Interval Chromatic Motion Direction of Motion Distance Bet. Common Melodic Intervals Duration of Melodic Arcs Melodic Fifths Melodic Octaves Melodic Thirds Melodic Tritones Most Common Melodic Interval Most Common Melodic Interval Prevalence Number of Common Melodic Intervals Relative Strength Most Common Intervals Repeated Notes Size of Melodic Arcs Stepwise Motion	Average Note To Note Dynamics Change Overall Dynamic Range Variation of Dynamics Variation of Dynamics In Each Voice

Table 12.1.: Features extracted with JSymbolic (McKay and Fujinaga, 2006) that were analysed in the first experiment

12.4.1. Objective

After building a first version of the knowledge base grounded solely on literature of Music Psychology we set the objective of making this knowledge base in an automatic way without worrying about defining manually what would be the right features and their respective weights. This step from the manual to the automatic building of the knowledge base gave us much more confidence on its reliability. So, the first experiment was the first step towards automatic bi-dimensional classification of MIDI music by emotional content.

We tested the hypothesis that there is only a small group of features emotionally-relevant from a larger group of various features. We worked to obtain the first two sets of features emotionally relevant to valence and arousal. We intended to identify the emotional relevance of the 2 harmonic features (key mode and key) available from MIDI Toolbox (Eerola and Toiviainen, 2004) and 103 one-dimensional features available from JSymbolic (McKay and Fujinaga, 2006): instrumentation (18 features), texture (14 features), rhythm (28 features), dynamics (4 features), pitch (22 features) and melody (17 features). Table 12.1 shows the features extracted with JSymbolic. Detailed description of each one can be seen in McKay's thesis (McKay, 2004).

12.4.2. Method

We selected 9 MIDI files of western tonal music (pop and r&b genres) from a large database of pre-composed music of various genres. These files were selected based on its musical quality. The genres were randomly chosen from a group that included besides these, others like rap, rock and classical. Selected files went through the processes of segmentation and feature extraction. From this resulted a group of 412 segments labeled with musical features. The regression models built in the initial phase were used to classify each segment with an appropriate emotional label. From this group of 412 segments emotionally classified, we selected 16 segments to be used to update the regression models. This selection process was grounded on the purpose of covering all the bi-dimensional emotional space. Listeners were invited to classify the selected segments through the use of several mailing-lists (intended for discussion of subjects like music theory, music therapy and others).

12.4.3. Data

Data consists of the selected musical segments and obtained emotional answers from the listeners.

12.4.3.1. Music

The 16 musical segments lasted between 20 and 60 seconds and are available in this link[27].

12.4.3.2. Emotional answers

53 listeners answered to the questionnaire: 33 male and 20 female with ages between 14 and 56 years old (mean of 34, standard deviation of 12). They had background in informatics, technology and music. We calculated the mean and standard deviation for the emotional answers obtained in the questionnaire, as is shown in Figure 12.4.1. Mean and standard deviations were computed first between listeners, and then averaged over segments. We measured the agreement of the listeners on the emotional content of the music using the Cronbach's Alpha and obtained a value of 79.49% for arousal and a value of 78.19% for valence. These values give us an acceptable (if not good) internal consistency [28] of the obtained emotional answers. .

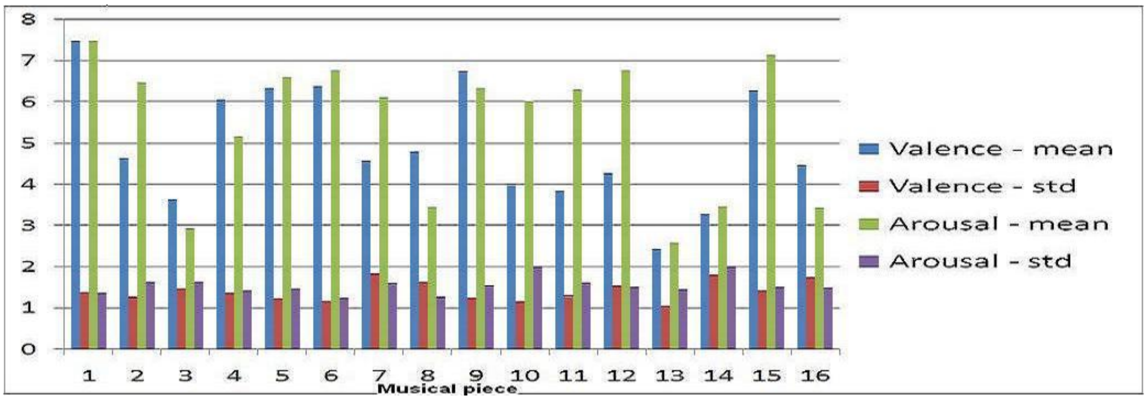


Figure 12.4.1.: Mean and standard deviations of the emotional responses in the first experiment

12.4.4. Results

The results of this experiment are divided into two subsubsections. One that consists of feature ranking and the other that consists of feature selection and classification.

[27] http://student.dei.uc.pt/%7Eapsimoes/PhD/Music/icmc08/index.html
[28] http://en.wikipedia.org/wiki/Cronbach's_alpha

12.4.4.1. Feature Ranking

We calculated individually the correlation coefficient between each feature and the two emotional dimensions. Table 12.2 presents, for each category, the features with the highest correlation with valence[29]. We can highlight rhythmic (e.g, variability of note duration, note density and polyrhythms), instrumentation (e.g., string ensemble fraction), melodic (e.g., melodic tritones) and textural features (e.g., average number of independent voices) as the most relevant ones to valence.

Category	Feature	CC
Instrumental	String Ensemble Fraction	-0.50
	Saxophone Fraction	0.36
	Electric Guitar Fraction	0.31
Textural	Average Number of Independent Voices	0.40
	Variability Number of Independent Voices	0.38
Rhythmic	Variability of Note Duration	-0.66
	Note Density	0.57
	Polyrhythms	-0.54
	Average Note Duration	-0.52
	Average Time Between Attacks	-0.52
Dynamics	Variation of Dynamics of Each Voice	0.22
Melodic	Melodic Tritones	0.47
	Most Common Melodic Interval Prevalence	-0.37
	Relative Strength Common Intervals	0.37
	Relative Strength of Top Pitches	0.37
	Relative Strength of Top Pitch Classes	0.35
	Importance of Middle Register	0.33
Harmonic	Key mode	-0.23

Table 12.2.: Best features of each category - valence

The corresponding results for arousal are represented in Table 12.3. We can highlight rhythmic (e.g., average time between attacks, average note duration and note density), instrumentation (e.g., number of unpitched instruments and percussion prevalence), melodic features (e.g., importance of high register and primary register), textural (e.g., range of highest line) and dynamics (e.g., variation of dynamics) as the most relevant ones to arousal.

[29]The meaning of each musical feature present in this table and in the following tables is described in the Glossary, section A.2

Category	Feature	CC
Instrumental	Number of Unpitched Instruments	0.60
	Percussion Prevalence	0.53
Textural	Range of Highest Line	-0.51
	Variability Number of Independent Voices	0.45
Rhythmic	Average Time Between Attacks	-0.73
	Average Note Duration	-0.72
	Note Density	0.68
	Variability of Time Between Attacks	-0.65
	Strength of Strongest Rhythmic Pulse	-0.61
Dynamics	Variation of Dynamics	0.49
	Average Note to Note Dynamics Change	0.46
Melodic	Importance of High Register	-0.55
	Primary Register	-0.50
	Stepwise Motion	-0.40
	Most Common Melodic Interval Prevalence	0.36
Harmonic	Key mode	-0.13

Table 12.3.: Best features of each group - arousal

12.4.4.2. Feature Selection and Classification

We applied the best first search method Witten et al. (1999) on the features with the highest ranking (Tables 12.2 and 12.3) to select the set of features that better discriminate the emotional content of music. We made a compromise between the number of features and the quality of the results. Then, we applied 10-fold cross-validation on the most discriminant features with the results presented in Table 12.4. From the analysis of this table, we have the two best features in the classification of valence with similar weights, and average note duration, average time between attacks and importance of high register with the highest weights in the classification of arousal.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.76	0.75	0.91	Average time between attacks	-0.50
				Variability of note duration	-0.55
Arousal	0.77	0.86	1.06	Average note duration	-0.48
				Average time between attacks	-0.35
				Importance of high register	-0.45
				Note density	0.09

Table 12.4.: Results of 10-fold cross-validation for valence and arousal – first experiment

12.4.5. Discussion

We presented a preliminary experiment that undertook music emotion classification as a regression problem. SVM regression obtained the best results in the classification of the dimensions of valence and arousal. N-fold cross-validation results using the coefficient of correlation showed that the performance of the predictive models for classification of arousal (0.77) and for the classification of valence (0.76) are similar and positive.. Rhythmic features proved to be very important to valence and arousal (e.g., average time between attacks, note density and average note duration). Melodic features (e.g., importance of high register and primary register) were also important to classify arousal.

Regarding the instrumentation not too much could be concluded because of the lack of music pieces with similar instruments. Moreover, more instrumentation features were needed (e.g., spectral sharpness, spectral dissonance and analysis of the frequency spectrum of samples). It was also important to implement some features of dynamics (e.g., average loudness) for arousal prediction and harmony (e.g., spectral consonance) for valence. Concerning the texture and melodic features there was the need of more tests. Therefore, it was our goal to extend this study to a statistical significant number of music files.

12.5. Second Experiment - Extended Evaluation of the Classification Module

The second experiment was divided into two parts. The first part is described in this section; the second part is described in section 12.6. This part consisted in an extended evaluation of the classification module and its contents were published in the proceedings of the 2008 Sound and Music Computing Conference (Oliveira and Cardoso, 2008c).

Features per Category	
Instrumental (JSymbolic)	Textural (MIR Toolbox)
Note Prevalence of Acoustic Grand Piano Note Prevalence of Bright Acoustic Piano Note Prevalence of Electric Grand Piano ... Note Prevalence of Helicopter Note Prevalence of Applause Note Prevalence of Gunshot Time Prevalence of Acoustic Grand Piano Time Prevalence of Bright Acoustic Piano Time Prevalence of Electric Grand Piano ... Time Prevalence of Telephone Ring Time Prevalence of Helicopter Time Prevalence of Applause Time Prevalence of Gunshot Slap Bass Fraction Harpsichord Fraction	Spectral Texture MFCC 1 Spectral Texture MFCC 2 Spectral Texture MFCC 3 Spectral Texture MFCC 4 Spectral Texture MFCC 5 Spectral Texture MFCC 6 Spectral Texture MFCC 7 Spectral Texture MFCC 8 Spectral Texture MFCC 9 Spectral Texture MFCC 10 Spectral Texture MFCC 11 Spectral Texture MFCC 12 Spectral Texture MFCC 13
Rhythmic (MIDI Toolbox and JMusic)	Melodic (MIDI Toolbox and JMusic)
Average Meter Accent Synchrony Maximum Meter Accent Synchrony Concurrent Onsets Average Duration Accent Meter Average Metrical Hierarchy Variability of Events Onset Autocorrelation Consecutive Identical Rhythms Distinct Rhythm Count Repeated Rhythmic Value Density Rhythm Range Same Direction Interval Count Syncopation	Average Melodic Complexity Maximum Melodic Complexity Average Melodic Originality Maximum Melodic Originality Average Melodiousness Maximum Melodiousness Average Melodic Accent Maximum Melodic Accent Average Melodic Attraction Maximum Melodic Attraction Average Melodic Mobility Maximum Melodic Mobility Average IR Narmour Maximum IR Narmour Average Melodic Tessitura Maximum Melodic Tessitura Big Jump Big Jump Followed By Step Back Climax Position Climax Strength Consecutive Identical Pitches Leap Compensation Melodic Direction Stability Overall Pitch Direction Repeated Pitch Density
Harmonic (JMusic)	
Dissonance	

Table 12.5.: Features analysed in the second experiment that were not analysed in the first experiment

12.5.1. Objective

The second experiment had the same objective as the first one. In this part of the second experiment we tried to overcome two problems: the limited number of music files being classified and the limited number of features extracted from the music. The first problem was surmounted by extending the number of musical files (from 16) to a statistical significant number (96). The second problem was overcome by extracting a larger number of features (from 105 to 414); we extracted the 105 features of the first experiment plus other 309 features not extracted in the first experiment by using other third-party software. Thus, the second experiment was the second step towards automatic bi-dimensional classification of MIDI music by emotional content. This experiment was dedicated to the refinement of the knowledge base built in the first experiment.

Just as in the first experiment, we came up with the hypothesis that there is a small amount of features that may predict arousal/valence. We worked to refine the two set of features relevant to valence and arousal. We intended to identify the emotional relevance of 148 unidimensional features and 3 multidimensional ones (note prevalence of instruments, time prevalence of instruments and spectral texture) that were categorized into six groups: instrumentation (20), texture (15), rhythm (42), dynamics (4), melody (64) and harmony (3). We used JSymbolic (McKay and Fujinaga, 2006), MIDI Toolbox (Eerola and Toiviainen, 2004), JMusic (Sorensen and Brown, 2000) and MIR Toolbox (Lartillot and Toiviainen, 2007). Special attention was devoted to the identification of the emotional relevance of new features and important ones from the first experiment: the importance (volume*time) of 13 Mel Frequency Cepstral Coefficients[30] of each sample used to synthesize musical instruments, the prevalence (by note or time) of specific groups and individual instruments, tempo, notes density, duration of notes, rhythmic variability, melodic complexity, prevalence of repeated notes, prevalence of the most common melodic intervals, pitch classes and pitches, and mode (major or minor). Table 12.5 shows the ”new” features not present in Table 12.1. In the instrumental category note prevalence of instruments and time prevalence of instruments are multidimensional features, each one with the size that corresponds to the number of standard pitched instruments of the General MIDI. To avoid a longer table, we only mention the first and last three instruments separated by ”...” to be representative of all the 128 instruments. Detailed description of each feature can be consulted in the reference (McKay, 2004) as in the case of JSymbolic features, or on the following links as in the cases of MIDI Toolbox[31] , JMusic[32] and MIR Toolbox[33].

[30] http://en.wikipedia.org/wiki/Mel-frequency_cepstrum
[31] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/miditoolbox/Manual
[32] http://ses.library.usyd.edu.au/bitstream/2123/6205/1/abrown.pdf
[33] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/MIRtoolboxUsersGuide1.3.3

12.5.2. Method

The steps of the method of this experiment were designed in order to accomplish the objectives of the two parts of this experiment. We selected 90 MIDI files of western tonal music (film music genre) from a large database of pre-composed music of various genres. These files were selected based on their musical quality. We selected this genre, because of its closer connection with the emotional dimension, as it is proved by the fact that the composers usually guide the process of producing music by an emotional specification subjacent to the film scenes. Another reason which led to this selection was to diversify the genre of music being analysed, as we want a system that works with all the genres of western tonal music. Selected files were put through the processes of segmentation and feature extraction. From this resulted a group of 5238 segments labeled with musical features. The regression models built in the first experiment were used to classify each segment with an appropriate emotional label. From this group of 5238 segments emotionally classified, we selected 96 segments to be used to update regression models. This selection process was grounded on the purpose of covering all the bi-dimensional emotional space. Listeners were invited to classify a subgroup of 16 segments from the group of 96 selected segments through the use of several mailing-lists (intended for discussion of subjects like music theory, music therapy and others).

12.5.3. Data

Data consists of the selected musical segments and obtained emotional answers from the listeners.

12.5.3.1. Music

The 96 musical segments lasted between 20 and 60 seconds and are available in this link [34]. We extended the first experiment by increasing the number of music pieces (from 16 to 96) and features (from 105 to 414).

12.5.3.2. Emotional Answers

80 listeners answered to the questionnaire: 34 male and 46 female aged between 17 and 69 years old (mean of 38, standard deviation of 11). They had background in informatics, technology and music. We calculated the mean and standard deviation for the

[34] http://student.dei.uc.pt/%7Eapsimoes/PhD/Music/smc08/index.html

emotional responses obtained in the questionnaire, as is shown in Figures 12.5.1 and 12.5.2. Mean and standard deviations were computed first between listeners, and then averaged over segments. We measured the agreement of the listeners on the emotional content of the music using the Cronbach's Alpha and obtained a value of 7.44% for arousal and a value of 25.16% for valence. These values give us an unacceptable internal consistency [35] of the obtained emotional answers. This may be explained by the fact that each listener answered to its own subgroup of 16 segments (from the group of 96 selected segments) - as explained in subsection 12.5.2. Nobody answered to the same subgroup of musical segments as this subgroup was randomly chosen.

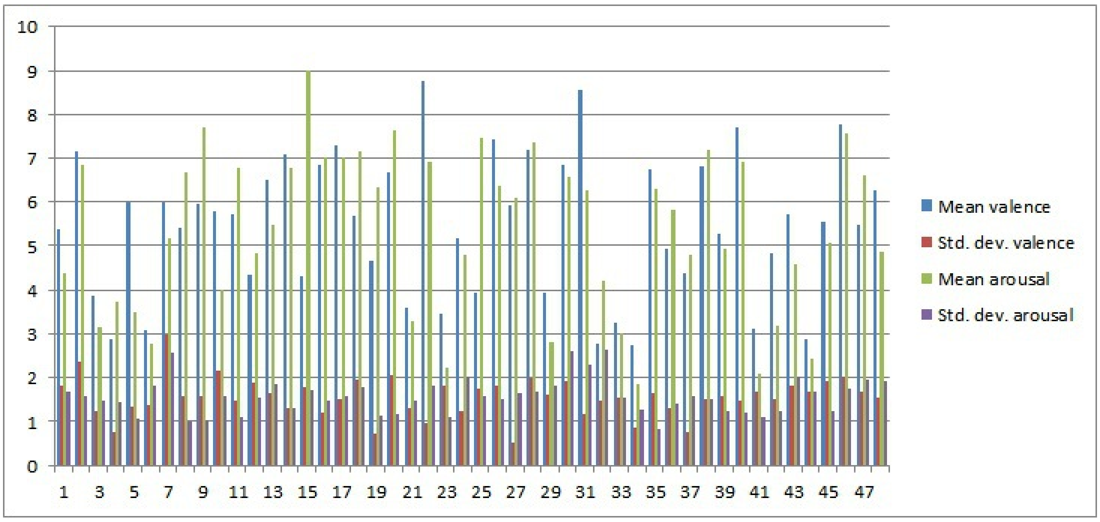


Figure 12.5.1.: Mean and standard deviations of the first 48 emotional responses in the second experiment

[35] http://en.wikipedia.org/wiki/Cronbach's_alpha

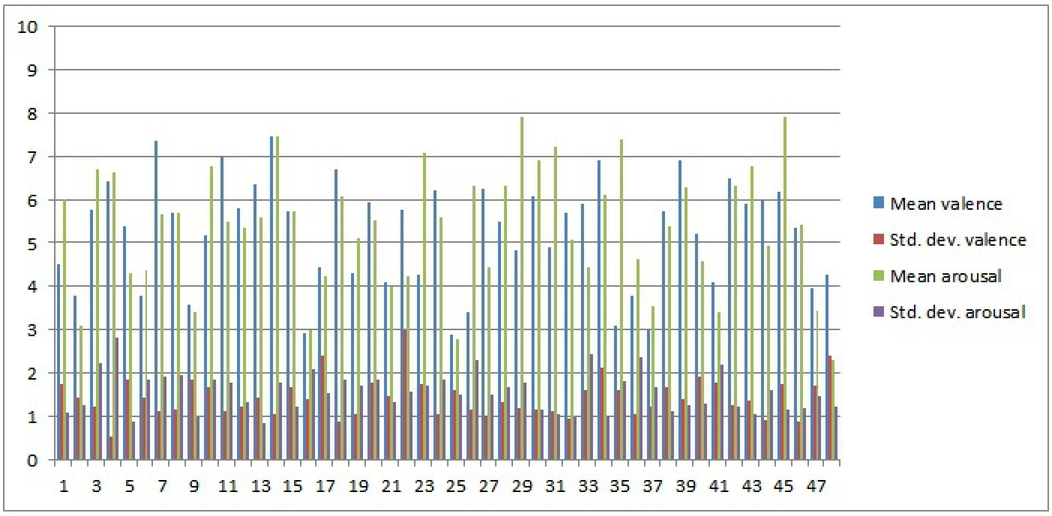


Figure 12.5.2.: Mean and standard deviations of the second 48 emotional responses in the second experiment

12.5.4. Results

The results of this experiment are divided into two subsubsections. One that consist of feature ranking and the other that consists of feature selection and classification.

12.5.4.1. Feature Ranking

We calculated individually the correlation coefficient between each feature and the two emotional dimensions. The features with the highest correlation with valence in each category are presented in Table 12.6. We can highlight rhythmic (e.g, tempo, average note duration and average time between attacks), harmonic (e.g., key mode and key), instrumentation (e.g., note prevalence of muted guitar) and melodic features (e.g., climax position) as the most relevant ones to the valence. We started by applying feature selection algorithms (Witten et al., 1999) to reduce the number of features and to improve classification results. From this resulted a group of 26 features. We applied 8-fold cross validation with these features and obtained a correlation coefficient of 0.81. After this, we selected manually the best group of features to know the most important features in the stage of selection, but also for the stage of transformation. From this resulted a group of four features. The correlation coefficient obtained with the application of 10-fold cross validation with these features is available in Table 12.8. We also determined the correlation coefficient (0.57) obtained by using the regression models of the first experiment.

Category	Feature	CC
Instrumental	Note Prevalence Muted Guitar	0.37
	Electric Instrument Fraction	0.34
	Note Prevalence Steel Drums	0.33
	Time Prevalence Marimba	0.31
	Note Prevalence Fretless Bass	0.31
	Note Prevalence Timpani	-0.27
	Electric Guitar Fraction	0.23
	String Ensemble Fraction	-0.22
Textural	Spectral Texture MFCC 4	0.23
	Spectral Texture MFCC 6	0.22
	Spectral Texture MFCC 7	0.21
	Number of Unpitched Instruments	0.20
Rhythmic	Tempo	0.63
	Average note duration	-0.49
	Average time between attacks	-0.48
	Strength of strong. rhythmic pulse	-0.42
	Variability of note duration	-0.42
	Note density	0.40
	Strength of two strong. rhythmic pulses	-0.37
	Variability of time between attacks	-0.36
	Number of relatively strong pulses	0.30
	Distinct rhythm count	0.29
	Rhythmic variety	-0.28
	Strength sec. strong. rhythmic pulse	-0.25
	Strongest rhythmic pulse	0.20
Dynamics	Staccato incidence	0.15
Melodic	Climax position	0.32
	Average melodic complexity	0.24
	Interval strong. pitch classes	0.20
	Dominant spread	0.20
Harmonic	Key mode	-0.43
	Key	-0.37

Table 12.6.: Best features of each group - valence

Table 12.7 presents for each category the features with the highest correlation with arousal. We can highlight rhythmic (e.g., average note duration, note density and variability of note duration), melodic (e.g., climax position and average melodic complexity) and dynamics (e.g., staccato incidence) as the most relevant ones to the arousal. We started by applying feature selection algorithms (Witten et al., 1999) to reduce the number of features and to improve classification results. From this resulted a group of 23 features. We applied 8-fold cross validation with these features and obtained a correlation coefficient of 0.84. After this, we manually selected the best group of features to know the most important features in the stage of selection, but also for the stage of transformation. From this resulted a group of four features. The correlation coefficient obtained with the application of 10-fold cross validation with these features is available in Table 12.8. We also determined the correlation coefficient (0.77) obtained by using the regression models of the first experiment.

Category	Feature	CC
Instrumental	Electric instrument fraction	0.28
	String ensemble fraction	-0.27
	Note prevalence english horn	-0.26
	Number of unpitched instruments	0.25
	Note prevalence flute	-0.25
	Brass fraction	0.25
	Note prevalence orchestra hit	0.22
	Electric guitar fraction	0.21
Textural	Spectral texture MFCC 2	0.28
	Variab. prevalence unpitched instruments	0.25
	Spectral texture MFCC 4	0.24
Rhythmic	Average note duration	-0.68
	Note density	0.63
	Variability of note duration	-0.57
	Tempo	0.55
	Average time between attacks	-0.55
	Variability of time between attacks	-0.54
	Average duration accent	-0.53
	Strength strongest rhythmic pulse	-0.47
	Number of relatively strong pulses	0.43
	Strength two strong. rhythmic pulse	-0.41
	Polyrhythms	-0.38
Dynamics	Staccato incidence	0.35
Melodic	Climax position	0.45
	Average melodic complexity	0.38
	Consecutive identical pitches	0.37
	Climax strength	-0.33
	Repeated notes	0.32
	Most common pitch class prevalence	0.31
	Relative strength of top pitch classes	-0.30
	Amount of arpeggiation	0.29
	Same direction interval	0.27
	Repeated pitch density	0.24
Harmonic	Key mode	-0.22

Table 12.7.: Best features of each group - arousal

12.5.4.2. Feature Selection and Classification

We applied the best-first search method Witten et al. (1999) on the features with the highest ranking (Tables 12.6 and 12.7) to select the set of features features emotionally more discriminant. We made a compromise between the number of features and the quality of the results. Then, we applied 10-fold cross-validation on the most discriminant features with the results presented in Table 12.8. From the analysis of this table, we have average note duration, tempo and note density with the highest weights in the classification of valence, and average note duration with the highest weight in the classification of arousal.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.70	0.85	1.04	Average note duration	0.31
				Tempo	-0.48
				Key mode	-0.18
				Note density	0.34
Arousal	0.77	0.84	1.04	Average note duration	-0.84
				Tempo	0.42
				Note density	0.41

Table 12.8.: Results of 10-fold cross-validation for valence and arousal – second experiment

12.5.5. Discussion

We were expecting that the importance (volume*time) of the 13 Mel Frequency Cepstral Coefficients (van de Laar, 2006) of each sample used to synthesize musical instruments and that the prevalence (by note or time) of specific groups and individual instruments would have a higher relevance on the emotional discrimination of the musical output. However, rhythmic, harmonic and melodic features seemed to have a higher importance on the emotional discrimination. Nonetheless, instrumentation features like the prevalence of muted guitar still have an emotional relevance, for example, for valence as it is expressed by the correlation coefficient of 0.37.

We presented an extension of the first experiment that undertook music emotion classification as a regression problem. SVM regression obtained the best results in the classification of the dimensions of valence and arousal. Validation results using the coefficient of correlation confirmed that the classification of arousal (0.84) is easier than the classification of valence (0.81). Rhythmic (e.g., tempo, note density and average/variability of note duration) and melodic (e.g., climax position and melodic complexity) features proved to be very important to valence and arousal. Harmonic (e.g., key mode) and dynamics features (e.g., staccato incidence) were also important to classify, respectively, the valence and arousal. As a matter of curiosity we calculated the correlation coefficient between valence and arousal and obtained a value of 0.63, which indicates that there is some colinearity amongst these two dimensions.

With similar goals to (Kuo et al., 2005; Muyuan et al., 2004), we developed a knowledge base with relations between music features and emotions, Kuo et al. developed an affinity graph and Muyuan and Naiyao a SVM classifier. We used continuous dimensions (valence and arousal) instead of discrete emotions (Kuo et al., 2005; Muyuan et al., 2004). The results of our model (81% for valence and 84% for arousal) surpass the results of Kuo et al. (80%) and Muyuan and Naiyao (70%) when using a higher number of features (>20).

With these satisfactory results, we felt ready to move to the third experiment of our work, which consisted in the transformation of the emotional content of selected music to approximate even further its emotional content to an intended emotion.

12.6. Second Experiment - Analysis of Audio Features

This section describes the second part of the second experiment. This part consisted in the analysis of audio features and its contents were presented in the 2008 Portuguese Audio Engineering Society Conference (Oliveira and Cardoso, 2008b).

12.6.1. Objective

This part of the second experiment intended to understand the importance of audio features in the emotional expression, as well as to understand their relation with emotionally-relevant MIDI features. The data of the second experiment (described in subsection 12.5.3) was used to analyse the importance of 18 audio features. These features were extracted with MIR Toolbox (Lartillot and Toiviainen, 2007) and Psysound Toolbox (Cabrera, 1999). This was done with the objective of identifying the audio features emotionally more relevant for: the selection of instruments in the synthesis module (subsection 11.5.3); and the analysis of the spectral characteristics of the musical audio output. We also worked on bridging the gap between the MIDI and audio domains, by analysing the influence of MIDI features on audio features.

12.6.2. Method

The method used in this part of the experiment is described in subsection 12.5.2.

12.6.3. Data

Data used in this part of the experiment is described in subsection 12.5.3.

12.6.4. Results

The results of this experiment are divided into two subsubsections. One that consist of feature ranking and the other that consists of feature selection and classification.

12.6.4.1. Feature Ranking

We calculated the correlation between the 18 audio features and valence (Table 12.9) and arousal (Table 12.10). In bold font we have the features with the highest correlation coefficients.

Audio feature	CC
Spectral sharpness (Ambres)	**0.42**
Spectral dissonance (Sethares)	**0.28**
Spectral sharpness (Zwickler)	**0.37**
Timbral width	**0.32**
Volume	-0.22
Tonal dissonance (Sethares)	-0.25
Spectral dissonance (H&K)	-0.04
Tonal dissonance (H&K)	-0.11
Loudness	**0.41**
Spectral similarity	-0.26
Brightness (>1500Hz)	0.21
Brightness (>4000Hz)	0.17
Brightness (>400Hz)	0.06
Inharmonicity	0.05
Harmonic mode	0.13
Energy	0.23
ADSR envelope	-0.04
Register	0.12

Table 12.9.: Correlation coefficients between audio features and valence

Audio feature	CC
Spectral sharpness (Ambres)	**0.36**
Spectral dissonance (Sethares)	**0.49**
Spectral sharpnes (Zwickler)	**0.34**
Timbral width	**0.29**
Volume	-0.26
Tonal dissonance (Sethares)	-0.29
Spectral dissonance (H&K)	0.17
Tonal dissonance (H&K)	-0.06
Loudness	0.28
Spectral similarity	**-0.58**
Brightness (> 1500Hz)	0.18
Brightness (> 4000Hz)	0.21
Brightness (> 400Hz)	-0.03
Inharmonicity	0.06
Harmonic mode	0.08
Energy	0.29
ADSR envelope	-0.13
Register	0.02

Table 12.10.: Correlation coefficients between audio features and arousal

We calculated the correlation coefficients between the audio features with the highest importance on emotional discrimination (spectral similarity, spectral dissonance and spectral sharpness) and some of the most emotionally-relevant MIDI features (Table 12.11).

Audio and symbolic features	Correlation Coefficient
Spectral similarity and average duration accent	0.61
Spectral similarity and average note duration	0.50
Spectral similarity and average time between attacks	0.44
Spectral similarity and variability of time between attacks	0.43
Spectral similarity and strength of strongest rhythmic pulse	0.42
Spectral dissonance and variab. of note prev. of unpitched instruments	0.46
Spectral dissonance and percussion prevalence	0.45
Spectral dissonance and number of unpitched instruments	0.43
Spectral dissonance and bass drum prevalence	0.42
Spectral dissonance and melodic complexity	0.41
Spectral sharpness and harpsichord fraction	0.41
Spectral sharpness and number of unpitched instruments	0.40
Spectral sharpness and variab. of note prev. of unpitched instruments	0.35
Spectral sharpness and climax position	0.33

Table 12.11.: Correlation coefficients between relevant audio and symbolic features

12.6.4.2. Feature Selection and Classification

To have a first idea about the importance of the best audio features in the classification of the emotional content, we proceeded to a manual selection of features according to their correlations coefficients with the two emotional dimensions (Tables 12.9 and 12.10). The best features were spectral sharpness - Ambres, spectral dissonance - Sethares, loudness, spectral similarity, timbral width and tonal dissonance. Then, we calculated the correlation coefficient between the values of valence/arousal of the emotional answers (subsubsection 12.5.3.2) and the best features and obtained, respectively, the values of ~0.61/~0.75 for valence/arousal.

12.6.5. Discussion

The correlation between audio and MIDI features (subsubsection 12.6.4.1) allowed us to drawn some interesting conclusions. Longer duration of notes and longer time between the attacks (onsets) of the notes contribute to a more homogeneous frequency spectrum more homogeneous (similar). This is a plausible conclusion, because it is intuitive to conclude that less variations in a musical piece contribute to less variation in the frequency spectrum and as a result to a high degree of spectral similarity. Another conclusion drawn is that the use of unpitched (percussion) instruments contribute to a high degree of spectral dissonance. Melodic complexity also influences spectral dissonance. Other conclusions could be drawn, but these seemed to us the most important.

The values of the correlation coefficients presented in subsubsection 12.6.4.2 are close to the ones obtained with MIDI features: 0.76/0.77 (Table 12.4) in the first experiment and 0.70/0.77 (Table 12.8) of the first part of the second experiment. This gave us precious indication about the importance of the audio features, especially for the arousal, which had the closer results. We made an ad-hoc preliminary test on the effect of classifying music with both audio and emotionally-relevant MIDI features (e.g., average note duration and note density) and verified that the inclusion of audio features in the process contribute to an increase in the classification performance (measured with the help of the correlation coefficient). A detailed analysis of the effect of using both audio and MIDI features in the classification performance was left as an object of study to the third experiment (section 12.7) and to the experiments of the calibration/validation (chapter 15) of the EDME. Special attention would be devoted to the six features with the highest correlation coefficients (Tables 12.9 and 12.10).

We came to some conclusions based only on the results obtained in this part of the second experiment. For instance, there is some emotional content of the musical pieces which is only controlled in the audio domain. This is particularly true in the selection of instruments, more exactly in the selection of the samples for each note of the MIDI file. We verified the existence of colinearity among the features of the MIDI and audio domains. We can also infer that timbre/sound is an important musical feature that can be used to control/influence the emotional expression in music. This is visible not only on the results of this part of the second experiment, but also on the results of the first part of the second experiment and of the first experiment.

12.7. Third Experiment – Improvement of the Classification Module

The third experiment was divided into three parts. The first part is described in this section; the second part is described in section 12.8; the third part is described in section 12.9. This part consisted in the improvement of the classification module. The contents of this part of the experiment were published in the proceedings of the 2009 Sound and Music Computing Conference (Oliveira and Cardoso, 2009).

Category	Feature	Category	Feature
Instrumental (JSymbolic)	Note Prevalence of Bass Drum 2	Rhythmic (made with JSymbolic)	Average Rhythmic Pulse
	Note Prevalence of Bass Drum 1		Average Dynamics
	Note Prevalence of Side Stick/Rimshot		
	...		
	Note Prevalence of Open Cuíca	Harmonic (MIR Toolbox)	Spectral centroid
	Note Prevalence of Mute Triangle		
	Note Prevalence of Open Triangle		

Table 12.12.: Features analysed in the third experiment that were not analysed in the first and second experiments

12.7.1. Objective

This part of the third experiment had the same objective as the first experiment and of the first part and second parts of the second experiment. We worked toward the systematization of the emotionally-relevant group of features and their respective weights. This part of the third experiment was the third step towards automatic bi-dimensional classification of MIDI music by emotional content. This experiment was dedicated to the refinement of the knowledge base built with the data coming from the first and second experiments.

Like in the first and second experiments, we came up with the hypothesis that there is a small amount of features that may predict arousal/valence. We worked to refine the two sets of features relevant to valence and arousal. We intended to identify the emotional relevance of 482 features. We used JSymbolic (McKay and Fujinaga, 2006), MIDI Toolbox (Eerola and Toiviainen, 2004) and JMusic (Sorensen and Brown, 2000) to extract the MIDI features and MIR Toolbox (Lartillot and Toiviainen, 2007) and Psysound Toolbox (Cabrera, 1999) to extract the audio features. Special attention was put on the identification of the emotional relevance of new features and important ones from the first and second experiments: average time between attacks, variability of note duration, average note duration, importance of high register, note density, tempo, key mode, spectral sharpness, spectral dissonance and spectral similarity. Table 12.12 shows the features not present in Tables 12.1, 12.5 and 12.9. In the instrumental category, note prevalence of instruments is a multidimensional feature with size of 48 that corresponds to the number of standard unpitched instruments of the General MIDI. To avoid a longer table, we only mention the first and last three instruments separated by "..." to be representative of all the 48 percussion instruments. Detailed description of each feature can be consulted on the reference (McKay, 2004) as in the case of JSymbolic features, or on the following link as in the cases of MIR Toolbox[36].

12.7.2. Method

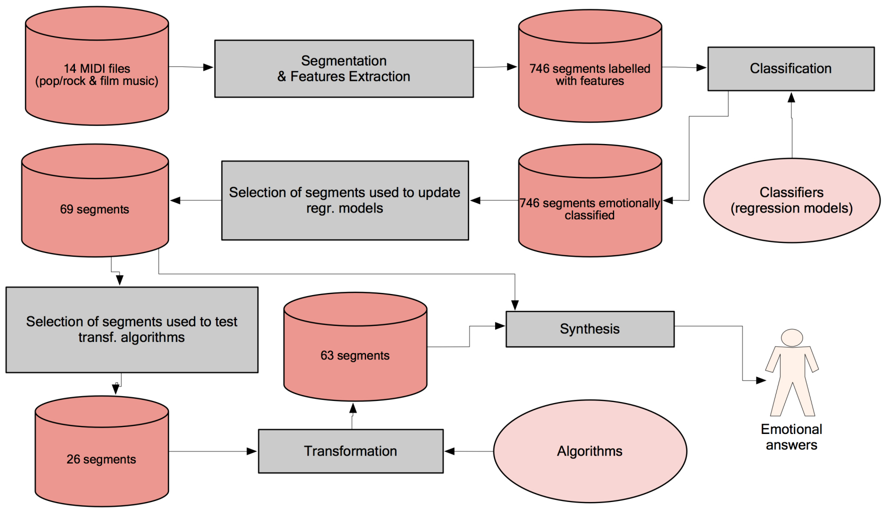


Figure 12.7.1.: Experimental steps of the third experiment

Figure 12.7.1 presents the steps of this experiment. These steps were designed in order to accomplish the objectives of the three parts of this experiment. We selected 14 MIDI files of western tonal music (pop/rock music genre) from a large database of pre-composed music of various genres. These files were selected based on their musical quality. The genres were randomly chosen from a group that included besides these, others like rap, R&B and classical. Selected files went through the processes of segmentation and feature extraction. From this resulted a group of 746 segments labelled with 482 musical features. The regression models built in the second experiment were

[36]https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/MIRtoolboxUsersGuide1.3.3

used to classify each segment with an appropriate emotional label. From this group of 746 segments emotionally classified, we selected 69 segments to be used to update regression models. This selection process was grounded on the purpose of covering all the bi-dimensional emotional space. Then, we selected a small group of 26 segments from the group of 69 pieces. These segments were used to test the effectiveness of transformation algorithms (see section 12.8). After transforming tempo, pitch, scale and articulation, we obtained a group of 63 segments (26 original segments + 37 transformed segments). Listeners were invited to classify a subgroup of 22 segments from the group of 132 selected segments (69 to update regression models + 63 to test transformation algorithms) through the use of several mailing-lists (intended for discussion of subjects like music theory, music therapy and others).

12.7.3. Data

Data consists of the selected musical segments and obtained emotional answers from the listeners.

12.7.3.1. Music

The 132 musical segments lasted between 10 and 15 seconds and are available in this link [37]. We extended the first experiment by increasing the number of musical pieces (from 16 of the first experiment and 96 of the first and second parts of the second experiment) and features (482, from 105 of the first experiment and from 414 of the first part of the second experiment).

12.7.3.2. Emotional Answers

37 listeners answered to the questionnaire: 28 male and 9 female aged between 14 and 63 years old (mean of 33, standard deviation of 13). They had background in informatics, technology and music. We calculated the mean and standard deviation for the emotional responses obtained in the questionnaire, as is shown in Figures 12.7.2, 12.7.3 and 12.7.4. Mean and standard deviations were computed first between listeners, and then averaged over segments. We measured the agreement of the listeners on the emotional content of the music using the Cronbach's Alpha and obtained a value of 35.74% for arousal and a value of 63.77% for valence. The value obtained for the emotional answers of arousal give us an unacceptable internal consistency [38] ; the value

[37] http://student.dei.uc.pt/~apsimoes/PhD/Music/smc09/
[38] http://en.wikipedia.org/wiki/Cronbach's_alpha

obtained for the emotional answers of valence is a bit better but with a questionable internal consistency [39]. Again,

this may be explained by the fact that each listener answered to its own subgroup of 22 segments (from the group of 132 selected segments) - as explained in subsection 12.7.2. Nobody answered to the same subgroup of musical segments as this subgroup was randomly chosen.

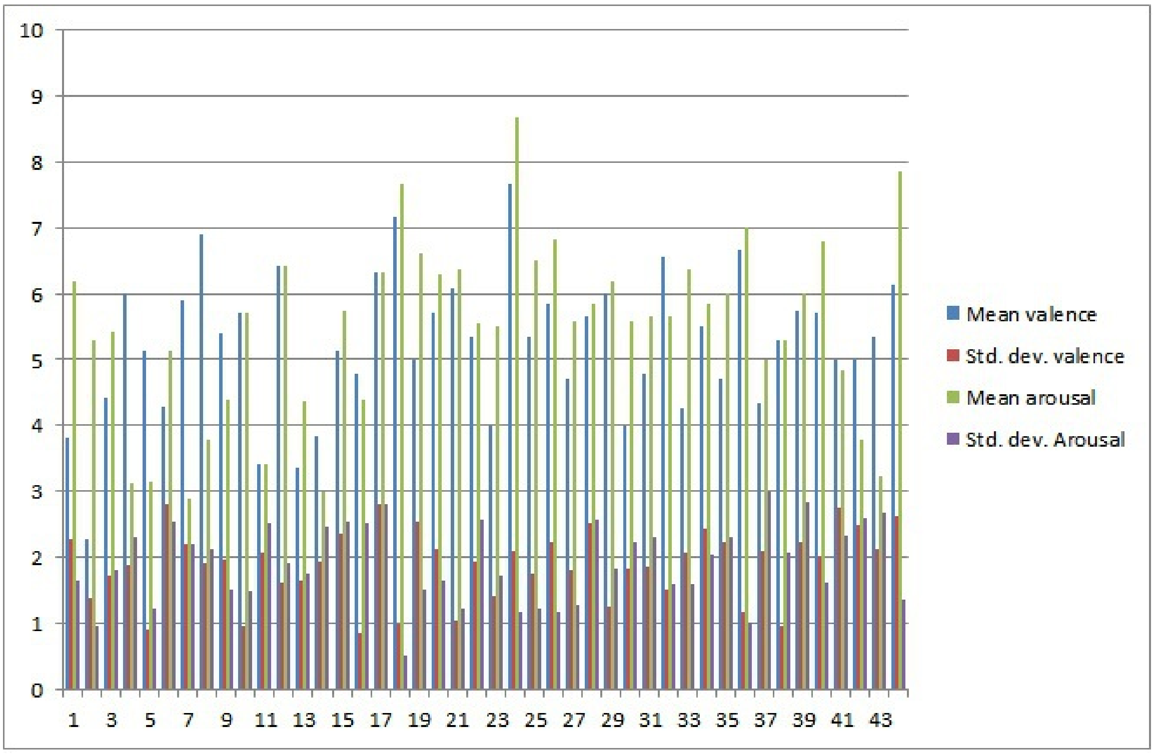


Figure 12.7.2.: Mean and standard deviations of the first 44 emotional responses in the third experiment

[39]http://en.wikipedia.org/wiki/Cronbach's_alpha

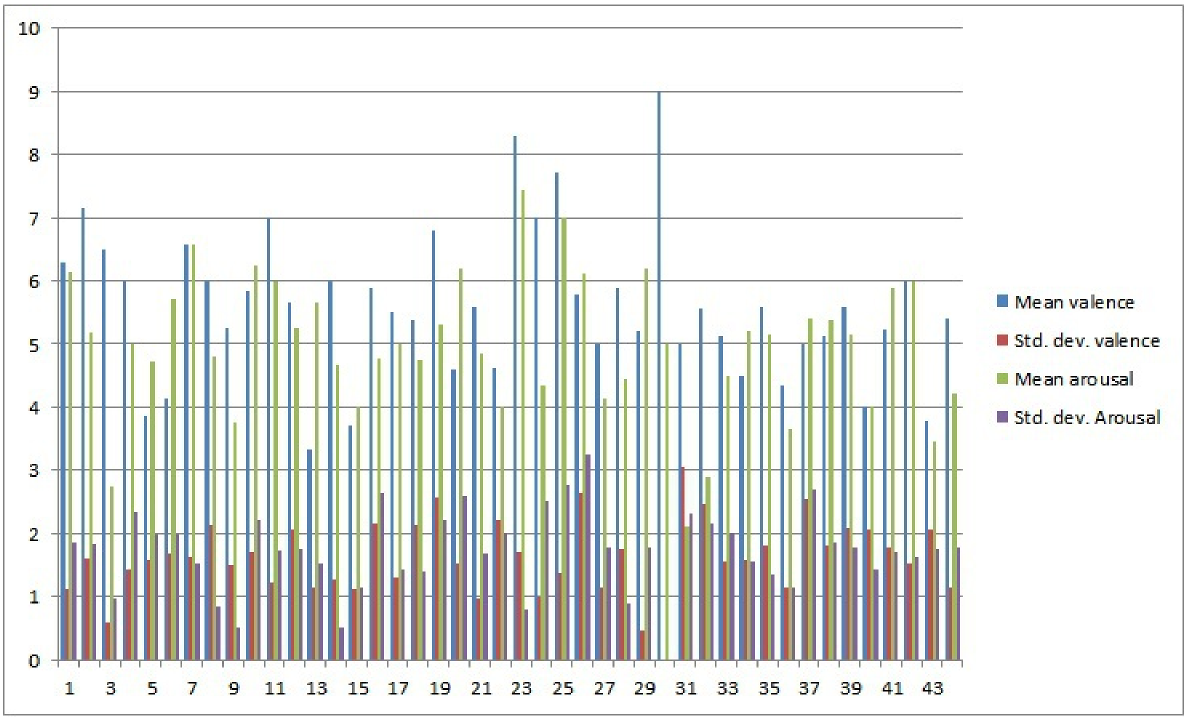


Figure 12.7.3.: Mean and standard deviations of the second 44 emotional responses in the third experiment

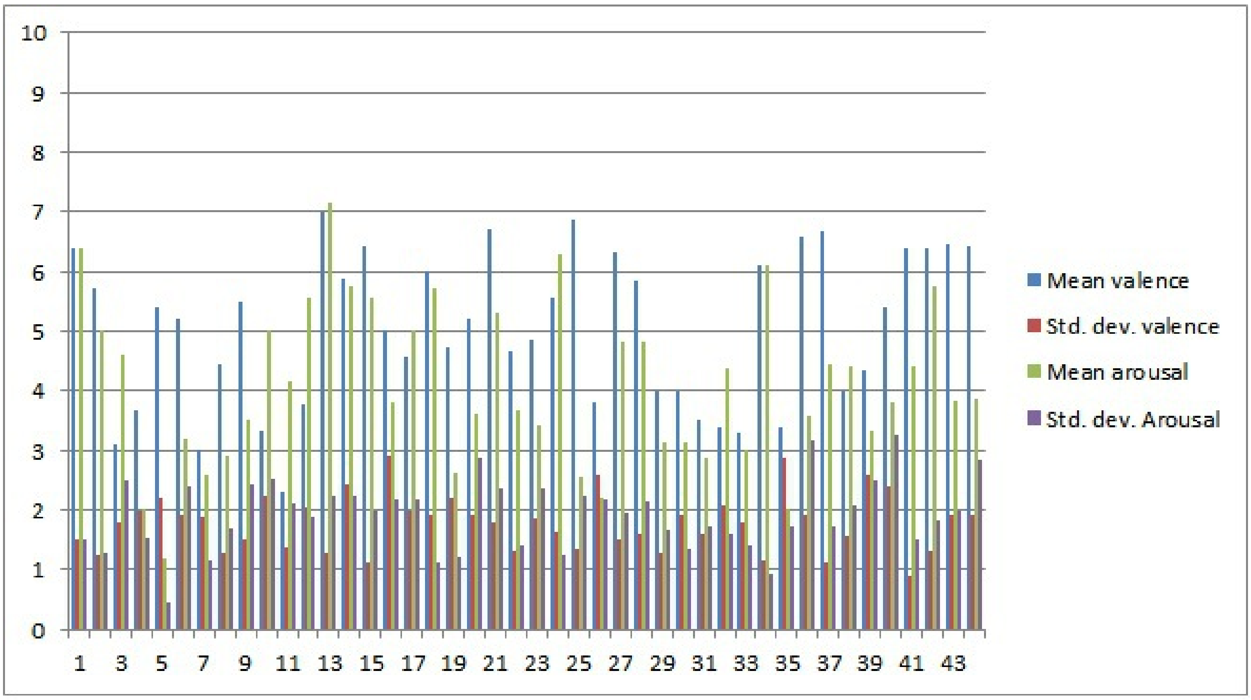


Figure 12.7.4.: Mean and standard deviations of the third 44 emotional responses in the third experiment

12.7.4. Results

The results of this experiment are divided into two subsubsections. One that consist of feature ranking and the other that consists of feature selection and classification.

12.7.4.1. Feature Ranking

We calculated individually the correlation coefficient between each feature and the two emotional dimensions. Table 12.13 and Table 12.14 present the correlation between the best features and valence and arousal. On the one hand, we can highlight rhythmic (e.g, staccato incidence, note density, average note duration and average time between attacks) and instrumentation (e.g., number of unpitched instruments) as the most relevant ones to the valence. On the other hand, we can highlight rhythmic (e.g, note density and staccato incidence) and instrumentation (e.g., variability of note prevalence of unpitched instruments, percussion prevalence and number of unpitched instruments) as the most relevant ones to the arousal.

Musical feature	CC
Staccato incidence	0.57
Number of unpitched instruments	0.53
Note density	0.52
Average note duration	-0.50
Average time between attacks	-0.50
Overall dynamic range	0.48
Variability of note duration	-0.46
Melodic fifths	0.45
Pitch variety	0.43
Note prevalence of closed hi-hat	0.42
Rhythmic looseness	0.41
Percussion prevalence	0.40

Table 12.13.: Features emotionally more discriminant for valence

Musical feature	CC
Var. note prev. unpitched instruments	0.70
Percussion prevalence	0.69
Note density	0.66
Number of unpitched instruments	0.58
Staccato incidence	0.56
Importance of loudest voice	0.55
Variation of dynamics	0.48
Note prevalence of snare drum	0.47
Overall dynamic range	0.46
Variability of note prevalence of pitched instruments	0.45
Note prevalence of bass drum	0.45
Note prevalence of closed hi-hat	0.43

Table 12.14.: Features emotionally more discriminant for arousal

12.7.4.2. Feature Selection and Classification

We applied the best first search method Witten et al. (1999) on the features with the highest ranking (Tables 12.13 and 12.14) to select the set of features features emotionally more discriminant. We made a compromise between the number of features and the quality of the results. Then, we applied 10-fold cross-validation on the most discriminant features with the results presented in Table 12.15. From the analysis of this table, we have the features overall dynamic range and variability of note duration with the highest weights in the classification of valence, and note density with the highest weight in the classification of arousal.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.69	0.76	0.97	Average time between attacks	-0.18
				Number of unpitched instruments	0.20
				Overall dynamic range	0.32
				Percussion prevalence	-0.12
				Variability of note duration	-0.31
Arousal	0.71	0.81	0.99	Note density	0.29
				Percussion prevalence	0.12
				Variability of unpitched instruments	0.15

Table 12.15.: Results of 10-fold cross-validation for valence and arousal – third experiment

12.8. Third Experiment - Evaluation of the Transformation Algorithms

The second part of the third experiment is described in this section. It consisted in the evaluation of the transformation algorithms. The contents of this part of the experiment were published in the proceedings of the 2009 Sound and Music Computing Conference (Oliveira and Cardoso, 2009).

12.8.1. Objective

Despite of the lower importance of the role of the transformation module when compared with the classification module, we also dedicated an experiment to test the effectiveness of its algorithms. This experiment was focused on the automatic transformation of two emotional dimensions of music (valence and arousal) by changing five musical features: tempo, pitch register, musical scales, instruments and articulation. We verified the effectiveness of the five algorithms in approximating the emotional content of music segments to the desired emotion of the listener.

12.8.2. Methods, Results and Discussion

The method used in this experiment is the one described in subsection 12.7.2. We present more details about the method, results and discussion for each algorithm in the following subsubsections.

12.8.2.1. Tempo

Algorithm The transformation of tempo starts by obtaining the original tempo of the music piece. Then, it changes the tempo parameter in the MIDI metadata, or, alternatively, it changes note onsets and/or increases/decreases the duration of notes.

Method We transformed six segments by accelerating their tempo in 50% and slowing it down in 30%, obtaining three versions with different tempos for each one: fast, normal and slow. For each of the resulting six groups of three segments, we correlated the tempo of each version with the emotional data obtained in the third experiment (see section 12.7).

Results Table 12.16 presents the correlation coefficients for each of the six groups. The last column of this table contains the absolute mean for all of these groups.

Group	1	2	3	4	5	6	Absolute Mean
Valence	0.94	0.96	0.91	-0.07	0.62	1.00	0.75
Arousal	1.00	0.98	0.98	-0.16	0.97	-0.36	0.74

Table 12.16.: Correlation coefficients between tempo and valence and arousal for the six groups of segments

Discussion The expected high positive coefficients were confirmed by most of the results. However, the fourth group of segments obtained small negative coefficients for both valence and arousal, and the sixth group for arousal. This may be explained by the presence of an imperceptible transformation, because of the presence of very long notes (> four seconds) in the original segment. A higher percentage of acceleration and slowing down of the original segment would be needed. The result of 100% for valence in the sixth group is not very reliable because the answers were very close: 3.5, 3.3 and 3.2. Emotional transformations contributed to an increase of 0.4/0.2 in valence/arousal with changes from low to normal tempo, and an increase of 1/0.8 in valence/arousal with changes from normal to high tempo.

12.8.2.2. Pitch Register

Algorithm The algorithm that transforms pitch register transposes up/down pitched instruments[40] (percussion instruments don't change) by a specific number of octaves to increase/decrease valence/arousal. We chose octaves, because they are the intervallic transformation more consonant (Vassilakis, 2005) with audible repercussion in the frequency spectrum. The system adds positive/negative multiples of twelve to the pitch of all the notes.

Method We transformed five segments by transposing them up and down two octaves, obtaining three versions of different registers for each one: high, normal and low. For each of the resulting five groups of three segments, we correlated the register of each version with the emotional data obtained in the third experiment (see section 12.7).

[40]http://wiki.answers.com/Q/What_is_the_difference_between_pitched_and_non-pitched_instruments

Results Table 12.17 presents the correlation coefficients for each of the five groups. The last column of this table contains the absolute mean for all of these groups.

Group	1	2	3	4	5	Absolute Mean
Valence	0.79	1.00	0.15	0.60	0.33	0.57
Arousal	-0.39	-0.63	-0.98	-0.92	-0.62	0.71

Table 12.17.: Correlation coefficients between pitch register and valence and arousal for the five groups of segments

Discussion Generally speaking, the increase of register correlates positively with valence and negatively with arousal. A more detailed analysis of the results in groups three and five showed lower correlation for valence, which revealed that the change from normal to high register contributes to a decrease in valence. From the analysis of the mean pitch of the segments, we can observe that, for these cases, the increase in register affects valence positively only till we have mean values of MIDI pitch around 80, whilst higher values contribute to a decrease in valence. We assisted to a similar situation in this first group for arousal: values of MIDI pitch higher than 80 do not seem to affect the arousal of music. Emotional transformations contributed to an increase of 2/-0.6 in valence/arousal with changes from low to normal register, and an increase of 0.7/-0.4 in valence/arousal with changes from normal to high register.

12.8.2.3. Musical Scales

Algorithm The algorithm that transforms musical scales finds the original scale of the MIDI file using MIDI toolbox (Eerola and Toiviainen, 2004), which takes into account only pitched instruments[41] (percussion instruments are not considered), and selects a target scale according to emotional tags to be defined for each scale. Once the scale is chosen, it finds the pitch distance relative to the tonic for each note in the original scale. If this distance is not found in the target scale, it finds the closest pitch distance that is present in the target scale and changes the pitch of the note, accordingly. Suppose we want to transform from a ragha madhuri scale (pitch distances of 4, 5, 7, 9, 10 and 11 semitones to the tonic) to a minor gipsy scale (pitch distances of 2, 3, 6, 7, 8 and 11 semitones to the tonic). A note distant four semitones from the tonic in the ragha madhuri scale would have its pitch decreased by one semitone to be distant three semitones from the tonic in the minor gipsy scale. This happens because the interval of four semitones is not present in the minor gipsy scale. We used a group of

[41] http://wiki.answers.com/Q/What_is_the_difference_between_pitched_and_non-pitched_instruments

27 twelve-tone scales[42]. We chose this group and not others because it has a higher variety of number of notes and intervals: scales have between two and seven notes and intervals vary from one to seven semitones.

Method We transformed one segment by changing the original major scale to other 27 musical scales. We used feature selection algorithms in the process of finding the features that best characterize the emotional variation when changing the scale.

Results We calculated the weights for the most important features for valence: number of semitones in scale (-0.17), difference between successive intervals of the scale (-0.15), spectral dissonance (0.18) and spectral sharpness (-0.14). We made the same for arousal: number of semitones in scale (-0.19), difference between successive intervals of the scale (-0.07), spectral dissonance (0.14) and stepwise motion (0.24). Table 12.18 presents the correlation coefficients between the most discriminant features and the emotional dimensions.

Feature	Valence	Arousal
Spectral dissonance	0.46	0.31
Tonal dissonance	0.28	-
Timbral width	-0.32	-
Spectral sharpness	0.34	-0.20
Stepwise motion	0.24	0.33
Melodic thirds	-0.34	-0.18
Number of semitones in scale	-0.40	-0.23
Differenc. successive intervals in scale	-0.28	-0.16
Correlation coefficient	0.61	0.45

Table 12.18.: Correlation coefficients between musical features and valence and arousal for the 27 versions of the segment

Discussion It is not an easy task to find features that can be helpful in defining a musical as is shown by the low correlation coefficients of the features shown in Table 12.18. However, some of the features despite of its low correlation coefficients can be helpful in finding scales more appropriate to some emotions.

[42]http://papersao.googlepages.com/musicalscales

12.8.2.4. Instruments

Algorithm The algorithm used to transform the set of instruments used by the music, obtains original MIDI instruments specification and selects new instruments according to the emotional tags of each timbre. These tags are pre-computed, offline, through a weighed sum of audio features (e.g., spectral dissonance and spectral sharpness), with the help of a vector of weights defined in the knowledge base for each emotional dimension. Transformations are done by taking into consideration spectral features (Lartillot and Toiviainen, 2007), to allow the transformation to be done with compatible instruments, for example, it is preferable to change an acoustic piano to an electric piano, instead of changing it into a trumpet.

Method We changed the instruments of the original group of 69 segments (not subject to any type of transformation). This change consisted only in modifying the MIDI patch (instrument) of the musical piece. We tried to have each of the General Midi 1 (GM1) instruments present in, at least, one of the segments, in order to analyse the emotional impact of every GM1 instrument. This test was an extension of what was done in the second experiment on the analysis of audio features (section 12.6).

Results Table 12.19 presents the correlation coefficients between audio features and the valence and arousal of the segments.

Audio Feature	Valence	Arousal
Spectral dissonance	0.28	0.72
Timbral width	-	0.54
Tonal dissonance	0.19	0.27
Spectral sharpness	-	0.44

Table 12.19.: Correlation coefficients between musical features and valence and arousal for the 69 segments.

Discussion We can infer that instruments are essentially relevant to the arousal, because for valence the correlation coefficients obtained in this experiment are lower than the ones obtained in section 12.6. Spectral dissonance and spectral sharpness obtained the highest values of correlation coefficients (as in section 12.6). So, these features can be said to be the more relevant in the emotional analysis of the sound/timbre of instruments. We found that violin, string ensembles, choirs and piccolo contribute to low valence; and percussion instruments contribute to high valence/arousal.

As in these tests, the instruments that we intended to evaluate did not appear alone, despite its high presence in the music, the results shall be analysed with caution. However, the results give us indications about some tendencies.

12.8.2.5. Articulation

Algorithm The algorithm that transforms normal to staccato articulation decreases the duration of all notes by a specific percentage. If we consider 75%, notes with a duration of X would have a new duration of X-X*0.75.

Method We transformed 14 segments by changing their articulation to staccato and obtained two versions for each one: normal and staccato. We correlated the articulation of the 28 versions with the emotional data obtained in our experiment.

Results We found that the change from normal to staccato articulation is 40% correlated with the increase of valence and has no impact in arousal.

12.8.3. Overall discussion

We successfully tested the effectiveness of algorithms of music transformation. Change of tempo was positively related to both valence and arousal. Change of pitch register was positively related to valence and negatively related to arousal. The presence of semitones in musical scales was found to be an important feature negatively related to valence. Spectral dissonance, timbral width and spectral sharpness were found to be important features for instruments and are positively related to arousal. Staccato articulation was found to be positively related to valence.

If we look at all the experiments carried out to evaluate the transformation module (Oliveira and Cardoso, 2008a,b, 2009), we can conclude that the transformation of tempo, note density, pitch register, spectral sharpness (Ambres), spectral sharpness (Zwickler), timbral width (spectral flatness) and loudness contribute to a direct influence on valence; and that the transformation of tempo, note density, spectral sharpness (Ambres), spectral sharpness (Zwickler) and spectral dissonance (Sethares) contribute to a direct influence on arousal. The transformation of pitch register and spectral similarity influenced arousal in an inverse way.

12.9. Third Experiment - Melodic Analysis

The third part of the third experiment is described in this section. It consisted on the analysis of the influence of the melody in the emotional content of music. The contents of this part of the third experiment were published in the journal of Knowledge-Based Systems (Oliveira and Cardoso, 2010).

12.9.1. Objective

In the first and second experiments, as well as in the first and second parts of the third experiment, we proceeded to the extraction of features from the whole musical pieces. We thought that we could gain in the classification performance of the emotional content, if we moved from the analysis of the whole piece (bass line, harmonic line, melodic line and percussion line) to the analysis of only the melodic line of the piece. The part of melodic analysis of the third experiment aimed to verify the importance of the melody in the expression of emotions by using the data of the three experiments. We came up with the hypothesis that by analysing solely the melodic line it would be easier to find features with a high degree of emotional discrimination.

12.9.2. Method

We manually extracted the melodic lines from the musical pieces used in the experiments. We guided this extraction by considering the loudness and pitch of the notes: notes with high loudness and pitch were considered as having a high probability of belonging to the melodic line. We extracted from the melodic lines the features analysed in the third experiment and used the listeners' answers obtained with the questionnaires.

12.9.3. Data

We used data coming from the first experiment, first part of the second experiment and first part of the third experiment. This data includes music and emotional answers (subsections 12.4.3, 12.5.3 and12.7.3).

12.9.4. Results

We did not proceeded to feature ranking in this part of the third experiment, because we already had an idea of which were the more emotionally relevant features after analysing the results of the first experiment, three parts of the second experiment and the first two parts of the third experiment.

12.9.4.1. Feature Selection and Classification

In order to select the set of features features emotionally more discriminant in each of the six cases of classification presented in Table 12.20, we applied a mixture of manual and automatic selection. Manual selection was guided by the emotional importance of features based on the results of the previous two experiments and results of the first two parts of the third experiment, as well as based on the results from the literature of Music Psychology (chapter 6). Automatic selection was done with the application of the best first search method Witten et al. (1999). We made a compromise between the number of features and the quality of the results. Then, we applied 10-fold cross-validation on the most discriminant features with the results presented in Table 12.20. Both feature selection (automatic and manual) and classification (using 10-fold cross-validation) where applied separately, for each of the six cases presented in the mentioned table.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence - data of first experiment	0.79	0.61	0.87	Average note duration	-0.04
				Rhythmic variability	-0.35
				Staccato incidence	0.18
				Time prevalence of koto	0.22
				Variability of note duration	-0.50
Valence - data of second experiment	0.62	0.94	1.18	Average time between attacks	-0.47
				Tempo	0.59
				Maximum note duration	-0.06
				Variability of note duration	0.04
				Variation of dynamics	0.25
Valence - data of third experiment	0.41	1.00	1.25	Average note duration	-0.16
				Comb. streng. two strong. pulses	-0.25
				Minimum note duration	-0.32
				Strength strong. rhythmic pulse	0.11
Arousal - data of first experiment	0.85	0.64	0.85	Average note duration	-0.48
				Tempo	0.29
				Maximum note duration	-0.25
				Most common pitch prevalence	0.29
Arousal - data of second experiment	0.72	0.89	1.14	Average note duration	-0.09
				Average time between attacks	-0.83
				Tempo	0.41
				Variation of dynamics	0.43
Arousal - data of third experiment	0.54	0.94	1.20	Number of common pitches	0.05
				Rel. streng. common mel. interval	0.14
				Variation of dynamics	0.40

Table 12.20.: Results of 10-fold cross-validation for valence and arousal – melodic analysis

12.9.5. Discussion

An interesting aspect found through all these six cases of classification is that there are many ”new” best features that did not appeared in the first experiment, in the first two parts of the second experiment and in the first part of the third experiment. In the case of the classification of valence with the data of the first experiment we have two ”new” features: rhythmic variability and time prevalence of koto. Four ”new” features appeared in the classification of valence with data of second and third experiments: maximum note duration, combined strength of the two strongest pulses, minimum note duration and strength of the strongest rhythmic pulse. The classification of arousal with the data of first, second and third experiments gave rise to another four ”new” features: maximum note duration, most common pitch prevalence, number of common pitches and relative strength of common melodic interval. Ten ”new” features out of 25 features were used in six cases of classification. We also observed that there is variability on the best features for each of the experiments. The conditions that vary across the experiments are basically the style and the duration of the musical pieces. We believe that this variability may be explained by the style differences.

After analysing the correlation coefficients of Table 12.20, we can conclude that the melody does not reflect much of the emotional content expressed in the music analysed in the third experiment. The correlation coefficients of 0.41 for valence and of 0.54 for arousal are low. The same happens in the classification of valence with the data of the second experiment, where we obtained a correlation coefficient of 0.62. On the other side, we obtained high correlations coefficients (0.79 and 0.85) in the classification of, respectively, valence and arousal with the data of the first experiment. Other relatively high correlation coefficient (0.72) was obtained in the classification of arousal with the data of the second experiment. The correlation coefficients also revealed to be lower when using just the melody for most of the cases. These results do not concur with our initial thought that the classification performance of the emotional content could be improved if we were focused on the extraction of features from only the melodic line. These results may indicate a lower relevance of the melody in discriminating the emotional content of music. Therefore, we decided to keep considering the whole information about the music in our approach. The hypothesis that by analysing solely the melodic line it would be easier to find features was not confirmed. What was confirmed was that we gain from analysing the whole piece in order to find emotionally-relevant features.

13. Knowledge Base Systematization

Before proceeding to the calibration/validation of the EDME system, we decided to systematize the knowledge base. This systematization consisted in making a careful analysis through all the collected data (musical and emotional). This data was obtained in three sequential experiments that were divided, in some cases, into several parts, as was the case of the second and third experiments. The main purpose of the experiments was to contribute to a better control of the emotions being expressed in music.

In order to build the knowledge base, we decided to collect the most discriminant features in previous experiments (tables 12.4, 12.8, 12.10, 12.9, 12.15 and 12.20), in literature (chapter 6) and in similar works (section 8.4). We obtained a group of 13 features for both valence and arousal. Similarly to what was done in the described experiments we went through a stage of feature ranking followed by a stage of feature selection and classification.

Musical feature	CC - First Experiment	CC-Sec. Experiment	CC - Third Experiment
Average Note duration	-0.52	-0.50	-0.50
Average Time Between Attacks	-0.52	-0.49	-0.50
Importance of Bass Register	-0.03	-0.09	0.00
Tempo	-0.06	0.63	-
Note Density	0.57	0.43	0.52
Percussion Prevalence	0.16	0.06	0.40
Repeated Notes	-0.24	0.1	-0.04
Variation of Dynamics	0.01	0.05	0.00
Key mode	-0.14	-0.44	0.00
Spectral loudness	0.26	-	0.17
Spectral dissonance (Sethares)	0.02	-	0.28
Spectral sharpness (Ambres)	0.07	-	0.09
Spectral similarity	0.17	-	-0.13

Table 13.1.: Correlation between features and valence

13.1. Feature Ranking

As a term of comparison, we decided to calculate the correlation coefficients for these features with the data of each of the three experiments. Table 13.1 presents the correlation between the features and valence. Table 13.2 presents the correlation between the features and arousal. We can highlight average note duration, average time between attacks and note density as being the most relevant ones for both valence and arousal. It is also important to mention the relevance of percussion prevalence to arousal.

Musical feature	CC - First Experiment	CC-Sec. Experiment	CC - Third Experiment
Average Note duration	-0.72	-0.69	-0.26
Average Time Between Attacks	-0.74	-0.56	-0.21
Importance of Bass Register	-0.32	0.13	0.08
Tempo	-0.40	0.56	-
Note Density	0.68	0.64	0.66
Percussion Prevalence	0.53	0.20	0.69
Repeated Notes	-0.04	0.32	0.35
Variation of Dynamics	0.50	0.00	0.48
Key mode	-0.40	-0.23	-0.14
Spectral loudness	0.03	-	0.53
Spectral dissonance (Sethares)	-0.02	-	0.72
Spectral sharpness (Ambres)	-0.14	-	0.44
Spectral similarity	0.18	-	-0.38

Table 13.2.: Correlation between features and arousal

13.2. Feature Selection and Classification

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.78	0.66	0.88	Average Note Duration	-0.61
				Average Time Between Attacks	-0.28
				Tempo	-0.33
				Note Density	0.45
				Variation of Dynamics	-0.34
				Key Mode	0.26
				Spectral Sharpness	0.07
				Spectral Loudness	0.37
				Spectral Similarity	-0.14
Arousal	0.74	0.66	1.08	Average Note Duration	-0.49
				Average Time Between Attacks	-0.33
				Importance of Bass Register	-0.18
				Note Density	0.07
				Variation of Dynamics	0.35
				Spectral Dissonance	-0.15

Table 13.3.: Results of 10-fold cross-validation for valence and arousal – first experiment

We proceeded to a phase of feature selection by using both manual selection and the best first search method Witten et al. (1999) in the group of 13 features, in order to find a smaller set of features features, always having in mind the compromise between the number of features and the quality of the results. Using the data of the first experiment, we obtained a group of nine features for valence and a group of six features for arousal. We applied 10-fold cross-validation with these features with the results presented in Table 13.3. From the analysis of this table, we have average note duration and note density with the highest weights in the classification of valence, and average note duration, average time between attacks and variation of dynamics with the highest weights in the classification of arousal. Using the data of the second experiment, we obtained a group of six features for valence and a group of five features for arousal. We applied 10-fold cross-validation with these features with the results presented in Table 13.4. From the analysis of this table, we have average note duration and tempo with the highest weights in the classification of valence, and average note duration and note density with the highest weights in the classification of arousal. Using the data of the third experiment, we obtained a group of five features for valence and a group of three features for arousal. We applied 10-fold cross-validation with these features with the results presented in Table 13.5. From the analysis of this table, we have average note duration and average time between attacks with the highest weights in the classification of valence, and spectral dissonance with the highest weight in the classification of arousal.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.68	0.89	1.09	Average Note Duration	-0.26
				Average Time Between Attacks	-0.15
				Importance of Bass Register	-0.19
				Tempo	0.47
				Note Density	0.14
				Key Mode	-0.14
Arousal	0.81	0.74	0.96	Average Note Duration	-0.93
				Average Time Between Attacks	0.22
				Tempo	0.33
				Note Density	0.54
				Repeated Notes	0.36

Table 13.4.: Results of 10-fold cross-validation for valence and arousal – second experiment

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.58	0.85	1.10	Average Note Duration	-0.39
				Average Time Between Attacks	-0.31
				Importance of Bass Register	-0.18
				Percussion Prevalence	0.12
				Spectral Loudness	0.10
Arousal	0.75	0.72	0.95	Percussion Prevalence	0.13
				Spectral Dissonance	0.38
				Spectral Similarity	-0.21

Table 13.5.: Results of 10-fold cross-validation for valence and arousal – third experiment

We went further and joined the musical and emotional data (Figure 13.2.1) [43]. We used, again, both manual selection and the best first search method Witten et al. (1999) on the group of 13 features. We obtained a group of seven features for valence and a group of six features for arousal. We applied 10-fold cross-validation with these features with the results presented in Table 13.6. From the analysis of this table, we have average note duration and tempo with the highest weights in the classification of valence, and tempo and note density with the highest weights in the classification of arousal. We also calculated the percentage of correct predictions and obtained results of 79,0% for valence and 84,5% for arousal. We considered a correct prediction the one that falls in the interval of the mean value of the emotional answer, given by the listeners, plus or minus the standard deviation of this answer.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.62	0.88	1.11	Average note duration	-0.45
				Average time between attacks	-0.21
				Importance of Bass Register	-0.16
				Tempo	0.38
				Note Density	0.15
				Variation of Dynamics	0.12
				Key mode	-0.09
Arousal	0.77	0.83	1.02	Average note duration	-0.20
				Tempo	0.39
				Note density	0.58
				Percussion prevalence	0.15
				Repeated Notes	0.18
				Variation of Dynamics	0.14

Table 13.6.: Results of 10-fold cross-validation for valence and arousal after joining the data of all the experiments

[43] As explained in section 8.2, we are using bidimensional plots for representing emotions, with the horizontal axis representing valence and the vertical axis arousal. Each point represents the mean values of valence and arousal obtained from the listeners for each piece of music.

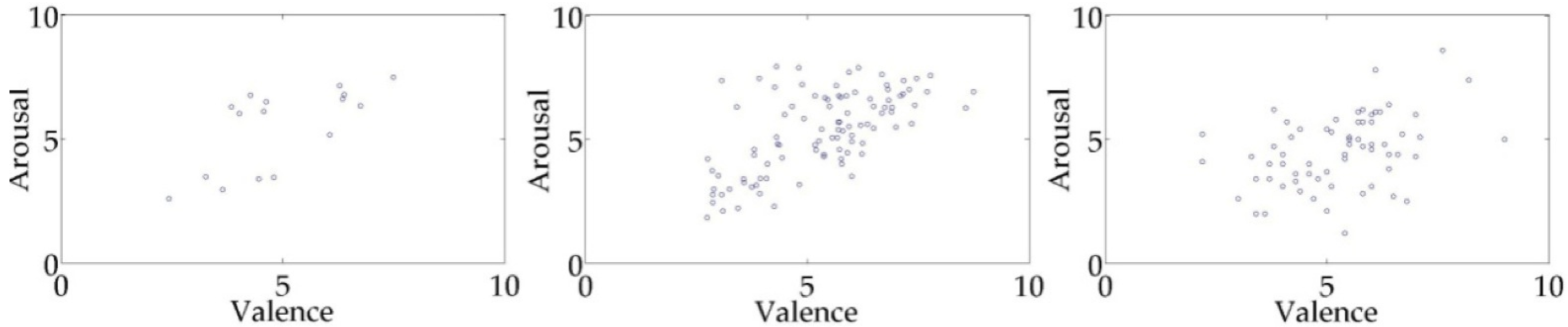


Figure 13.2.1.: Scatterplot of emotional data of first, second and third experiments

13.3. Discussion

For the first experiment (Table 13.3), average note duration, average time between attacks, note density and variation of dynamics are common features used in the classification of the emotional dimensions. Note density has a positive influence as a result of the positive weights; average note duration and average time between attacks have a negative influence as a result of the negative weights. Then, for valence, we have tempo, variation of dynamics and spectral similarity with a negative influence, and key mode, spectral sharpness and spectral loudness with a positive influence; for arousal, we have importance of bass register and spectral dissonance with a negative influence, and variation of dynamics with a positive influence.

For the second experiment (Table 13.4), average note duration, average time between attacks and tempo are common features used in the classification of the emotional dimensions. Tempo and average time between attacks (for arousal) have a positive influence as a result of the positive weights; average note duration and average time between attacks (for valence) have a negative influence as a result of the negative weights. Then, for valence, we have importance of bass register and key mode with a negative influence, and note density with a positive influence; for arousal, we have note density and repeated notes with a positive influence.

For the third experiment (Table 13.5), percussion prevalence is the only common feature used in the classification of the emotional dimensions. It has a positive influence as a result of the positive weights. Then, for valence, we have average note duration, average time between attacks and importance of bass register with a negative influence, and spectral loudness with a positive influence; for arousal, we have spectral dissonance with a positive influence; and spectral similarity with a negative influence.

The results of the classification using the data of all the three experiments were presented in Table 13.6. After analysing this table, we observed that average note duration, tempo, note density and variation of dynamics are common features used in the classification of the emotional dimensions. Average note duration has a negative influence as a result of the negative weight; tempo, note density and variation of dynamics have a positive influence as a result of the positive weights. Then, for valence, we have average time between attacks, importance of bass register and key mode with a negative influence; for arousal, we have percussion prevalence and repeated notes with a positive influence.

The analysis of the contents of these first four paragraphs of this section allows us to make further observations. Average note duration and average time between attacks are always used in the classification of valence with a negative influence as a results of the negative weight.

14. Evaluation of Classifiers' Performance

Musical and emotional features are given to a classifier in order to obtain a model that relates the musical and the emotional domains. We tested different models and methods of optimization available on Weka (Witten et al., 1999). We considered five categories of classifiers: function-like, instance-based, mixed, rule-based and tree-based. Having in mind that each classifier intends to learn a mapping model, we explain each one briefly. More details about each of the classifiers, models and algorithms following referred can be obtained in (Witten and Frank, 2005).

1. **Function-like**. *Gaussian Process regression* uses gaussian functions to map an input vector to an output vector. It allows the normalization and standardization of input vector. Polynomial and gaussian support vector kernels can be used: normalized poly kernel, poly kernel, pre-computed kernel matrix kernel, puk, RBF kernel and string kernel . Isotonic regression uses the least square error method to pick the best feature and estimate the isotonic regressive function. *Linear regression* fits input to output vector by using a specific optimization, like least mean square. It uses the Akaike criterion to select the best function. It allows feature selection using M5 and greedy methods; and elimination of collinear features. *Least mean square* is a steeper descent algorithm with a stochastic method of optimization. It uses the *Linear regression method* to develop least median square regression functions. These functions are generated from random subsamples of the input vector. This method selects the function with the lowest median square error. *Multilayer perceptron* is a type of neural network that uses various layers of non-linear functions. It trains data using backpropagation algorithm. Various parameters can be defined: number of hidden layers, learning rate, momentum, training time and others. *Pace regression* is an improvement of the ordinary least squares that estimates the effect of each input feature and uses cluster analysis to improve the estimation of a mapping function. It consists of a group of estimators that can be of various types: empirical bayes, nested model selector, subset selector, PACE2, PACE4, PACE6, ordinary least squares selection, AIC, BIC and RIC. It is adequate when there are many features, because it determines very well which ones to discard. *Radial Basis Function* method uses functions of this type to find a mapping function. It uses the k-means clustering algorithm to provide the basis functions and learns a linear regression on top of that. Symmetric multivari-

ate gaussians are fit to the data from each cluster. It standardizes all the features and uses k parameter to define the number of clusters being generated. *Simple linear regression* uses ordinary least squares methods. It obtains the feature with the lowest squared error. *Sequential Minimization Optimization (SMO) regression* is a type of Support Vector Machine regression that uses the SMO algorithm for training a support vector classifier. As with the method of *Gaussian Process regression* it uses polynomial or gaussian kernels and allows the normalization and standardization of input data.

2. **Instance-based**. *Instance-based k-nearest neighbor* uses a k-nearest neighbor algorithm to find a solution from a space with part of the input vector. Four algorithms can be used: ball tree, cover tree, KD tree and linear NN search. It allows the selection of the best K value using cross-validation. Euclidean distance is the distance metric being used. The number of neighbours is another parameter to be defined. *K** calculates an entropic distance between instances and the variable to be classified. It uses a generalized distance function based on transformations. *LWL* weights each instance using local distance functions. Weighed instances are used to build a classifier that can be any of the other classifiers here described. Like the *Instance-based k-nearest neighbour* it allows the used of four algorithms of search for the nearest neighbour.
3. **Mixed**. These kinds of methods use various types of classifiers. *Additive regression* is a boosting algorithm that is used to improve the performance of regression classifiers. It uses two parameters: the shrinkage, which governs the learning rate; and the maximum number of models to generate. Each iteration fits a model to the residuals left by the classifier in the previous iteration. *The bagging method* divides the input vector into various input vectors with a lower dimension which are given to different classifiers. *Ensemble selection* uses the average prediction of several classifiers to predict the output value. It allows the use of five different types of algorithms to optimize the ensemble: forward selection, backward elimination, forward selection + backward elimination, best model and build library only. Seven metrics can be used to optimize the chosen ensemble: accuracy, RMSE, ROC, precision, recall, fscore and all the referred metrics. *Random Subspace* divides input vector into different subspaces that are used by different tree-type classifiers. *Regression by discretization* converts continuous input vector into a discrete input vector that is used by any type of the classifiers here described.
4. **Rule-based**. *Conjunctive rule method* establishes rules composed by conjunctions of different variables of the input vector. This method calculates the information gain of each variable and prunes the generated rule using Reduced Error

Pruning (REP) or simple pre-pruning based on the number of variables of each rule. *Decision tables method* uses a set of features and a set of labeled instances to predict the output of new instances. It uses the root mean square error to evaluate the performance of feature combinations used in the decision table. It applies best-first search to evaluate the subsets of features and can use cross-validation for evaluation. *M5Rules* build various trees using M5'. It obtains regression rules using the best leaf from each tree.

5. **Tree-based**. *Decision stump* is a predictive model which uses a binary tree with only one level. *M5P* builds trees' models with the help of the divide and conquer method. *REP tree* builds regression trees' models using information gain/variance reduction criterion. Trees are pruned using reduce-error.

With the systematization of the knowledge base, we were ready to evaluate the performance of various classifiers in the classification of valence and arousal (Figure 14.0.1 and 14.0.2). A description of the acronyms of the classifiers presented in these figures is available[44]. The performance was evaluated by applying training/test split (66%/34%) and 10-fold cross-validation. Each classifier was evaluated with their default parameters (Witten et al., 1999). We considered three metrics: correlation coefficient (CC), mean absolute error (MAE) and root mean square error (RMSE).

[44]GP – Gaussian Process; IR – Isotonic Regression; LMS – Least Mean Square; LR – Linear Regression; MP – Multilayer Perceptron; PR – Pace Regression; RBF – Radial Basis Function; SLR – Simple Linear Regression; SMO – SVM Regression; IBK – Instance-Based K-Nearest Neighbor; KS – K Star; LWL – Locally-weighted Learning; AR – Additive Regression; BAG – Bagging; ES – Ensemble Selection; RSS – Random SubSpace; RD – Regression By Discretization; CR – Conjunctive Rule; DT – Decision Table; M5R – M5 Rules; DS – Decision Stump; M5P – M5 Trees; REP – REP Tree.

Classifier	GP	IR	LMS	LR	MP	PR	RBF	SLR	SMO	IBK	KS	LWL	AR	BAG	ES	RSS	RD	CR	DT	M5R	DS	M5P	REP	Measure
First experiment cross-validation	0.69 0.89 1.07	0.66 **0.83** 1.07	**0.72** **0.83** **0.97**	**0.73** **0.78** **0.95**	0.63 0.97 1.15	- - -	0.19 1.31 1.41	0.55 1.00 1.19	**0.76** **0.75** **0.91**	0.54 1.15 1.27	0.68 0.98 1.12	0.49 1.04 1.27	**0.71** 0.87 1.11	0.60 0.99 1.11	0.61 0.95 1.11	0.59 1.00 1.17	**0.72** **0.83** **0.99**	0.45 1.07 1.27	0.39 1.11 1.30	0.69 0.84 1.01	0.42 1.12 1.37	**0.73** **0.79** **0.95**	0.52 1.00 1.21	CC MAE RMSE
First experiment training/test split	**0.81** 0.96 **1.03**	0.46 1.45 1.54	**0.85** 1.30 1.41	0.54 0.93 **1.05**	0.67 1.41 1.50	- - -	0.70 **0.79** **1.09**	0.20 1.62 1.82	0.79 **0.84** **1.02**	0.53 **0.75** **0.97**	**0.88** 1.14 1.20	**0.91** 0.96 1.12	**0.86** 1.06 1.17	**0.81** 1.29 1.34	0.70 0.92 **1.06**	0 1.54 1.61	**0.90** 1.13 1.18	0 1.12 1.21	0 1.43 1.51	0.38 1.37 1.65	0.70 **0.85** 1.09	0.37 1.33 1.60	0 1.54 1.61	CC MAE RMSE
Second experiment cross-validation	**0.72** **0.84** **1.01**	0.59 0.98 1.16	**0.71** **0.84** 1.03	**0.71** **0.83** **1.02**	0.68 0.87 1.11	**0.71** **0.83** **1.01**	0.53 1.01 1.22	0.61 0.98 1.14	**0.70** 0.85 1.04	0.57 1.09 1.40	0.56 0.98 1.22	0.50 1.04 1.26	0.61 0.95 1.20	0.63 0.93 1.12	0.61 0.95 1.15	0.63 0.93 1.11	0.50 1.10 1.34	0.39 1.12 1.34	0.53 1.08 1.24	**0.71** **0.83** **1.02**	0.34 1.14 1.39	**0.71** **0.83** **1.02**	0.55 0.97 1.24	CC MAE RMSE
Second experiment training/test split	**0.83** 0.91 **1.02**	0.66 1.04 1.21	**0.83** **0.82** **0.96**	**0.80** **0.88** **1.00**	0.63 1.06 1.26	**0.82** 0.95 **0.97**	0.59 1.14 1.35	0.72 1.06 1.17	**0.82** **0.87** **1.00**	0.61 1.10 1.36	0.56 1.08 1.32	0.66 1.04 1.21	0.73 0.90 1.08	0.77 0.98 1.12	0.59 1.15 1.31	0.74 1.01 1.17	0.71 0.91 1.12	0.48 1.13 1.39	0.64 1.08 1.22	**0.80** **0.88** **1.00**	0.54 1.09 1.33	**0.80** 0.88 **1.00**	0.64 1.06 1.22	CC MAE RMSE
Third experiment cross-validation	**0.66** **0.77** **1.00**	0.49 1.00 1.20	**0.64** 0.83 **1.04**	**0.68** **0.77** **0.98**	**0.66** 0.83 1.08	**0.68** **0.78** **0.97**	0.40 1.00 1.23	0.37 1.01 1.28	**0.69** **0.76** **0.97**	0.55 1.00 1.27	0.60 0.87 1.13	0.62 0.87 **1.06**	0.48 0.93 1.28	0.56 0.83 1.10	0.49 0.90 1.20	0.58 0.89 1.09	0.41 0.96 1.31	0.15 1.06 1.40	0.36 1.07 1.33	**0.65** **0.80** **1.02**	0.12 1.08 1.44	**0.68** **0.78** **0.98**	0.51 0.90 1.16	CC MAE RMSE
Third experiment training/test split	**0.71** 0.90 **1.08**	0.45 1.07 1.25	**0.75** **0.87** **1.07**	**0.75** **0.77** **0.97**	0.50 1.23 1.60	**0.76** **0.75** **0.95**	0.38 1.13 1.30	0.41 1.10 1.28	**0.77** **0.78** **0.99**	0.62 0.96 1.18	0.65 **0.87** 1.11	0.69 **0.87** **1.05**	0.51 1.09 1.26	0.67 0.91 1.11	0.69 0.95 1.15	0.51 1.12 1.31	0.51 1.09 1.32	0 1.27 1.39	0.33 1.11 1.35	0.44 1.09 1.26	0.37 1.18 1.31	0.44 1.09 1.26	0 1.24 1.42	CC MAE RMSE
Overall performance	**0.74** **0.88** **1.04**	0.55 1.06 1.24	**0.75** **0.92** **1.08**	**0.70** **0.83** **1.00**	0.64 1.06 1.28	**0.74** **0.83** **0.98**	0.47 1.06 1.27	0.48 1.13 1.31	**0.76** **0.81** **0.99**	0.57 1.00 1.24	0.66 0.99 1.18	0.65 0.97 1.16	0.65 0.97 1.18	0.67 0.99 1.15	0.62 0.97 1.16	0.51 1.08 1.24	0.53 1.00 1.21	0.25 1.13 1.33	0.38 1.15 1.33	0.61 0.97 1.16	0.42 1.08 1.32	0.62 **0.95** **1.14**	0.37 1.12 1.32	CC MAE RMSE

Figure 14.0.1.: Classifiers performance for valence

If we analyse carefully Figure 14.0.1, which presents the performance of various classifiers for valence, we conclude the following: support vector regression, least mean squares and regression by discretization obtained the best performances in the first experiment; linear regression, M5R and least mean squares obtained the best performances in the second experiment; linear regression, pace regression and support vector regression obtained the best performances in the third experiment. In general, if we consider the mean of the results obtained in the three experiments, support vector regression, pace regression and linear regression obtained the best performances.

If we analyse carefully Figure 14.0.2, which presents the performance of various classifiers for valence, we conclude the following: Gaussian process, multilayer perception and support vector regression obtained the best performances in the first experiment; least mean squares, additive regression and bagging obtained the best performances in the second experiment; Gaussian process, support vector regression and linear regression obtained the best performances in the third experiment. In general, if we consider the mean of the results obtained in the three experiments, support vector regression, Gaussian process and radial basis function obtained the best performances

An overall analysis allows us to conclude that, on the one hand, function-based models like support vector regression and Gaussian processes are the ones that perform better; and on the other hand, rule-based and tree-based models are the ones that perform worst. This may be explained by the robustness of the function-based models and lack of it on the other models.

Classifier	GP	IR	LMS	LR	MP	PR	RBF	SLR	SMO	IBK	KS	LWL	AR	BAG	ES	RSS	RD	CR	DT	M5R	DS	M5P	REP	Measure
First experiment cross-validation	**0.82**	0.16	0.26	0.64	**0.80**	-	0.62	0.44	**0.77**	**0.75**	0.63	0.55	**0.76**	0.49	0.42	0.14	0.50	0.17	0.63	0.56	0.20	0.56	0.10	CC
	0.89	1.32	1.75	0.98	**0.77**	-	0.98	1.15	**0.86**	**0.67**	**0.85**	0.92	**0.67**	1.07	1.21	1.34	1.22	1.44	0.97	1.11	1.22	1.11	1.39	MAE
	1.13	1.86	2.68	1.33	**1.09**	-	1.29	1.52	**1.06**	**1.11**	1.31	1.46	**1.09**	1.42	1.48	1.64	1.52	1.73	1.28	1.44	1.81	1.44	1.68	RMSE
First experiment training/test split	**0.89**	0.38	**0.81**	**0.84**	**0.95**	-	**0.98**	0.30	**0.94**	0.40	0.21	0.38	0.37	0.61	0.39	0.54	0.24	0.52	0.44	**0.88**	0.52	**0.88**	0.52	CC
	0.95	1.07	1.09	1.02	**0.86**	-	**0.25**	1.32	**0.95**	1.26	1.45	1.15	1.15	**0.89**	**0.98**	1.05	1.32	**0.96**	1.11	0.99	**0.93**	0.99	**0.95**	MAE
	1.26	1.80	**1.28**	**1.25**	**1.02**	-	**0.30**	1.93	**1.13**	1.77	1.99	1.73	1.90	1.47	1.63	1.51	1.79	1.56	1.61	**1.23**	1.61	**1.23**	1.64	RMSE
Second experiment cross-validation	**0.81**	**0.80**	**0.79**	**0.77**	0.74	**0.77**	0.74	0.61	**0.77**	0.71	0.73	0.75	**0.80**	**0.79**	**0.80**	0.76	0.75	0.69	0.68	0.76	0.73	**0.77**	0.76	CC
	0.79	**0.76**	**0.81**	**0.81**	0.90	0.82	0.88	0.97	0.84	0.96	0.89	0.86	**0.75**	**0.77**	**0.75**	0.88	0.86	0.96	0.97	0.85	0.90	0.82	0.83	MAE
	0.96	**0.97**	**1.02**	**1.02**	1.12	**1.03**	1.07	1.32	1.05	1.20	1.10	1.05	**0.96**	**0.97**	**0.96**	1.04	1.09	1.15	1.18	1.06	1.08	**1.03**	1.04	RMSE
Second experiment training/test split	**0.85**	0.79	0.81	0.84	**0.86**	0.84	**0.85**	0.76	**0.85**	0.79	0.83	0.81	0.81	0.83	0.79	0.79	**0.86**	0.79	0.57	**0.86**	0.79	**0.87**	0.78	CC
	0.92	0.89	**0.86**	0.90	0.96	0.90	**0.84**	1.12	0.89	0.89	0.87	0.88	**0.82**	**0.85**	0.91	0.94	**0.74**	0.92	1.20	**0.85**	0.90	**0.85**	0.93	MAE
	1.07	1.08	**1.04**	1.06	1.15	1.06	1.05	1.27	**1.04**	1.08	1.05	1.08	**1.04**	**1.02**	1.09	1.11	**0.92**	1.13	1.45	**0.98**	1.11	**0.98**	1.12	RMSE
Third experiment cross-validation	**0.71**	0.62	**0.68**	**0.70**	0.56	**0.70**	**0.68**	0.64	**0.71**	0.44	0.67	0.66	0.60	0.63	0.67	0.67	0.64	**0.68**	0.63	**0.70**	0.61	**0.70**	0.62	CC
	0.79	0.90	0.86	**0.82**	1.04	**0.82**	**0.82**	0.87	**0.81**	1.27	0.87	0.84	0.94	0.88	0.83	0.86	0.91	**0.81**	0.90	**0.82**	0.88	**0.82**	0.90	MAE
	1.00	1.11	1.04	**1.00**	1.30	**1.01**	1.03	1.09	**0.99**	1.52	1.07	1.07	1.18	1.10	1.05	1.05	1.13	1.04	1.11	**1.00**	1.12	**1.00**	1.11	RMSE
Third experiment training/test split	**0.71**	0.67	**0.70**	**0.70**	0.66	**0.70**	0.66	0.63	**0.72**	0.48	0.69	0.64	0.59	**0.71**	**0.70**	0.62	0.58	0.62	0.63	**0.70**	0.65	**0.70**	0.59	CC
	0.69	0.72	0.72	**0.70**	0.77	**0.70**	0.72	0.76	**0.69**	1.24	0.74	0.77	0.83	**0.70**	**0.69**	0.79	0.79	0.74	0.78	**0.70**	0.72	**0.70**	0.82	MAE
	0.86	0.92	0.89	**0.88**	0.96	**0.88**	0.93	0.95	**0.86**	1.48	0.95	1.01	1.11	**0.87**	0.89	1.00	1.04	0.95	0.99	**0.88**	0.93	**0.88**	1.04	RMSE
Overall performance	**0.80**	0.57	0.68	**0.75**	**0.76**	**0.75**	**0.76**	0.56	**0.79**	0.60	0.63	0.63	0.66	0.68	0.63	0.59	0.60	0.58	0.60	0.74	0.58	**0.75**	0.56	CC
	0.84	0.94	1.02	0.87	0.88	**0.81**	**0.85**	1.03	**0.84**	1.05	0.95	0.90	**0.86**	**0.86**	0.90	0.98	0.97	0.97	0.99	0.89	0.93	0.88	0.97	MAE
	1.05	1.29	1.33	**1.09**	1.11	**1.00**	**0.95**	1.35	**1.02**	1.36	1.25	1.23	1.21	1.14	1.18	1.23	1.25	1.26	1.27	1.10	1.28	**1.09**	1.27	RMSE

Figure 14.0.2.: Classifiers performance for arousal

15. Calibration and Validation

The system was calibrated/validated in two types of experiments: ratings and physiological. Unlike previous experiments, in this new series of experiments we intended to obtain experimental data in a controlled environment. Emotional data obtained from these experiments was used to refine the knowledge base. Special attention was devoted to the identification of the weights of the musical features that compose it.

15.1. Rating Experiment

The rating experiment was developed with the objective of calibrating/validating the musical output of the system. We prepared a sample of music that, according to EDME's classification, covered all the quadrants of the bi-dimensional space. We used Superlab software (Haxby et al., 1993) to prepare the experiments.

15.1.1. First Experiment

15.1.1.1. Objective

We intended to verify the accuracy of EDME in classifying valence and arousal by using experimental data obtained in a controlled environment. This experiment aimed to examine the fit of the expected locations of music to its observed locations. This was done by classifying music through ratings made with naive listening subjects. This experiment was also dedicated to the refinement of the knowledge base. We came up with the hypothesis that a maximum of 13 features, which resulted from the phase of knowledge base systematization (chapter 13), was enough to discriminate valence and arousal of music. We intended to identify from this group of features a smaller subset for each emotional dimension, as well as to identify the weights for the features.

15.1.1.2. Data

Data consists of the selected musical segments and obtained emotional answers from the listeners.

Music The 30 musical segments used in this experiment lasted between 5 and 30 seconds. These segments belonged to four different genres of music: 16 of classical music, 4 of pop music, 7 of rock music and 3 of soundtrack.

Emotional Answers 30 listeners participated in this experiment: 20 male and 10 female aged between 18 and 23 years old (mean of 20, standard deviation of 2). They had background in informatics and technology. We calculated the mean and standard deviation for the emotional responses obtained in the questionnaire, as is shown in Figures 12.5.1 and 12.5.2. Mean and standard deviations were computed first amongst listeners, and then averaged over segments.

15.1.1.3. Method

We selected 40 MIDI files of western tonal music (classical, rock, pop and soundtrack genres) from a large database of pre-composed music of various genres. These files were selected based on its musical quality. Selected files went through the processes of segmentation and feature extraction. From this, a group of 193 segments labeled with musical features was obtained. The regression models built in the phase of knowledge base systematization (chapter 13) were used to classify each segment with an appropriate emotional label. From this group of 193 segments emotionally classified, we selected 30 segments covering all the quadrants of the bi-dimensional emotional space.

This experiment was carried out in a room with six desktops and respective headphones. Each user had one individual session that lasted, approximately, 10 minutes. In each session the user was guided by five screens of instructions (in portuguese). The first screen (Figure 15.1.1) gave general instructions about each session. It said that each user has to listen to several musical segments and afterwards he had to evaluate that segment in two dimensions represented in figures of the Self-Assessment Manikin (described in section 8.3) (Bradley and Lang, 1994). One dimension was related to the positive/negative effect of music that corresponded to valence. The other dimensions were related to the calm/exciting effect of music that corresponds to arousal.

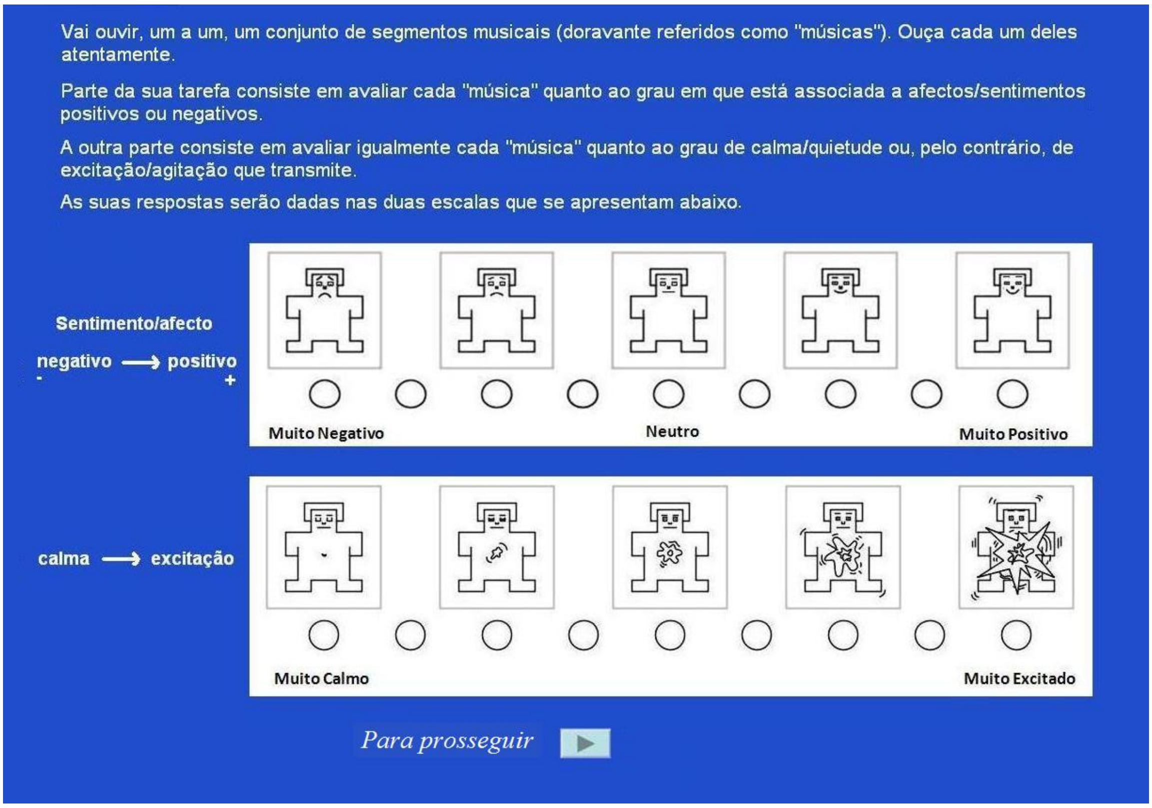


Figure 15.1.1.: General instructions giving information about what each session consists in

The second screen (Figure 15.1.2) gave detailed instructions about the dimension of valence. It emphasized the differences in the mouth and eyebrows of each of the five pictures of the Self-Assessment Manikin. Based on these differences it guided the user on the selection of the circles below the pictures that best reflected his evaluation of the valence of the listened music.

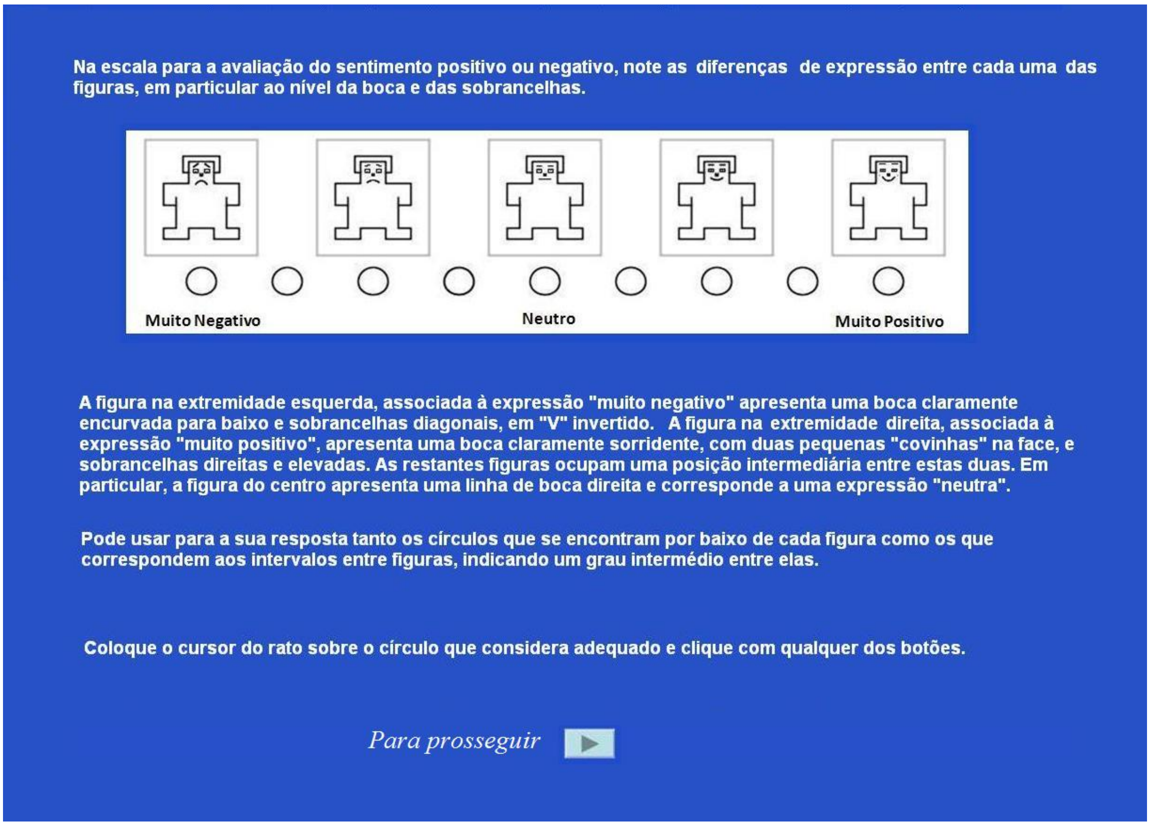


Figure 15.1.2.: Instructions about the selection of the valence of music

The third screen (Figure 15.1.3) gave detailed instructions about the dimension of arousal, emphasizing the differences in the eyes, eyebrows and ”lines of energy” of each of the five pictures of the Self-Assessment Manikin. As for valence, based on these differences it guided the user on the selection of the circles below the pictures that best reflect his evaluation of the arousal of the listened music.

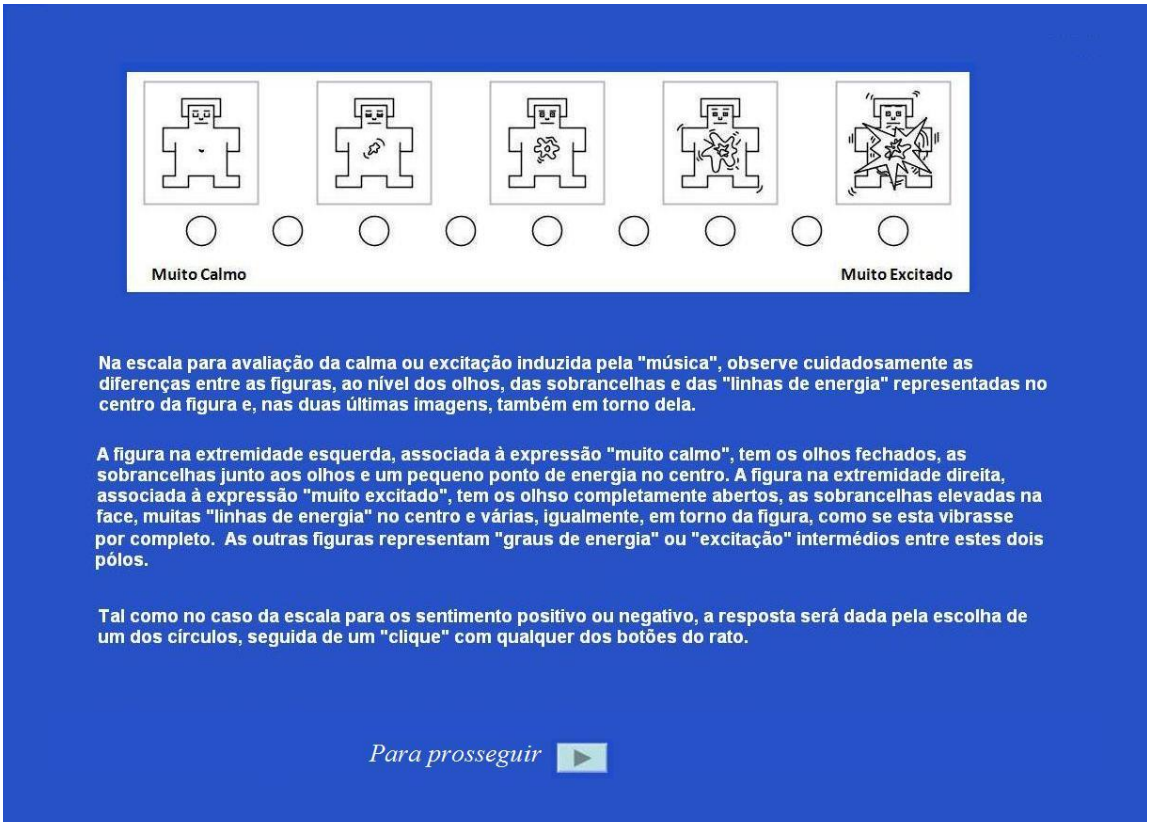


Figure 15.1.3.: Instructions about the selection of the arousal of music

The fourth screen (Figure 15.1.4) gave instructions about the process of listening to the music and skipping from one music piece to another one. It is possible to listen to the piece one or more times. This all depends if the user selected the button to listen again to the music again, or if he/she selected the other button which would guide the person to answer the valence and arousal dimension of the music. These buttons were inside a text box. This screen also shown to the user that he/she had a period of training that consists in listening to five pieces of music and answering the corresponding values of the emotional dimensions. After training, the user had to tag each of 30 musical segments with the desired ratings for valence and arousal.

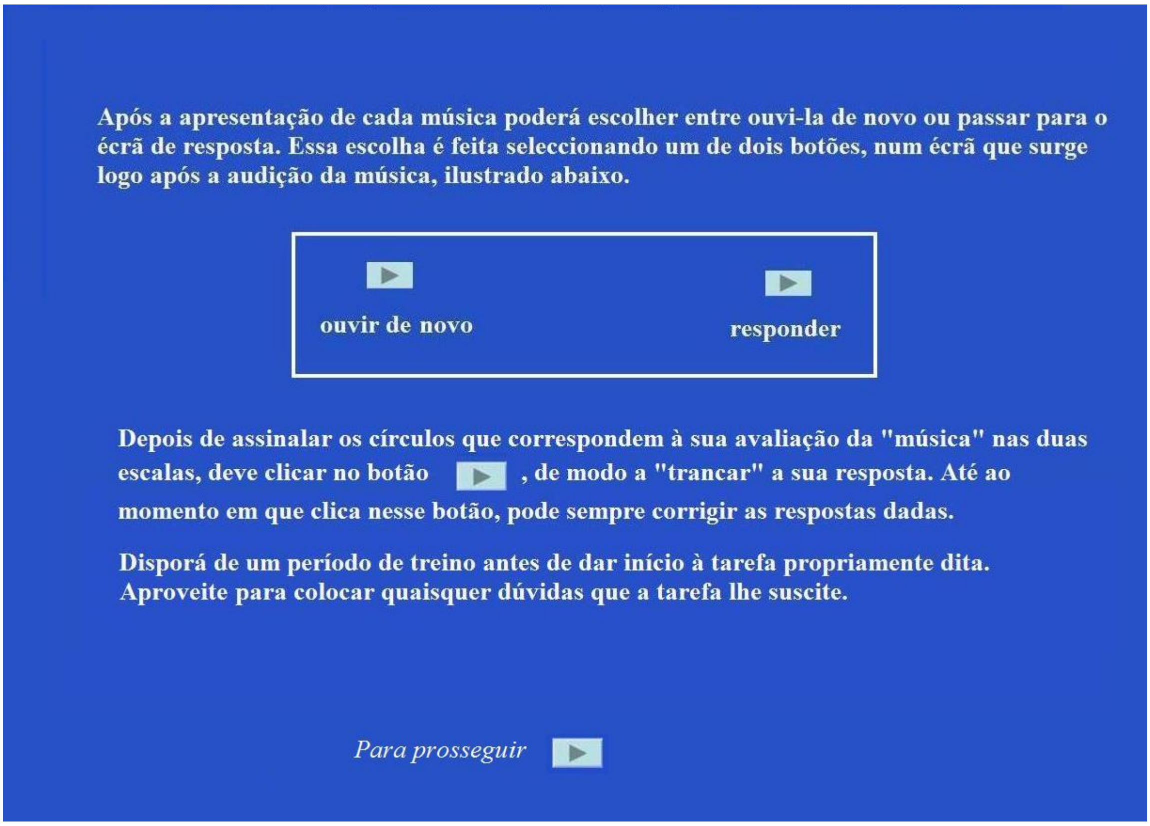


Figure 15.1.4.: Screen that guides the user while listening to one music piece and skipping to the next one

The fifth screen (Figure 15.1.5) appeared after listening to each music sample. In this screen the user had to select the circle that best fitted the desired value for valence (above in the Figure 15.1.5) and arousal (below in the Figure 15.1.5). There were nine possible choices for each of the dimensions. After selecting the desired ratings for each dimension, the user had to click in the button shown below the pictures, in order to listen to other music.

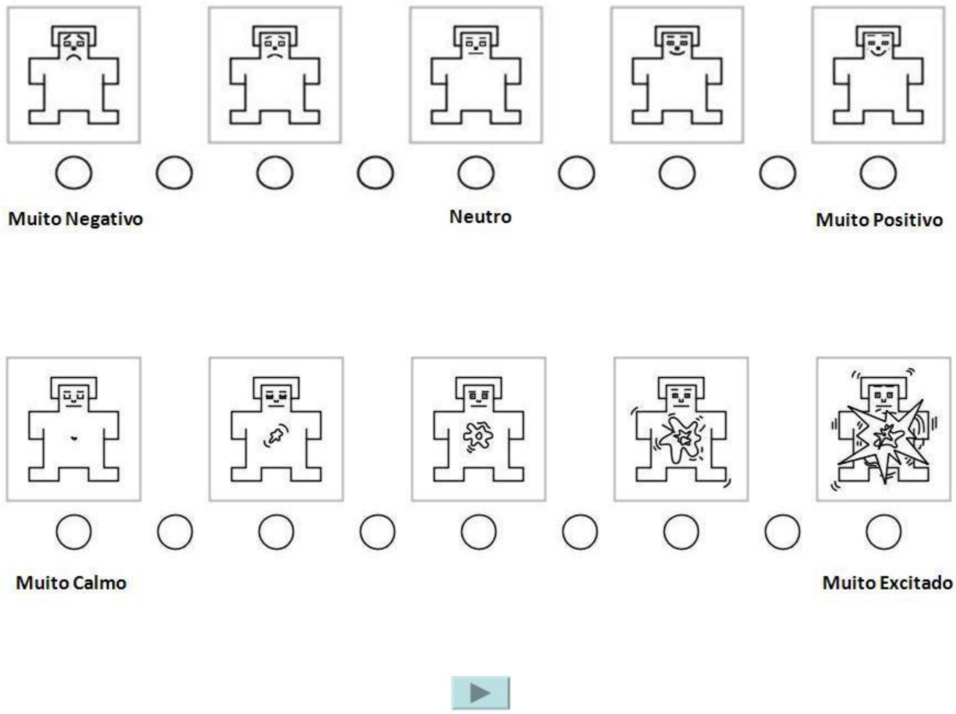


Figure 15.1.5.: Screen where the user rates valence and arousal of each music

The preparation of the musical material, getting the experimental data, the analysis of both the music material and experimental data, and other stages followed the method described in chapter 12 and presented in Figure 12.1.1.

15.1.1.4. Statistical Data

We calculated the mean and standard deviation for the emotional responses obtained with the Self Assessment Manikin (Bradley and Lang, 1994, described in section 8.3), as is shown in Figure 15.1.6, which presents the mean and standard deviation for emotional responses obtained. Mean and standard deviations were computed first amongst listeners, and then averaged over segments.

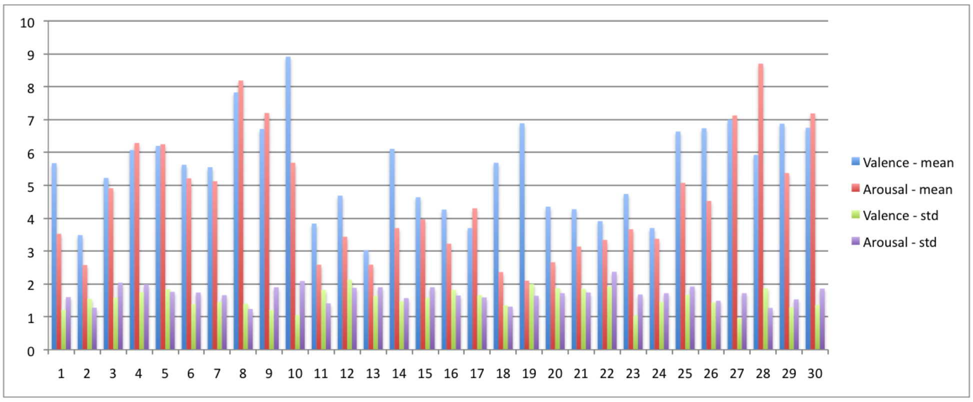


Figure 15.1.6.: Mean and standard deviations of the emotional responses in the first experiment of calibration/validation

In order to have a visual idea of the emotional distributions obtained with the listeners' answers and system's answers we built the scatter chart for both. Figure 15.1.7 presents the scatter chart for the listeners' answers, Figure 15.1.8 presents the scatter chart for system's answers. These charts allow us to see how the emotional space is covered, but also to discover similarities and differences among them. The main visible difference is that the users do not tend to answer with values of valence close to 0, which sometimes happens with the system. The scatter chart of system's answers is similar to the ones obtained in web-based experiments (Figure 13.2.1). In order to have numeric relations between the emotional distributions of listeners' answers and system's answers, we have calculated the correlation coefficients, mean absolute error and root mean square error between the listeners' and system's answers for each axis (valence and arousal). We obtained, respectively, the values of 0.75, 0.81 and 0.97 for valence and the values of 0.85, 0.86 and 1.08 for arousal.

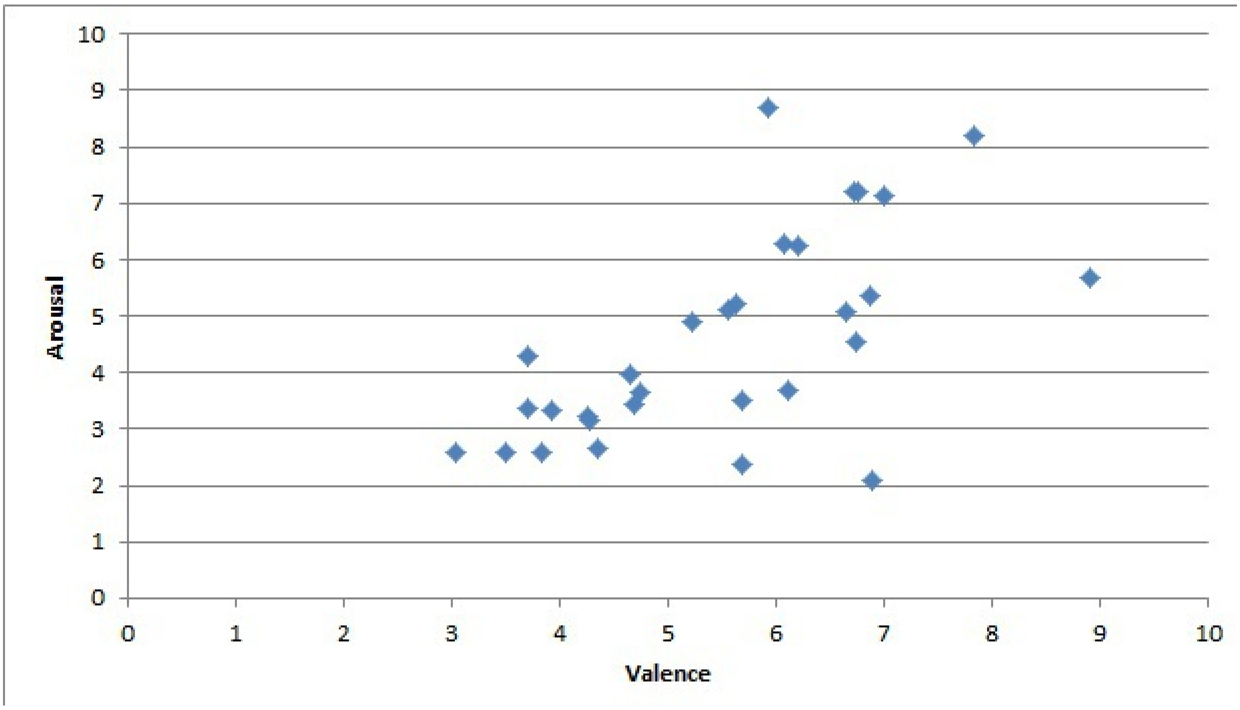


Figure 15.1.7.: Emotional distribution of listeners' answers (points represent mean values for each piece of music)

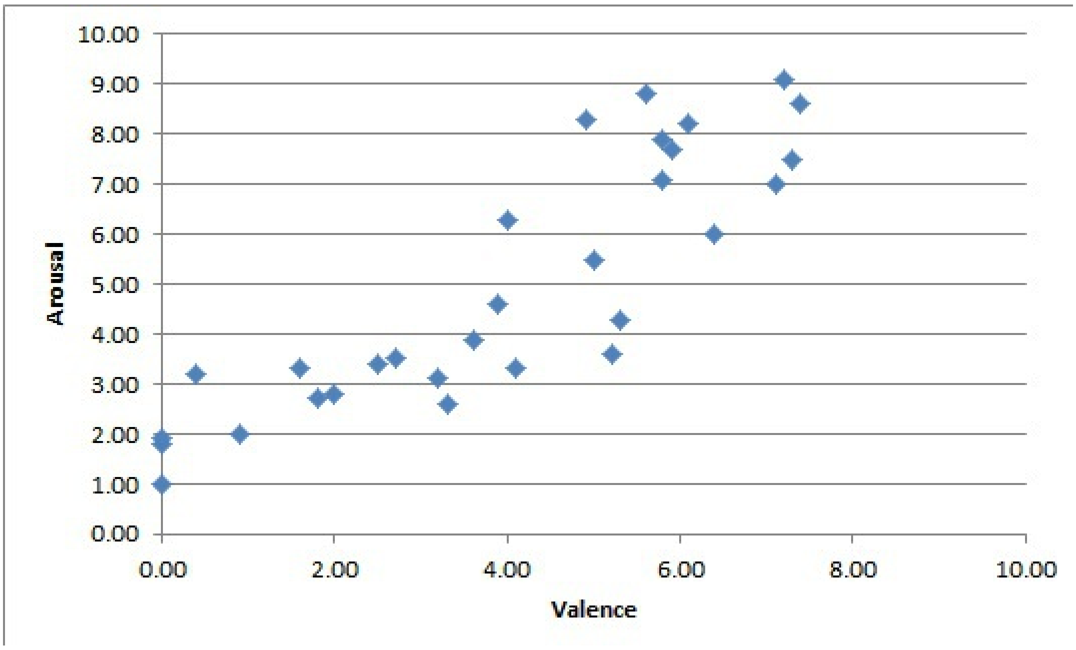


Figure 15.1.8.: Emotional distribution of system's answers (points represent values for each piece of music)

15.1.1.5. Results

In this experiment we identified the emotional relevance of 13 features. These features resulted from a phase of systematization of the knowledge base (described in chapter 13) and were the ones considered the most discriminant in previous experiments (tables 12.4, 12.8, 12.10, 12.9, 12.15 and 12.20), in literature (chapter 6) and in similar studies (section 8.4).

Feature Ranking We have calculated individually the correlation coefficient between each feature and the two emotional dimensions. Table 15.1 presents the feature, its description and the correlation coefficient between the feature and valence. Rhythmic (e.g, average note duration and average time between attacks) and texture features (e.g., spectral dissonance) are of particular relevance to valence, because of its high values of correlation.

Musical feature	Correlation Coefficient
Average Note duration	-0.63
Average Time Between Attacks	**-0.78**
Importance of Bass Register	0.40
Tempo	**0.50**
Note Density	0.54
Percussion Prevalence	0.44
Repeated Notes	**0.41**
Variation of Dynamics	**0.27**
Key mode	**-0.21**
Spectral loudness	0.31
Spectral dissonance (Sethares)	**0.72**
Spectral sharpness (Ambres)	**0.39**
Spectral similarity	-0.47

Table 15.1.: Correlation between features and valence, in bold style we have the best features of Table 15.4

Table 15.2 presents the feature, its description and the correlation coefficient between the feature and arousal. Rhythmic (e.g., average note duration, average time between attacks and tempo), melodic (e.g., repeated notes) and texture features (e.g., spectral dissonance, spectral sharpness and spectral similarity) are of particular relevance to arousal, because of its high values of correlation.

Musical feature	Correlation Coefficient
Average Note duration	**-0.55**
Average Time Between Attacks	-0.61
Importance of Bass Register	**0.56**
Tempo	**0.68**
Note Density	**0.47**
Percussion Prevalence	0.41
Repeated Notes	0.63
Variation of Dynamics	0.31
Key mode	-0.10
Spectral loudness	**0.53**
Spectral dissonance (Sethares)	**0.62**
Spectral sharpness (Ambres)	0.62
Spectral similarity	-0.60

Table 15.2.: Correlation between features and arousal, in bold style we have the best features of Table 15.4

As a matter of curiosity, we calculated the correlation among all the features and presented in Table 15.3 those with the highest correlations (>0.5 or <-0.5). There are several features with high correlation, which indicates that there is a relatively high colinearity among them.

Musical feature	Musical feature	Correlation Coefficient
Average Note Duration	Average Time Between Attacks	0.81
	Spectral Sharpnes (Ambres)	-0.64
	Spectral Similarity	0.55
Average Time Between Attacks	Note Density	-0.53
	Percussion Prevalence	-0.52
	Spectral Loudness	-0.51
	Spectral Dissonance (Sethares)	-0.62
	Spectral Sharpness (Ambres)	-0.73
	Spectral Similarity	0.54
Importance of Bass Register	Percussion Prevalence	0.58
	Repeated Notes	0.59
	Spectral Dissonance (Sethares)	0.69
Tempo	Repeated Notes	0.51
Note Density	Spectral Dissonance (Sethares)	0.51
Percussion Prevalence	Repeated Notes	0.7
	Spectral Dissonance (Sethares)	0.82
	Spectral Sharpness (Ambres)	0.52
	Spectral Similarity	-0.5
Repeated Notes	Spectral Dissonance (Sethares)	0.70
	Spectral Sharpness (Ambres)	0.54
	Spectral Similarity	-0.51
Spectral Loudness	Spectral Sharpness (Ambres)	0.72
Spectral Sharpness (Ambres)	Spectral Similarity	-0.54

Table 15.3.: Correlation between features emotionally more discriminant

Feature Selection and Classification We applied the best first search method Witten et al. (1999) on the set of features to select those emotionally more discriminant. We made a compromise between the number of features and the quality of the results. Then, we applied 10-fold cross-validation on the set of features emotionally more discriminant. The results are presented in Table 15.4. The features considered in this table are in bold style in Tables 15.1 and 15.2.

Emotional dimension	CC	MAE	RMSE	Best features	Weight
Valence	0.85	0.61	0.74	Average Time Between Attacks	-0.54
				Tempo	0.23
				Repeated Notes	-0.16
				Variation of Dynamics	0.17
				Key Mode	-0.06
				Spectral Dissonance	0.37
				Spectral Sharpness	-0.37
Arousal	0.83	0.77	1.01	Average Note duration	-0.19
				Importance of Bass Register	0.37
				Tempo	0.37
				Note Density	0.33
				Spectral Loudness	0.14
				Spectral Dissonance	0.06

Table 15.4.: Results of 10-fold cross-validation for valence and arousal – first experiment of calibration/validation

15.1.1.6. Statistical Analysis

We proceeded to the statistical analysis of the system classification and listeners' classification of the quadrants of the 30 musical pieces. We used SPSS Statistics software to do this (Field, 2009). Kappa and Cramer's V were used as statistical measures. The results of the interrater analysis are Kappa = 0.688 with $p < 0.0001$. This measure of agreement, while statistically significant, is substantially convincing (Landis and Koch, 1977). We obtained a value of 0.766 for Cramer's V, which according to the literature[45] shows us that the two variables (classification of the system and classification of the listeners) are probably measuring the same concept.

15.1.1.7. Discussion

From the analysis of Tables 15.1 and 15.2 it seems that the variation (of some features), as expressed by repeated notes, variation of dynamics and spectral (di)similarity, contribute to an increase of both valence and arousal.

The high values of correlation of Table 15.3, allow us to make some conclusions with a degree of confidence. The percussion line of a musical pieces seems to be more important than the melodic, harmonic and bass lines in dictating the rhythm of the music. The higher the prevalence of percussion, the lower the time between attacks. A high presence of percussion and repeated notes increase dissonance of music.

[45] http://homes.chass.utoronto.ca/~josephf/pol242/LM-3A#Stage%20I:%20%20Phi

After analysing Table 15.4 we came to some conclusions. In general, the set of features for both valence and arousal included the features emotionally more discriminant when correlated alone with the respective emotional dimensions (Tables 15.1 and 15.2). From these results, it seems that the 13 considered features can discriminate well both valence and arousal of each music. As a result, we can infer that the experiments conducted via online have a high degree of reliability, despite the fact of being made in a non-controlled context. The correlations coefficients of 0.85 and 0.83, respectively, for the classification of valence and arousal are significant.

From the analysis of Table 15.4, we have average time between attacks, spectral dissonance and spectral sharpness with the highest weights in the classification of valence, and importance of bass register, tempo and note density with the highest weights in the classification of arousal. Tempo and spectral dissonance are common features used in the classification of the emotional dimensions. Tempo and spectral dissonance have a positive influence as a result of the positive weights. Then, for valence, we have average time between attacks, repeated notes, key mode and spectral sharpness with a negative influence, and variation of dynamics with a positive influence; for arousal, we have average note duration with a negative influence, and importance of bass register, note density and spectral loudness with a positive influence.

We can compare the results of this experiment with the results obtained in the chapter 13 of knowledge base systematization, which used this same group of features in the classification of emotional dimensions using experimental data obtained via web. Average time between attacks and tempo were always used in the classification of valence. In the case of the classification of arousal there are no features that are always used. Focusing only on the correlation coefficient, the classification of valence obtained the following results: 0.78, 0.68, 0.58, 0.62 and 0.85, which give us a mean of 0.70 which is a satisfactory value. Concerning the classification, we obtained the following correlation coefficients: 0.74, 0.81, 0.75, 0.77 and 0.83, which gives a mean of 0.78 which is also a satisfactory result.

Similar distributions to the ones presented in Figures 15.1.7 and 15.1.8 were obtained in the three experiments carried out via online. This is another point that allow us to have more confidence in the reliability of the experiments done online.

The statistical results using Kappa and Cramer's V do not only confirm the reliability of this calibration/validation study but also the reliability of the experiments conducted with the help of online questionnaires (chapter 12). This is visible in the similarity of the results obtained for the best features, as well as for the results of 10-fold cross-validation.

15.2. Physiological and Behavioral Experiment

This experiment intended to obtain additional data, in a controlled environment, to assess the relationship between the emotional output (valence and arousal rating) of the computational system for the emotional control (EDME) and the physiological response to the sounds. As a result, the text of this section was authored by them. Emotional reactions to different sounds were evaluated by behavioral (pleasure and arousal rating), and physiological measures (heart rate, skin conductance and facial electromyographic -EMG).

15.2.1. Method

This experiment was led by Alba Grieco[46] and Armando Oliveira[47] at the Faculdade de Psicologia e de Ciências da Educação. Our contribution was on the selection of music fragments to be used in the experiment, which should cover all the quadrants of the bi-dimensional emotional space according to EDME's classification, and on the analysis of the results in terms of the quality of EDME's classification. The details of the experiment are available in (Grieco and Oliveira, 2012).

15.2.1.1. Participants

A group of 27 (25 female) undergraduate subjects participated at the experiment.

15.2.1.2. Materials, Design and Procedure

Of the 48 sounds used in the experiment, 10 files were selected from the International Affective Digitalized Sounds (IADS) and the remaining 38 files were pieces of music selected from different musical styles (classical, soundtrack, pop and rock). The valence and arousal of the IADS files were selected in order to cover the four quadrants of the valence-arousal space. The same happened with the remaining 38 segments, according to EDME's classification.

Valence and arousal were obtained in a similar manner as in the previous experiment, using the paper and pencil version of the affective rating system Self-Assessment Manikin (SAM) (Lang, 1980).

Physiological data were acquired during 9s on each trial, corresponding to 500 ms of registration preceding the sound presentation, then the registration during the presentation of the sound (with duration varying from 4973 to 7580ms), and a registration

[46] http://vision.psy.unipd.it/grieco.htm
[47] https://woc.uc.pt/fpce/person/ppgeral.do?idpessoa=14

performed after the sound was turned off, with variable duration (from 500 to 3500 ms). After subjects evaluated valence and arousal a relaxing screen was presented for 15 seconds. The next sound was presented 2s after the starting-button was pressed. We collected data of facial EMG, heart rate and galvanic skin response.

15.2.2. Results

The results obtained during the experiment were subject to analysis in three perspectives. A detailed description can be consulted in (Grieco and Oliveira, 2012).

The first analysis focused on the variation of physiological measures evocated by sounds from IADS with the emotional a priori valence and arousal (Bradley and Lang, 1999). The aim of this analysis was to assess the quality of the results obtained in this experiment.

The second analysis focused on the variation of physiological measures obtained with the sounds from IADS with the emotional output from the EDME system. The analysis concludes that mean CORR amplitude decrease with valence; ZIG activity modestly increase with valence, BPM increase with valence, and the increase of BPM with arousal depend on the valence. Finally, outcomes show that GSR values increase with arousal.

The third analysis focused on the variation of physiological measures with affective EDME system' emotional output (valence and arousal). Overall the outcomes shown that the physiological responses elicited when listening to sounds correlates with a priori valence and arousal values, in agreements with Bradley and Lang (2000) results. When the relation between physiological responses elicited when listening to elaborates pieces of music (classical, pop, rock and soundtracks) and the emotional output from EDME system is evaluated the results show that they are weakly correlated.

Part IV.

Conclusion

16. Discussion

This thesis was a journey with different and complementary stages. All began with a motivation very close to the statement presented in the beginning of the thesis “Music can change the world because it can change people.”, whose author was Bono (U2). The idea of joining two multidimensional worlds (music and emotions) with a very significant impact in the society gave the ”fuel” for the beginning of the thesis. This ”fuel” led to the exploration of these two worlds. Different works in the areas of Music Psychology, Music Computing and Affective Computing were studied in order to discover possible contributions to the state of the art in these areas, as well as to have a clear idea of what the aim of the thesis would be.

16.1. State of the art

We have found many studies on the area of Music Psychology that helped us in bridging the gap between the two worlds. The empirical results of these works contributed to the development of the first knowledge base. This knowledge base as was said is composed by regression models that relate musical features the emotional dimensions of valence and arousal. We found particularly practical the model of emotions representation proposed by Russell (1989). Because musical features were represented numerically, a numeric representation of emotions would be desirable. This was made possible with the Russell’s model.

The works of Music Computing that were studied were particular useful to the definition of the architecture of the EDME. Different tasks of Music Computing led to the development of the modules of EDME: segmentation, features extraction, classification, selection, transformation, sequencing and synthesis. The discovery of third party software that could facilitate the accomplishment of the objective of this thesis occupied a relevant portion of time. The module of feature extraction was the one that gained more from using third party software (McKay and Fujinaga, 2006; Eerola and Toiviainen, 2004; Sorensen and Brown, 2000; Lartillot and Toiviainen, 2007; Cabrera, 1999).

The works on Affective Computing which we studied were useful in making a clear vision of what was already done in order to accomplish the objective proposed in this thesis. There are four proposed approaches in order to accomplish it. The works based on the automatic composition are generally conceived for a bounded range of musical styles, and sometimes do not tackle the whole composition process. We desired to have the flexibility of producing complete music pieces in a wide range of styles, so this approach was not very suitable. The studies grounded on classification of pre-composed music and subsequent selection were scalable, but the quality of their answers is very dependent of the original music base. This one is, actually, a finite database, and thus cannot cover entirely the whole emotional spectrum. Therefore, one has to expect to select pieces that do not match exactly the intended emotion. The approach based on transformation has the disadvantage of producing outputs with low quality when the original music has characteristics very different from the desired ones. None of these three approaches, alone, gives an entirely satisfactory response to our requirements. The fourth approach consists in the hybrid combination of the former ones in order to overcome some of their weaknesses. For the purpose of our work, we found especially promising a particular hybrid approach that consists in combining classification/selection with transformation. In fact, the transformation can improve the classification/selection result when there is not a solution in the music base close to the emotional specification. On the other hand, as the selection tends to produce an output with characteristics close to the desired ones, the transformation assumes less risks of degrading music quality, because the adjustments needed to get the music characteristics fit the emotional specification are limited.

16.2. Experiments

After having a clear idea of the works more relevant to this thesis, from the areas of Music Psychology, Music Computing and Affective Computing, and after doing a first version of EDME ready to be used, we proceeded to conduct some experiments with the objective of improving the knowledge base, and more properly to obtain experimental data that could allow us to relate emotional and music domains with the help of the Weka software (Witten et al., 1999). We carried out three experiments, spread in different parts, in some cases. The first experiment consisted in analysing and selecting the first set of (MIDI) features emotionally relevant. The second experiment was an extension of the first experiment, which allowed used to analyse and select another set of features from a bigger group of features. The number of listeners and the number of music files were other variables that were extended. This second experiment also consisted in the analysing the first set of audio features with emotional impact, particularly the selection of the instruments samples used in the synthesis of the MIDI music. Another part of this experiment consisted in a preliminary evaluation of the selection and transformation modules. The third experiment also consisted in trying to find the set of (MIDI and audio) features with the most impact of the emotional dimensions. This experiment also contributed to the verification of the effectiveness of the algorithms of transformation. We were particularly successful in this aspect, as was observed from the analysis of the results of this experiment. Tempo, pitch register, musical scales, instruments and articulation all have a degree of importance in shifting the emotional content of music.

16.3. Systematization and Evaluation

By the end of the experiments we felt ready to two other stages needed before proceeding to the phase of calibration/validation of the EDME system. The first stage consisted in systematizing the knowledge base. We analysed all the results obtained in the three experiments and made a knowledge base that best bridged the semantic gap between the emotional and musical domains. The other stage consisted in making an extensive evaluation of different types of classifiers. After making a systematization of a small group of features emotionally relevant we were ready to make this evaluation. Function-based classifiers (Witten and Frank, 2005) were the ones that achieved the best results. From this group of classifiers we highlight the SVM regression classifier, because of its better results.

16.4. Calibration/Validation

The last experimental stage consisted in calibrating/validating EDME. This stage was divided in two experiments. The first experiment collected data with questionnaires based on Self-Assessment Manikin (Bradley and Lang, 1994) that were developed in the Superlab software (Haxby et al., 1993). Unlike previous web-based experiments, in this first experiment of calibration/validation we intended to obtained experimental data in a controlled environment. We intended to verify the accuracy of EDME in classifying valence and arousal by using experimental data obtained in a controlled environment. From the results of these experiments, it seems that the 13 considered features can discriminate well both valence and arousal of each music. As a result, we inferred that the experiments conducted via online had a high degree of reliability, despite the fact of being done in a non-controlled context. We also obtained similar distributions between the ones obtained from the emotional answers of the web-based experiments and the ones obtained in the controlled context. This is another point that allowed us to have more confidence in the reliability of the experiments done online. The statistical results using Kappa and Cramer's V also confirmed the reliability of the first experiment of calibration/validation and of the experiments made with the help of online questionnaires.

At a later stage, we assessed the experienced emotions in listeners by collecting psychophysiological data and by recording facial expressions.

The second experiment called "physiological and behavioral" led us to the conclusion that the emotional output from EDME system is weakly correlated. However, although the effects were not significant (with $p \leq 0.050$) the data show that corrugator muscle activity increase with arousal; heart rate measure in beats per minute increase with arousal, and galvanic skin response increase with both valence and arousal. Only for zigomatic muscle activity there is a significant increase with both, valence and arousal.

16.5. Application

In the meantime we also dedicated some time to the development of an installation that could test the interactive abilities of EDME (Ventura et al., 2009). At the core of the installation there was an affective computer system that selected appropriate music and images to express its emotional state. Music was selected using EDME, images were selected with another engine. The installation allowed people to experience and influence the emotional behavior the affective computer system. We conducted two experiments where people were able to ascribe emotions to the the system in a natural way. From the preliminary results, we carefully concluded that both music and images were effective and important in transmitting the emotional state of the affective computer system. Extended experiments would be needed to have clear certainty of this conclusion.

16.6. Contributions

As a whole we can conclude that EDME is a music production system that expresses the desired emotions. From its implementation resulted several advances to the state-of-the-art. It implements algorithms that control emotional content of music in different levels: segmentation, classification, selection and transformation. The knowledge base, one of the auxiliary structures, systematizes relations between emotions and musical features. It is also composed by an interface that allows different types of emotional representation. The flexibility of the architecture and the use of parameterizable structures widen the areas of application of EDME. The system was already applied in an affective installation, but we also intend to demonstrate the usability of EDME in healthcare and soundtrack generation, which leads us to the next chapter.

17. Future Work

17.1. Update of the transformation module

We should design new algorithms in order to have one algorithm for each of the features used in the classification module. With these algorithms developed, a new experiment shall be designed to test their effectiveness in approximating the emotional content of music to the desired emotion of the listeners. The final stage of this process shall test the effectiveness of the regression models used by the classification module in the transformation. By doing this, the transformation module will be ready to be used by EDME in the way we designed it.

17.2. Soundtrack generation

One of the most promising fields of application of EDME is the production of soundtracks for narrative contexts. Music has become an integral part of the emotional, immersive gaming experience. One can envisage EDME managing the musical component of a computer game, adapting dynamically to the game conditions by matching music to action in real time. Soundtrack composition for movies can also become simpler: given a script annotated with the emotions, the system may produce music accordingly. We intend to demonstrate the applicability of our system in such contexts by integrating it with EmoTag (Francisco and Hervas, 2007), a system developed by the Instituto de Tecnología del Conocimiento team that makes automated mark up of emotional information in texts. This system is capable, for example, to annotate narrative texts like scripts with information about the emotions derived from the text. We intend to develop a prototype that will integrate both systems in order to be possible to demonstrate the feasibility of automatically producing music that is emotionally consistent with a given text.

17.3. Healthcare

The application of systems like ours has been done in the healthcare domain (Wingstedt et al., 2005). We intend to demonstrate the usability of EDME in a healthcare context. EDME will be tested with patients of the paediatrics service of the Hospital de Santo André de Leiria (HSAL). The several musicians that use this service do not have an automatic method to help in the selection of music for each healthcare situation. The use of EDME can be very important to overcome this problem and to promote the use of music as a medical tool in the whole hospital. It is a precious tool not only for patients, but also for the family, doctors and nurses by promoting a desired emotional ambiance.

The interface of EDME is prepared to be used by professionals of the HSAL in several experiments. In each experiment, the system reproduces in an audio format song-like structures formed by musical segments selected from a personalized database of pre-composed music. Experiments are focused in the analysis of the amount of deviation between the expected and obtained emotional effects on patients.

17.4. Emotionally-Driven Music Composition

Despite the fact of not using a module of music composition in our thesis we already did a review of the works developed in this area. Some of this review was already presented where we described the fourth approaches being used to accomplish the objective of this thesis. Another was presented in works done by me [48] (from pages 60 to 68) and Ivana Matic [49]. The work developed by Ivana Matic was grounded on the EDME system. She talked about composing melody that suits well to desired emotion using Neural Networks, Cellular Automata, tables with specific values and rules. After that, rhythm generation was also discussed.

[48] http://student.dei.uc.pt/~apsimoes/PhD/PhDThesisProposal.pdf
[49] http://eden.dei.uc.pt/~apsimoes/Automatic_composition.pdf

18. Accompanying CD-ROM

The source code of the system developed in this thesis has been put on a supplementary CD-ROM. Videos demonstrating the offline and online stages of our system is also part of the CD-ROM.

A. GLOSSARY

A.1. Music

MUSICAL FEATURE	DEFINITION
Tonality	Western tonal music rules
Rhythm	Variation of the duration of sounds over time
Melody	Series of linear events (pitches) or a succession
Harmony	Study of pitch simultaneity (e.g., chords)
Chord	Aggregate of musical pitches sounded simultaneously
Dynamics	Softness or loudness of a sound or note
Timbre	Perception of sound harmonics and onsets (attack transients)
Loudness	Sound pressure change (amplitude or intensity)
Pitch	Sound wave's frequency
Pitch range	Difference between highest and lowest pitches
Pitch variation	Amount of pitch change in the melody
Key	Pitch class from which the scale is built
Interval	Pitch step
Melodic contour/motion	Up and down pattern of pitch changes
Meter	Regular alternation of strong and weak beats in twos or threes, at many hierarchical temporal levels
Mode	Subset ot pitches used in a song
Articulation	Performance technique
Legato	Articulation used to play notes smoothly
Staccato	Articulation used to play notes distinctly
Vibrato	Quickly up and down of the pitch of notes
Attack/Note onset	Beginning of a musical note or other sound
Note	Musical sign used to represent the duration and pitch of sound
Texture	Overall sound (color) of a piece of music
Timing	Adjust the time of notes/beats to sound well

A.2. Description of music features

FEATURE	DESCRIPTION
ADSR envelope	Curve of Attack, Decay, Sustain and Release representative of the sound energy
Amount of arpeggiation	Fraction of horizontal intervals that are repeated notes, minor thirds, major thirds, perfect fifths, minor sevenths, major sevenths, octaves, minor tenths
Average duration accent	Average duration accent of the notes. It uses two variables. Tau variable represents saturation duration, which is proportional to the duration of the echoic store. Accent variable covers the minimum discriminable duration
Average melodic complexity	Expectancy-based model of melodic complexity based either on pitch or rhythm-related components or on an optimal combination of them together. It focus on tonal and accent coherence, and to the amount of pitch skips and contour self-similarity the melody exhibits
Average Note Duration	Average duration of notes in seconds
Average Note to Note Dynamics Change	Average change of loudness from one note to the next note in the same channel
Average Number of Independent Voices	Average number of different channels in which notes have sounded simultaneously. Rests are not included in this calculation
Average Time Between Attacks	Average time in seconds between Note On events (irregardless of channel)
Brass fraction	Fraction of Note Ons belonging to brass patches (including saxophones) (General MIDI patches 57 to 68)
Brightness (>1500Hz)	Amount of sound energy above the frequency of 1500 Hz
Brightness (>4000Hz)	Amount of sound energy above the frequency of 4000 Hz
Brightness (>400Hz)	Amount of sound energy above the frequency of 400 Hz
Climax position	Represents where the climax of the melody starts. The value is a percentage of the complete melody. The exact formula is the sum of the rhythm values of all notes prior to the climax, divided by the sum of all the rhythm values in the melody
Climax strength	Inverse of the count of the number of notes sharing the highest pitch
Consecutive identical pitches	Count of intervals whose size is 0 semitones
Distinct rhythm count	Number of rhythms that appear at least once
Dominant spread	Largest number of consecutive pitch classes separated by perfect 5ths that accounted for at least 9% each of the notes
Electric Guitar Fraction	Fraction of Note Ons belonging to electric guitar patches (General MIDI patches 27 to 32)

Electric Instrument Fraction	Fraction of Note Ons belonging to electric (non- “synth”) patches (General MIDI patches 5, 6, 17, 19, 27 to 32, 34 to 40)
Energy	The global energy of the signal x is computed simply by taking the root average of the square of the amplitude, also called root-mean-square
Harmonic mode	Estimation of the modality, i.e. major vs. minor, returned as a numerical value between -1 and +1
Importance of High Register	Fraction of Note Ons between MIDI pitches 73 and 127
Importance of Middle Register	Fraction of Note Ons between MIDI pitches 55 and 72
Importance of loudest voice	Difference between the average loudness of the loudest channel and the average loudness of the other channels that contain at least one note divided by 64
Inharmonicity	Amount of partials that are not multiples of the fundamental frequency, as a value between 0 and 1. More precisely, the inharmonicity considered here takes into account the amount of energy outside the ideal harmonic series
Interval strong. pitch classes	Absolute value of the difference between the pitches of the two most common pitch classes
Key	Returns the key according to the Krumhansl-Kessler algorithm. C major = 1, C# major = 2, ... c minor = 13, c# minor = 14, ...
Key Mode	Estimates the key mode (1=major, 2=minor) based on Krumhansl-Kessler key finding algorithm and pitch distribution
Loudness	Loudness is the subjective impression of the intensity of a sound, measured in sones. Specific loudness is the loudness attributable to an auditory filter. The specific loudness function extends from the low frequency filters to the high frequency filters
Melodic fifths	Fraction of melodic intervals that are perfect fifths
Melodic Tritones	Fraction of melodic intervals that are tritones
Most Common Melodic Interval Prevalence	Fraction of melodic intervals that belong to the most common interval
Most common pitch class prevalence	Fraction of Note Ons corresponding to the most common pitch class
Note Density	Average number of notes per second
Note prevalence english horn	Number of notes played using the MIDI patch corresponding to english horn divided by the total number of Note Ons in the piece
Note prevalence flute	Number of notes played using the MIDI patch corresponding to flute divided by the total number of Note Ons in the piece
Note Prevalence Fretless Bass	Number of notes played using the MIDI patch corresponding to fretless bass divided by the total number of Note Ons
Note Prevalence Muted Guitar	Number of notes played using the MIDI patch corresponding to muted guitar divided by the total number of Note Ons
Note prevalence orchestra hit	Number of notes played using the MIDI patch corresponding to orchestra hit by the total number of Note Ons in the piece

Note Prevalence Steel Drums	Number of notes played using the MIDI patch corresponding to steel drums divided by the total number of Note Ons
Note Prevalence Timpani	Number of notes played using the MIDI patch corresponding to timpani divided by the total number of Note Ons in the piece
Note prevalence of bass drum	Number of notes played using the MIDI patch corresponding to snare drum divided by the total number of Note Ons in the piece
Note prevalence of closed hi-hat	Number of notes played using the MIDI patch corresponding to closed hi-hat divided by the total number of Note Ons in the piece
Note prevalence of snare drum	Number of notes played using the MIDI patch corresponding to snare drum divided by the total number of Note Ons in the piece
Number of relatively strong pulses	Number of beat peaks with frequencies at least 30% as high as the magnitude of the bin with the highest magnitude
Number of Unpitched Instruments	Total number of MIDI Percussion Key Map patches that were used to play at least one note
Overall dynamic range	The maximum loudness minus the minimum loudness value
Percussion Prevalence	Total number of Note Ons belonging to percussion patches divided by total number of Note Ons in the recording
Pitch variety	Number of pitches used at least once
Polyrhythms	Number of beat peaks with frequencies at least 30% of the highest magnitude whose bin labels are not integer multiples or factors (using only multipliers of 1, 2, 3, 4, 6 and 8) (with an accepted error of +/- 3 bins) of the bin label of the peak with the highest magnitude. This number is then divided by the total number of beat bins with frequencies over 30% of the highest magnitude
Primary Register	Average MIDI pitch
Range of Highest Line	Difference between the highest note and the lowest note played in the channel with the highest average pitch divided by the difference between the highest note and the lowest note in the piece
Register	The octave position
Relative Strength Common Intervals	Fraction of melodic intervals that belong to the second most common interval divided by the fraction of melodic intervals belonging to the most common interval
Relative Strength of Top Pitch Classes	The magnitude of the 2nd most common pitch class divided by the magnitude of the most common pitch class
Relative Strength of Top Pitches	The magnitude of the 2nd most common pitch divided by the magnitude of the most common pitch
Repeated notes	Fraction of notes that are repeated

Repeated pitch density	Ratio between the count of consecutive notes of the same pitch and the count of all note to next note intervals
Rhythmic looseness	Average width of beat histogram peaks (in beats per minute). Width is measured for all peaks with frequencies at least 30% as high as the highest peak, and is defined by the distance between the points on the peak in question that are 30% of the height of the peak
Rhythmic variety	Ratio between the number of distinct rhythms and the total number of notes
Same direction interval	Count of consecutive intervals in the same direction
Saxophone Fraction	Fraction of Note Ons belonging to saxophone patches (General MIDI patches 65 to 68)
Spectral dissonance (H&K)	When applied to the compact spectrum, this feature measures the noisiness of the sound; when applied to the tonal components, it comes closer to measuring musical dissonance. This feature normalizes the results, and uses linear intensity
Spectral dissonance (Sethares)	When applied to the compact spectrum, this feature measures the noisiness of the sound; when applied to the tonal components, it comes closer to measuring musical dissonance. This feature does not normalize the results, and uses decibels
Spectral sharpness (Ambres)	Sharpness is a subjective measure of sound on a scale extending from dull to sharp. Aures' sharpness formula is a revision of Z&F's, so as to model the positive influence that loudness has on sharpness. Aures also uses a different g(z) function. Aures' formula is more sensitive to loudness than Zwickler formula
Spectral sharpness (Zwickler)	Sharpness is a subjective measure of sound on a scale extending from dull to sharp. Zwicker & Fastl's sharpness is calculated in the following manner - where N is loudness, N'(z) is specific loudness, z is the critical-band rate, and g(z) is a weighting function that emphasizes high frequencies
Spectral similarity	Spectral similarity calculates a similarity matrix with the help of MIR Toolbox in order to find the difference between consecutive frames of the frequency spectrum. It reflects the smoothness of the music (the changes of features along the music)
Spectral texture MFCC 2	Amount of energy presented on the second out of thirteen Mel-frequency cepstral coefficients
Spectral Texture MFCC 4	Amount of energy presented on the fourth out of thirteen Mel-frequency cepstral coefficients
Spectral Texture MFCC 6	Amount of energy presented on the sixth out of thirteen Mel-frequency cepstral coefficients
Spectral Texture MFCC 7	Amount of energy presented on the seventh out of thirteen Mel-frequency cepstral coefficients

Staccato incidence	Number of notes with durations of less than a 10th of a second divided by the total number of notes in the recording
Stepwise Motion	Fraction of melodic intervals that corresponded to a minor or major third
Strength of Strongest Rhythmic Pulse	Magnitude of the beat bin with the highest magnitude
Strength of two strong. rhythmic pulses	The magnitude of the higher (in terms of magnitude) of the two beat bins corresponding to the peaks with the highest magnitude divided by the magnitude of the lower.
Strength sec. strong. rhythmic pulse	Magnitude of the beat bin of the peak with the second highest magnitude
String Ensemble Fraction	Fraction of Note Ons belonging to orchestral string ensemble patches (General MIDI patches 49 to 52)
Strongest rhythmic pulse	Bin label of the beat bin with the highest magnitude
Tempo	Tempo in beats per minute
Timbral width	The width of the peak of the specific loudness spectrum is called the timbral width
Time Prevalence Marimba	The total time in seconds during which marimba was sounding notes divided by the total length in seconds of the piece
Tonal dissonance (H&K)	Tonal dissonance differs in that it only takes into account components of the spectrum that relate to tone. Tonalness is defined as “the degree to which a sound has the sensory properties of a single complex tone such as a speech vowel. As intonation gets increasingly worse, tonalness decreases.
Tonal dissonance (Sethares)	Tonal dissonance only takes into account components of the spectrum that relate to tone. Tonalness is defined as “the degree to which a sound has the sensory properties of a single complex tone such as a speech vowel. As intonation gets increasingly worse, tonalness decreases.
Variability of note prevalence of pitched instruments	Standard deviation of the fraction of notes played by each General MIDI instrument that is used to play at least one note
Variab. prevalence unpitched instruments	Standard deviation of the fraction of notes played by each MIDI Percussion Key Map instrument that is used to play at least one note
Variability of Note Duration	Standard deviation of note durations in seconds
Variability of Number of Independent Voices	Standard deviation of number of different channels in which notes have sounded simultaneously. Rests are not included in this calculation
Variability of time between attacks	Standard deviation of the times, in seconds, between Note On events (irregardless of channel)
Variation of Dynamics	Standard deviation of loudness levels of all notes
Variation of Dynamics of Each Voice	The average of the standard deviations of loudness levels within each channel that contains at least one note
Volume	Volume is a subjective measure of sound on a scale extending from small to large. Large volume is associated with low frequency, high intensity, and broad bandwidth

A.3. Affective Science

Terms	Definition
Happiness	Affective state characterized by feelings of enjoyment, pleasure, and satisfaction
Sadness	Affective state characterized by feelings of gloominess
Anger	Affective state characterized by a psychophysiological response to pain, perceived suffering or distress
Fear	Affective state characterized by a response to impending danger, that is tied to anxiety
Tension	Affective state characterized by physiological or mental stress
Relaxation	Affective state characterized by the absence of muscular tension and a non-active mind

A.4. Acronyms

Acronym	Definition
ADSR	Attack, Decay, Sustain, Release
BPM	Beats Per Minute
BVP	Blood Volume Pulse
CBR	Case-Based Reasoning
CC	Correlation Coefficient
EMG	Electromyography
GSR	Galvanic Skin Response
LPC	Linear Predictive Coding
MAE	Mean Absolute Error
MFCC	Mel Frequency Cepstral Coefficient
MIDI	Musical Instrument Digital Interface
RMS	Root Mean Square
RMSE	Root Mean Square Error
TPC	Tonal Pitch Class

Bibliography

Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., Kastner, T., Cremer, M., 2001. Content-based identification of audio material using mpeg-7 low level description. In: International Symposium on Music Information Retrieval (ISMIR).
URL http://ismir2001.ismir.net/pdf/allamanche.pdf

Amatriain, X., Bonada, J., Loscos, À., Arcos, J., Verfaille, V., 2003. Content-based transformations. Journal of New Music Research 32 (1), 95–114.

Arcos, J., de Mantaras, R., 2000. Combining ai techniques to perform expressive music by imitation. In: AAAI Workshop: Artificial Intelligence and Music. pp. 41–47.
URL http://citeseer.ist.psu.edu/cache/papers/cs/14277/http:zSzzSzwww.iiia.csic.eszSz~arcoszSzFuzzySaxex.pdf/combining-ai-techniques-to.pdf

Baraldi, F., 2003. An experiment on the communication of expressivity in piano improvisation and a study toward an interdisciplinary research framework of ethnomusicology and cognitive psychology of music. Tech. rep., Paris V University.
URL http://recherche.ircam.fr/equipes/repmus/MemoiresATIAM0203/Bonini.pdf

Barrington, L., Lyons, M. J., Diegmann, D., Abe, S., 2006. Ambient display using musical effects. In: International conference on Intelligent User Interfaces (IUI). Vol. 11. ACM Press, New York, NY, USA, pp. 372–374.
URL http://van.ucsd.edu/pubs/Barrington-AffectiveEffects-IUI2006.pdfhttp://van.ucsd.edu/pubs/AffectiveEffects.ppt

Bartneck, C., 2001. How convincing is mr. data's smile: Affective expressions of machines. User Modeling and User-Adapted Interaction 11 (4), 279–295.
URL http://www.bartneck.de/publications/2001/howConvincingIsMrDatasSmile/bartneckUMUAI2001.pdfhttp://citeseer.ist.psu.edu/cache/papers/cs/20773/http:zSzzSzwww.bartneck.dezSzworkzSzaem.pdf/bartneck00affective.pdf

Baum, D., 2006. Emomusic–classifying music according to emotion. In: Workshop on Data Analysis. Citeseer.
URL http://www.ifs.tuwien.ac.at/mir/pub/baum_wsom06.pdf

Berg, J., Wingstedt, J., 2005. Relations between selected musical parameters and expressed emotions - extending the potential of computer entertainment. In: International Conference on Advances in Computer Entertainment. p. 8.
URL http://sis.cms.livjm.ac.uk/library/AAA-GAMES-Conferences/ACM-ACE/ACE2005/FP3-8%20(a105).pdf

Birchfield, D., 2003. Generative model for the creation of musical emotion, meaning, and form. In: ACM SIGMM Workshop On Experiential Telepresence. ACM Press New York, NY, USA, pp. 99–104.
URL http://ame2.asu.edu/faculty/dab/research/publications/ETP03_Birchfield.pdf

Bod, R., 2002. Memory-based models of melodic analysis: Challenging the gestalt principles. Journal of New Music Research 31 (1), 27–36.

Bradley, M., Lang, P., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry 25 (1), 49–59.

Bradley, M., Lang, P., 1999. International affective digitized sounds (iads): Stimuli. Instruction Manual and Affective Ratings.

Bradley, M., Lang, P., 2000. Affective reactions to acoustic stimuli. Psychophysiology 37 (02), 204–215.
URL http://www.stanford.edu/~kateri/Becky/PDFs/Bradley%202000.pdf

Bresin, R., Friberg, A., 2000. Emotional coloring of computer-controlled music performances. Computer Music Journal 24 (4), 44–63.
URL http://www.speech.kth.se/prod/publications/files/724.pdf

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces. ACM, pp. 205–211.

Cabrera, D., 1999. Psysound: A computer program for psychoacoustical analysis. In: Australian Acoustical Society Conference. Vol. 24. pp. 47–54.
URL http://members.tripod.com/~densil/software/PsySound.PDF

Cambouropoulos, E., 1997. Music, Gestalt, and Computing-Studies in Cognitive and Systematic Musicology. Ch. Musical Rhythm: A Formal Model for Determining Local Boundaries, Accents and Metre in a Melodic Surface, pp. 277–293.

Cambouropoulos, E., 1998. Towards a general computational theory of musical structure. Ph.D. thesis, University of Edinburgh.

Carvalho, V., Chao, C., 2005. Sentiment retrieval in popular music based on sequential learning. In: SIGIR: Conference on Research and Development in Information Retrieval. Vol. 28.
URL http://www.andrew.cmu.edu/user/cchao/projects/carvalho_chao_sigir05.pdf

Casella, P., Paiva, A., 2001. Magenta: An architecture for real time automatic composition of background music. In: International Workshop on Intelligent Virtual Agents (IVA '01). Springer-Verlag, London, UK, pp. 224–232.
URL http://gaips.inesc-id.pt/gaips/shared/docs/Casella01Magenta.pdfhttp://liquidnarrative.csc.ncsu.edu:16080/classes/csc582/Presentations/waronke-magenta-presentation.pdf

Chattah, J., 2006. Semiotics, pragmatics, and metaphor in film music analysis. Ph.D. thesis, The Florida State University College Of Music.
URL http://etd.lib.fsu.edu/theses/available/etd-04042006-140957/unrestricted/JuanChattah_Dissertation.pdf

Chung, J., Vercoe, G., 2006. The affective remixer: Personalized music arranging. In: Conference on Human Factors in Computing Systems. ACM Press New York, NY, USA, pp. 393–398.
URL http://media.icu.ac.kr/park/id/readings/AffectRemix-Final.pdfhttp://media.icu.ac.kr/park/id/presentations/Remixer.ppthttp://courses.media.mit.edu/2005spring/mas630/05.projects/AffectListener/afflisten.ppt

Cliff, D., 2000. Hang the dj: Automatic sequencing and seamless mixing of dance-music tracks. Tech. rep., Hewlett-Packard Laboratories.

Collier, W., Hubbard, T., 2001. Musical scales and evaluations of happiness and awkwardness: Effects of pitch, direction, and scale mode. The American Journal of Psychology 114 (3), 355–375.
URL http://www.psy.tcu.edu/ColHub_AJP01.pdf

Corthaut, N., Govaerts, S., Duval, E., 2006. Moody tunes: The rockanango project. In: International Symposium on Music Information Retrieval (ISMIR). Vol. 7.
URL http://ismir2006.ismir.net/PAPERS/ISMIR0688_Paper.pdf

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. Feeltrace: An instrument for recording perceived emotion in real time. In: ISCA Workshop on Speech and Emotion. pp. 19–24.
URL http://www2.dfki.de/~schroed/articles/cowieetal2000.pdf

Dalla Bella, S., Peretz, I., Rousseau, L., Gosselin, N., 2001. A developmental study of the affective value of tempo and mode in music. Cognition 80, B1–B10.
URL http://www.brams.umontreal.ca/plab/downloads/Dalla_Bella_et_al_2001.pdf

Daly, E., Lancee, W., Polivy, J., 1983. A conical model for the taxonomy of emotional experience. Journal of Personality and Social Psychology 45 (2), 443–457.

Damásio, A., Sutherland, S., 1996. Descartes' error: Emotion, Reason and the Human Brain. Papermac London.

Desain, P., Honing, H., 2003. The formation of rhythmic categories and metric priming. Perception 32 (3), 341–365.
URL http://www.numerik.mathematik.uni-mainz.de/~schneid/optimaleMusik/Desain/mmm-28.pdf

Desain, P., Honing, H., et al., 1999. Computational models of beat induction: The rule-based approach. Journal of New Music Research 28 (1), 29–42.

Deutsch, D., 1982. The Psychology of Music. Academic Press.

DInca, G., Mion, L., June 2006. Expressive audio synthesis: From performances to sounds. In: International Conference on Auditory Display (ICAD). Vol. 12. University of London, UK.

URL http://www.dcs.qmul.ac.uk/research/imc/icad2006/proceedings/papers/f43.pdf

Dixon, S., 1997. Beat induction and rhythm recognition. Advanced Topics in Artificial Intelligence, 311–320.

Dornbush, S., Fisher, K., McKay, K., Prikhodko, A., Segall, Z., 2005. Xpod a human activity and emotion aware mobile music player. In: Proceedings of the International Conference on Mobile Technology, Applications and Systems. Citeseer, pp. 1–6.

Eerola, T., 2003. The dynamics of musical expectancy: Cross-cultural and statistical approaches to melodic expectations. Ph.D. thesis, University of Jyvaskila.
URL http://www.cc.jyu.fi/~ptee/publications/phd1.pdf

Eerola, T., Toiviainen, P., 2004. Mir in matlab: The midi toolbox. In: International Symposium on Music Information Retrieval (ISMIR).
URL http://www.cc.jyu.fi/~ptee/publications/3_2004.pdf

Ekman, P., 1999. Handbook of Cognition and Emotion. Sussex: John Wiley & Sons Ltd, Ch. Basic Emotions, pp. 45–60.

Ekman, P., Rosenberg, E., 2005. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA.

Eladhari, M., Nieuwdorp, R., Fridenfalk, M., 2006. The soundtrack of your mind: Mind music-adaptive audio for game characters. In: ACM SIGCHI international conference on Advances in computer entertainment technology. ACM Press New York, NY, USA.
URL http://student.dei.uc.pt/~apsimoes/PhD/References/soundtrackMind.pdf

Ellis, W., 1999. A source book of Gestalt psychology. Routledge.

Eronen, A., Klapuri, A., 2000. Musical instrument recognition using cepstral coefficients and temporal features 2, 753–756.

Etzel, J., 2006. Algorithms and procedures to analyze physiological signals in psychophysiological research. Ph.D. thesis, Iowa State University.
URL http://archives.ece.iastate.edu/archive/00000238/01/dissertation.pdf

Fabiani, M., Friberg, A., 2007. Expressive modifications of musical audio recordings: preliminary results. In: Proceedings of the 2007 International Computer Music Conference (ICMC 2007), Copenhagen (DK). Vol. 2. pp. 21–24.

Farbood, M., 2006. A quantitative, parametric model of musical tension. Ph.D. thesis, Massachusetts Institute of Technology.
URL http://web.media.mit.edu/~mary/thesis/files/MaryFarbood-PhD-Thesis-2006.pdf

Feng, Y., Zhuang, Y., Pan, Y., 2003. Popular music retrieval by detecting mood. In: SIGIR: Research and development in informaion retrieval. Vol. 26. ACM Press New York, NY, USA, pp. 375–376.

Field, A., 2009. Discovering statistics using SPSS. SAGE publications Ltd.

Francisco, V., Hervas, R., 2007. Emotag: Automated mark up of affective information in texts. In: Proceedings of the Doctoral Consortium in EUROLAN 2007 Summer School. pp. 5–12.

Friberg, A., October 2004. A fuzzy analyzer of emotional expression in music performance and body motion. Music and Music Science.
URL http://www.speech.kth.se/prod/publications/files/1346.pdf

Friberg, A., 2006. pdm: an expressive sequencer with real-time control of the kth music-performance rules. Computer Music Journal 30 (1), 37–48.

Friberg, A., Bresin, R., Sundberg, J., 2006. Overview of the kth rule system for musical performance. Advances in Cognitive Psychology 2 (2-3), 145–161.
URL http://www.speech.kth.se/prod/publications/files/1330.pdf

Friberg, A., Schoonderwaldt, E., Juslin, P., Bresin, R., 2002. Automatic real-time extraction of musical expression. In: International Computer Music Conference. pp. 365–367.
URL http://www.speech.kth.se/prod/publications/files/875.pdf

Friesen, W., Ekman, P., 1983. Emfacs-7: emotional facial action coding system, unpublished manuscript, University of California at San Francisco.

Frijda, N., 2000. Handbook of Emotions. New York: The Guilford Press, Ch. The Psychologists' Point of View, pp. 59–74.

Funk, M., Kuwabara, K., Lyons, M., 2005. Sonification of facial actions for musical expression. In: New Interfaces for Musical Expression (NIME). National University of Singapore Singapore, Singapore, pp. 127–131.
URL http://www.kasrl.org/lyons_nime2005_127.pdf

Gabrielsson, A., Lindstrom, E., 2001. Music and emotion: Theory and research. Ch. The Influence Of Musical Structure On Emotional Expression, pp. 223–248.

Gagnon, L., Peretz, I., 2003. Mode and tempo relative contributions to" happy-sad" judgements in equitone melodies. Cognition & Emotion 17 (1), 25–40.
URL http://www.brams.umontreal.ca/plab/downloads/CE_Gagnon.pdf

Gaye, L., Mazé, R., Holmquist, L., 2003. Sonic city: The urban environment as a musical interface. In: New Interfaces For Musical Expression (NIME). National University of Singapore Singapore, Singapore, pp. 109–115.
URL http://tii.se/reform/results/publications_2003/2003_nime.pdf

Goga, M., Goga, N., 2003. Aesthetic analyze of computer music. In: Generative Art Conference. Vol. 6.
URL http://www.generativeart.com/on/cic/papersga2003/a18.htm

Gómez, E., Peterschmitt, G., Amatriain, X., Herrera, P., 2003. Content-based melodic transformations of audio material for a music processing application. In: Proc. Int. Conf. Digital Audio Effects. Citeseer, pp. 333–338.

Grachten, M., July 2006. Expressivity-aware tempo transformations of music performances using case based reasoning. Ph.D. thesis, Universitat Pompeu Fabra.

Grieco, A., Oliveira, A. M., 2012. Physiological and behavioural reactions to acoustic stimuli. Tech. rep., Faculdade de Psicologia e Ciências da Educação da Universidade de Coimbra. URL http://dl.dropbox.com/u/64916421/Report2012.pdf

Grilo, C., 2002. Aplicação de algoritmos evolucionários à extracção de padrões musicais. Master's thesis.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182.

Haag, A., Goronzy, S., Schaich, P., Williams, J., 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. Lecture Notes in Computer Science 3068, 36–48. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2.1742&rep=rep1&type=pdf

Hashida, Y., Nakra, T., Katayose, H., Murao, Y., Hirata, K., Suzuki, K., Kitahara, T., 2008. Rencon: Performance rendering contest for automated music systems. In: Proceedings of the 10th Int. Conf. on Music Perception and Cognition (ICMPC 10), Sapporo, Japan. Citeseer, pp. 53–57.

Haxby, J., Parasuraman, R., Lalonde, F., Abboud, H., 1993. Superlab: General-purpose macintosh software for human experimental psychology and psychological testing. Behavior Research Methods 25 (3), 400–405.

Healey, J., Picard, R., Dabek, F., November 1998. A new affect-perceiving interface and its application to personalized music selection. In: Workshop Perceptual User Interfaces. URL http://www.acm.org/icmi/1998/Papers/Healey.pdf

Ilie, G., Thompson, W., 2006. A comparison of acoustic cues in music and speech for three dimensions of affect. Music Perception 23, 319–329. URL http://www.ccit.utoronto.ca/billt/papers/MUSIC.23_319-330.pdf

Janssen, J., van den Broek, E., Westerink, J., 2009. Personalized affective music player. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, pp. 1–6.

Jehan, T., September 2005. Creating music by listening. Ph.D. thesis, Massachusetts Institute of Technology, MA, USA. URL http://web.media.mit.edu/~tristan/phd/pdf/Tristan_PhD_MIT.pdfhttp://www.media.mit.edu/events/movies/video.php?id=tristan-2005-06-17

Juslin, P., 2001. Communicating emotion in music performance: A review and a theoretical framework. Music and Emotion: Theory and Research, 309–337.

Juslin, P., Laukka, P., 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. Journal of New Music Research 33 (3), 217–238.

Justus, T., Bharucha, J., 2002. Music Perception and Cognition in Stevens Handbook of Experimental Psychology, Volume 1: Sensation and Perception. Vol. 1. New York: Wiley, pp. 453–492.

URL http://ase.tufts.edu/psychology/music-cognition/pdfs/Justus_Bharucha_2002.pdf

Ka-Hing, J., Sze-Tsan, C., Kwok-Fung, C., Chi-Fai, H., 2006. Emotion-driven automatic music arrangement. In: International Conference on Computer Graphics and Interactive Techniques. ACM Press New York, NY, USA.
URL http://student.dei.uc.pt/~apsimoes/PhD/References/p108-ka-hing.pdf

Khalfa, S., Peretz, I., Blondin, J., Manon, R., 2002. Event-related skin conductance responses to musical emotions in humans. Neuroscience Letters 328 (2), 145–149.
URL http://skhalfa.com/doc/Khalfa_NeurLett.pdf

Kim, S., André, E., 2004. Composing affective music with a generate and sense approach. In: Proceedings of Flairs 2004 - Special Track on AI and Music. AAAI Press.
URL http://mm-werkstatt.informatik.uni-augsburg.de/files/publications/94/FLAIRS04KimS.pdf

Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D., 2010. Music emotion recognition: A state of the art review. In: International Symposium of Music Information Retrieval.

Kimura, J., 2002. Analysis of emotions in musical expression.
URL http://courses.media.mit.edu/2002spring/mas630/02.projects/kimura/

Klein, M., 2003. Psychophysiological and emotional dynamic responses to music: An exploration of a two-dimensional model. In: National Conferences on Undergraduate Research (NCUR).
URL http://www.bethelks.edu/academics/undergrad_research/files/13/Mike__NCUR2003_paper.pdf

Korhonen, M., 2004. Modeling continuous emotional appraisals of music using system identification. Master's thesis, University of Waterloo.
URL http://www.eng.uwaterloo.ca/~dclausi/Theses/MarkKorhonenMASc2004.pdf

Krumhansl, C., 2002. Music: A link between cognition and emotion. Current Directions in Psychological Science 11 (2), 45–50.
URL http://www.erin.utoronto.ca/~w3psyuli/MusicPerception.pdf

Kuo, F., Chiang, M., Shan, M., Lee, S., 2005. Emotion-based music recommendation by association discovery from film music. In: International Conference On Multimedia. Vol. 13. ACM Press New York, NY, USA, pp. 507–510.
URL http://users.cis.fiu.edu/~lli003/Music/am/4.pdfhttp://dblab.cs.nccu.edu.tw/presentation/Emotion-Based%20Music%20Recommendation%20By%20Assciation%20Discovery%20from%20Film.ppt

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159.

Lang, P., 1980. Self-assessment manikin. The Center for Research in Psychophysiology, University of Florida.

Larsen, J., Berntson, G., Poehlmann, K., Ito, T., Cacioppo, J., 2008. The psychophysiology of emotions. Handbook of Emotions, 180–195.

Larson, S., 2004. Musical forces and melodic expectations: Comparing computer models and experimental results. Music Perception 21 (4), 457–498.
URL http://www.informatics.indiana.edu/donbyrd/Teach/PapersEtcByOthers/Larson_MusicalForcesComparing.pdf

Lartillot, O., 2005. Efficient extraction of closed motivic patterns in multi-dimensional symbolic representations of music. In: International Conference on Web Intelligence. IEEE Computer Society Press.
URL http://ismir2005.ismir.net/proceedings/1082.pdf

Lartillot, O., Toiviainen, P., 2007. Mir in matlab (ii): A toolbox for musical feature extraction from audio. In: International Conference on Music Information Retrieval. pp. 237–244.

Lazarus, R., 1999. Handbook of Cognition and Emotion. Sussex: John Wiley & Sons Ltd, Ch. The Cognition-Emotion Debate: A Bit of History, pp. 3–19.

Legaspi, R., Hashimoto, Y., Moriyama, K., Kurihara, S., Numao, M., 2007. Music compositional intelligence with an affective flavor. In: International conference on Intelligent user interfaces. Vol. 12. ACM Press New York, NY, USA, pp. 216–224.
URL http://student.dei.uc.pt/~apsimoes/PhD/References/p216-legaspi.pdf

Leman, M., Lesaffre, M., Tanghe, K., 2001. Introduction to the ipem toolbox for perception-based music analysis. Mikropolyphonie-The Online Contemporary Music Journal 7.
URL http://www.ipem.ugent.be/Toolbox/IT_PaperMeeting.pdf

Leman, M., Vermeulen, V., De Voogdt, L., Taelman, J., Moelants, D., Lesaffre, M., 2003. Correlation of gestural musical audio cues and perceived expressive qualities. Gesture-based communication in human-computer interaction. Berlin, Heidelberg, Springer-Verlag, 40–54.
URL https://archive.ugent.be/retrieve/1424/2004LemanCorrelation.pdf

Lerdahl, F., Jackendoff, R., 1983. A Generative Theory Of Tonal Music. MIT Press Cambridge, Mass.
URL http://www.amazon.com/gp/reader/026262107X/ref=sib_dp_pt/102-0736134-1221739#reader-link

Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H., Martens, J., August 2003. User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. In: Stockholm Music Acoustics Conference. pp. 6–9.
URL http://www.ipem.ugent.be/staff/marc/Papers2003/2003LesaffreEtAl-Taxonomy-SMAC.pdf

Li, T., Ogihara, M., 2003. Detecting emotion in music. In: International Conference on Music Information Retrieval (ISMIR). Vol. 4. pp. 239–240.
URL http://ismir2003.ismir.net/papers/Li.PDF

Lisetti, C., Nasoz, F., 2004. Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP Journal on Applied Signal Processing 2004 (11), 1672–1687.
URL http://www.cs.cmu.edu/~cga/behavior/Lisetti.pdf

Liu, C., Yang, Y., Wu, P., Chen, H., 2006. Detecting and classifying emotion in popular music. In: Joint International Conference on Information Sciences. Vol. 9. pp. 996–999.
URL http://homepage.ntu.edu.tw/~d95942025/pub/JCIS06.pdf

Liu, D., Lu, L., Zhang, H., 2003. Automatic mood detection from acoustic music data. In: International Symposium on Music Information Retrieval (ISMIR). Vol. 4. pp. 81–7.
URL http://citeseer.ist.psu.edu/cache/papers/cs/30423/http:zSzzSzismir2003.ismir.netzSzpaperszSzLiu.PDF/liu03automatic.pdf

Liu, H., Singh, P., 2004. Conceptnet: A practical commonsense reasoning tool-kit. BT Technology Journal 22 (4), 211–226.

Livingstone, S., 2008. Changing musical emotion through score and performance with a computational rule system. Ph.D. thesis, The University of Queensland.
URL http://www.itee.uq.edu.au/~srl/CMERS/2008_LivingstoneSR_Changing_Musical_Emotion.pdf

Livingstone, S., Brown, A., 2005a. Dynamic response: real-time adaptation for music emotion. In: Australasian Conference On Interactive Entertainment. Vol. 2. Sydney, Australia, Creativity & Cognition Studios Press, pp. 105–111.
URL http://www.itee.uq.edu.au/~srl/IE_2005.pdf

Livingstone, S., Brown, A., 2005b. Influencing the perceived emotions of music with intent. In: International Conference on Generative Systems. Vol. 3.
URL http://www.itee.uq.edu.au/~srl/3rd_Iteration.pdf

Livingstone, S. R., Muhlberger, R., Brown, A. R., 2005. Playing with affect: Music performance with awareness of score and audience. In: Australasian Computer Music Conference.
URL http://www.itee.uq.edu.au/~srl/ACMC_05.pdf

Livingstone, S. R., Muhlberger, R., Brown, A. R., 2006. Influencing perceived musical emotions: The importance of performative and structural aspects in a rule system. In: Music as Human Communication: An HCSNet Workshop on the Science of Music Perception, Performance and Cognition. Vol. 1.
URL http://www.itee.uq.edu.au/%7Esrl/HCS_Abstract.pdfhttp://www.hcsnet.edu.au/files2/arch/music06slides/Steven_Livingstone.pdf?PHPSESSID=d89350b93f4bf17ef191b4c6766bdc1e

Livingstone, S. R., Muhlberger, R., Brown, A. R., Loch, A., 2007. Controlling musical emotionality: An affective computational architecture for influencing musical emotion. Digital Creativity 18.
URL http://www.itee.uq.edu.au/~srl/Controlling_Musical_Emotionality.pdf

Lopez, A., Oliveira, A., Cardoso, A., 2010. Real-time emotion-driven music engine. In: International Conference on Computational Creativity.

Lucassen, T., 2006. Emotions of musical instruments. In: Twente Student Conference on IT. Vol. 4.
URL http://referaat.ewi.utwente.nl/documents/2006_04_C-Intelligent_Interaction/2006_04_C_Lucassen,T-Emotions_of_Musical_Instruments.pdf

Margulis, E., 2005. A model of melodic expectation. Music Perception 22 (4), 663–713.
URL http://esf.ccarh.org/254/254_LiteraturePack1/MelFeatures2_ExpectTheoryt(Margulis).pdf

Martin, K., Scheirer, E., Vercoe, B., 1998. Music content analysis through models of audition. In: CM Multimedia Workshop on Content Processing of Music for Multimedia Applications. Vol. 12.
URL http://www.cs.princeton.edu/courses/archive/spring99/cs598b/scheirer.pdf

McCaig, G., Fels, S., 2002. Playing on heart-strings: experiences with the 2hearts system. In: NIME '02: Proceedings of the 2002 conference on New interfaces for musical expression. National University of Singapore, Singapore, Singapore, pp. 1–6.
URL http://student.dei.uc.pt/~apsimoes/PhD/References/p1-mccaig.pdf

McEnnis, D., McKay, C., Fujinaga, I., Depalle, P., 2005. Jaudio: A feature extraction library. In: International Symposium on Music Information Retrieval (ISMIR).
URL http://ismir2005.ismir.net/proceedings/2103.pdf

McKay, C., 2004. Automatic genre classification of midi recordings. Ph.D. thesis, McGill University.
URL http://www.music.mcgill.ca/~cmckay/papers/musictech/MA_Thesis.pdf

McKay, C., Fujinaga, I., 2006. jsymbolic: A feature extractor for midi files. In: International Computer Music Conference (ICMC).
URL http://www.music.mcgill.ca/~cmckay/papers/musictech/McKay_ICMC_06_jSymbolic.pdf

McKinney, M., Breebaart, J., 2003. Features for audio and music classification. In: International Symposium on Music Information Retrieval (ISMIR). Vol. 4.

Mehrabian, A., 1980. Basic dimensions for a general psychological theory. Cambridge, MA: Oelgeschlager, Gunn & Hain.

Meyer, L., 1956. Emotion and Meaning in Music. University of Chicago Press.
URL http://www.amazon.com/gp/reader/0226521397#reader-link

Meyers, O., 2007. A mood-based music classification and exploration system. Ph.D. thesis, Massachusetts Institute of Technology.

Moncrieff, S., Dorai, C., Venkatesh, S., 2001. Affect computing in film through sound energy dynamics. In: MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia. ACM Press, New York, NY, USA, pp. 525–527.
URL http://www.computing.edu.au/~svetha/cma-current/simon/acm2001.pdf

Monteith, K., Martinez, T., Ventura, D., 2010. Automatic generation of music for inducing emotive response. In: Proceedings of the International Conference on Computational Creativity. pp. 140–149.

Monteith, K., Martinez, T., Ventura, D., 2012. Automatic generation of melodic accompaniments for lyrics. In: International Conference on Computational Creativity. p. 87.

Moog, R., 1986. Midi: Musical instrument digital interface. Audio Engineering Society, 394–404.

Mosst, M., December 2006. Quantitative modeling of emotion perception in music. Master's thesis, University Of Southern California.

Muyuan, W., Naiyao, Z., Hancheng, Z., 2004. User-adaptive music emotion recognition. In: 7th International Conference on Signal Processing. Vol. 2. pp. 1352–1355.
URL http://www.ews.uiuc.edu/~mwang2/files/ICSP04.pdf

Nakra, T., 1999. Inside the conductors jacket: Analysis, interpretation and musical synthesis of expressive gesture. Ph.D. thesis, Massachusetts Institute of Technology.
URL http://vismod.media.mit.edu/pub/tech-reports/TR-518.pdf

Narmour, E., 1990. The Analysis and Cognition of Basic Melodic Structures: The Implication-realization Model. University of Chicago Press.

Numao, M., Kobayashi, M., Sakaniwa, K., 1997. Acquisition of human feelings in music arrangements. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 268–273.
URL http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/11032/http:zSzzSznumao-www.cs.titech.ac.jpzSzlabzSzpaperszSzNumao97b.pdf/numao97acquisition.pdf

Numao, M., Takagi, S., Nakamura, K., 2002. Constructive adaptive user interfaces - composing music based on human feelings. In: AAAI.
URL http://www.ai.sanken.osaka-u.ac.jp/files/Numao-caui.pdf

Oliveira, A., Cardoso, A., 2007. Towards affective-psychophysiological foundations for music production. In: Affective Computing and Intelligent Interaction. Vol. 4738. Springer, p. 511.

Oliveira, A., Cardoso, A., 2008a. Affective-driven music production: selection and transformation of music. In: International Conference on Digital Arts - ARTECH.

Oliveira, A., Cardoso, A., 2008b. Emotionally-controlled music synthesis. In: Encontro de Engenharia de Áudio da AES Portugal.

Oliveira, A., Cardoso, A., 2008c. Modeling affective content of music: A knowledge base approach. In: Sound and Music Computing Conference.

Oliveira, A., Cardoso, A., 2008d. Towards bi-dimensional classification of symbolic music by affective content. In: International Computer Music Conference.

Oliveira, A., Cardoso, A., 2009. Automatic manipulation of music to express desired emotions. In: Sound and Music Computing Conference.

Oliveira, A., Cardoso, A., 2010. A musical system for emotional expression. Knowledge-Based Systems 23, 901–913.

Oliver, N., Flores-Mangas, F., 2006. Mptrain: a mobile, music and physiology-based personal trainer. In: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services. ACM, pp. 21–28.

Ortony, A., Collins, A., 1988. The Cognitive Structure of Emotions. Cambridge University Press.

Ortony, A., Turner, T., 1990. What's basic about basic emotions? Psychology Review 97 (3), 315–31.

URL http://www.cs.northwestern.edu/~ortony/Andrew%20Ortony_files/Basic_Emotions.pdf

Pachet, F., Roy, P., Cazaly, D., 2000. A combinatorial approach to content-based music selection. Multimedia, IEEE 7 (1), 44–51.

Padova, A., Bianchini, L., Lupone, M., Belardinelli, M., September 2003. Influence of specific spectral variations of musical timbre on emotions in the listeners. In: Triennial ESCOM Conference. Vol. 5.
URL http://www.epos.uos.de/music/books/k/klww003/pdfs/164_Padova_Proc.pdf

Padova, A., Santoboni, R., Belardinelli, M., March 2005. Influence of timbre on emotions and recognition memory for music. In: Conference on Interdisciplinary Musicology.
URL http://www.oicm.umontreal.ca/doc/cim05/articles/PADOVA_A_CIM05.pdf

Paulus, J., Klapuri, A., 2006. Music structure analysis by finding repeated parts. In: First ACM Workshop on Audio and music computing multimedia. ACM, pp. 59–68.

Picard, R., 1997. Affective Computing. MIT Press Cambridge, MA, USA.

Plack, C., 2004. Auditory perception in Psychology: An International Perspective (PIP). Psychology Press.
URL http://www.psypress.co.uk/pip/resources/chapters/PIP_Auditory_Perception.pdf

Plutchik, R., 1980. A general psychoevolutionary theory of emotion. Emotion: Theory, research, and experience 1 (3), 3–33.

Pratt, C., 1948. Music as a language of emotions. Bulletin of the American Musicological Society 11 (1), 67–68.

Rentfrow, P., Gosling, S., 2003. The do re mis of everyday life: Examining the structure and personality correlates of music preferences. Journal of Personality and Social Psychology 84, 1236–56.
URL http://homepage.psy.utexas.edu/homepage/faculty/Gosling/reprints/jpsp03musicdimensions.pdf

Ritossa, D., Rickard, N., 2004. The relative utility of pleasantness and liking dimensions in predicting the emotions expressed by music. Psychology of Music 32 (1), 5–22.
URL http://music.ucsd.edu/~sdubnov/Mu175/Papers/Ritossaetal.pdf

Robertson, J., De Quincey, A., Stapleford, T., Wiggins, G., August 1998. Real-time music generation for a virtual environment. In: Workshop on AI/Alife and Entertainment. Vol. 24. p. 1998.
URL http://liquidnarrative.csc.ncsu.edu/classes/csc582/papers/real-time-music-generation.pdf

Russell, J., 1989. Measures of emotion. Emotion: Theory, research, and experience 4, 83–111.

Scaringella, N., Zoia, G., Mlynek, D., 2006. Automatic genre classification of music content: a survey. IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia.

Scheirer, E., 2000. Music-listening systems. Ph.D. thesis, Massachusetts Institute of Technology.
URL http://www.idiap.ch/~paiement/references/to_read/music/feature_extraction/eds-diss-full/eds-diss-full.pdf

Schellenberg, E., 1997. Simplifying the implication-realization model of melodic expectancy. Music Perception 14 (3), 295–318.

Schellenberg, E., Adachi, M., Purdy, K., McKinnon, M., 2002. Expectancy in melody: Tests of children and adults. Journal of Experimental Psychology: General 131 (4), 511–537.
URL https://www.erin.utoronto.ca/uploads/tx_researcherprofile/JEPGeneral.pdf

Schenker, H., 1973. Harmony. MIT Press.

Scherer, K., 2000. The Neuropsychology Of Emotion. Oxford University Press, Ch. Psychological models of emotion, pp. 137–162.
URL http://www.affective-sciences.org/system/files/2000_Scherer_Borod.pdf

Scherer, K., 2005. What are emotions? and how can they be measured? Social Science Information 44, 695–729.

Schubert, E., 1999. Measurement and time series analysis of emotion in music. Ph.D. thesis, University of New South Wales.
URL http://www.library.unsw.edu.au/~thesis/adt-NUN/uploads/approved/adt-NUN20021104.143221/public/02vol1.pdfhttp://www.library.unsw.edu.au/~thesis/adt-NUN/uploads/approved/adt-NUN20021104.143221/public/03vo2.pdf

Schwarz, D., 2004. Data-driven concatenative sound synthesis. Ph.D. thesis, Universite Paris.
URL http://recherche.ircam.fr/equipes/analyse-synthese/schwarz/thesis/report.pdf

Serra, X., Leman, M., Widmer, G., 2007. A roadmap for sound and music computing. The S2S Consortium.

Sloboda, J., 1991. Music structure and emotional response: Some empirical findings. Psychology of Music 19 (2), 110–120.

Sorensen, A., Brown, A., 2000. Introducing jmusic. In: Australasian Computer Music Conference. pp. 68–76.
URL http://eprints.qut.edu.au/archive/00006805/01/6805.pdf;http://jmusic.ci.qut.edu.au/

Steinbeis, N., Koelsch, S., Sloboda, J., 2006. The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. Journal of Cognitive Neuroscience 18 (8), 1380.
URL http://www.stefan-koelsch.de/papers/Steinbeis+JOCN2006_inpress.pdf

Sugimoto, T., Legaspi, R., Ota, A., Moriyama, K., Kurihara, S., Numao, M., 2008. Modelling affective-based music compositional intelligence with the aid of ans analyses. Knowledge-Based Systems 21 (3), 200–208.

Taylor, R., Boulanger, P., Torres, D., 2005. Visualizing emotion in musical performance using a virtual character. Lecture Notes in Computer Science 3638, 13.
URL http://www.cs.ualberta.ca/~pierreb/Papers-Thesis-2004-2005/SmartGraphics2005_13.pdf

Temperley, D., 2004. The Cognition of Basic Musical Structures. MIT Press.

Temperley, D., Sleator, D., 1999. Modeling meter and harmony: A preference-rule approach. Computer Music Journal 23 (1), 10–27.
URL http://www.cs.cmu.edu/~sleator/papers/music-modeling.pdf

Tenney, J., Polansky, L., 1980. Temporal gestalt perception in music. Journal of Music Theory 24 (2), 205–241.

Tillmann, B., Bharucha, J., Bigand, E., 2000. Implicit learning of tonality: A self-organizing approach. Psychological Review 107 (4), 885–913.
URL http://olfac.univ-lyon1.fr/unite/equipe-02/tillmann/download/Tillmann_etal_2000.pdf

Toiviainen, P., Krumhansl, C., 2003. Measuring and modeling real-time responses to music: The dynamics of tonality induction. Perception 32 (6), 741–766.
URL http://www.cc.jyu.fi/~ptoiviai/pdf/ToivKrumhPercep2003.pdf

Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I., 2008. Multilabel classification of music into emotions. In: International Conference on Music Information Retrieval. Vol. 2008.

Typke, R., Wiering, F., Veltkamp, R., 2004. A survey of music information retrieval systems. Retrieved April 12, 2004.

Tzanetakis, G., Cook, P., 2000a. Audio information retrieval (air) tools. In: International Symposium on Music Information Retrieval (ISMIR). Vol. 1.
URL http://www.ee.columbia.edu/~dpwe/papers/TzanC00-airtools.pdf

Tzanetakis, G., Cook, P., 2000b. Marsyas: a framework for audio analysis. Organised Sound 4 (03), 169–175.

Tzanetakis, G., Cook, P., 2002. Musical genre classification of audio signals. IEEE Transactions On Speech And Audio Processing 10 (5), 293.
URL http://www.ee.columbia.edu/~marios/courses/e6820y02/project/papers/Automatic%20Musical%20Genre%20Classification.pdf

van de Laar, B., 2006. Emotion detection in music, a survey. In: Twente Student Conference on IT. Vol. 4.
URL http://referaat.ewi.utwente.nl/documents/2006_04_C-Intelligent_Interaction/2006_04_C_Laar,B.L.A.van.de.,-Emotion_detection_in_music,_a_survey.pdf

Vassilakis, P., 2005. Auditory roughness as means of musical expression. Selected Reports in Ethnomusicology 12, 119–144.

Vavrille, F., 2006. Musicovery: An interactive webradio.
URL http://www.visualcomplexity.com/vc/project.cfm?id=329

Vayrynen, E., Seppanen, T., Toivanen, J., 2003. An experiment in emotional content classification of spoken finnish using prosodic features. In: Finnish Signal Processing Symposium. pp. 264–267.
URL http://www.mediateam.oulu.fi/publications/pdf/411.pdf

Ventura, F., Oliveira, A., Cardoso, A., 2009. An emotion-driven interactive system. In: 14th Portuguese Conference on Artificial Intelligence. pp. 167–178.

Vesterinen, E., 2001. Affective computing. In: Digital media research seminar.

Vickhoff, B., Malmgren, H., 2004. Why does music move us? Tech. rep., Dept. of Philosophy, Göteborg University, Sweden.
URL http://www.phil.gu.se/posters/musicmove.pdf

Vyzas, E., 1999. Recognition of emotional and cognitive states using physiological data. Ph.D. thesis, Massachusetts Institute Of Technology.
URL http://citeseer.ist.psu.edu/cache/papers/cs/11090/ftp:zSzzSzwhitechapel.media.mit.eduzSzpubzSztech-reportszSzTR-510.pdf/vyzas99recognition.pdf

Wallis, I., Ingalls, T., Campana, E., Goodman, J., 2011. A rule-based generative music system controlled by desired valence and arousal. In: Sound and Music Computing.

Wassermann, K., Eng, K., Verschure, P., Manzolli, J., 2003. Live soundscape composition based on synthetic emotions. IEEE Multimedia 10 (4), 82–90.
URL http://ada.ini.ethz.ch/presskit/papers/Wassermann-Emotions-2003.pdf

Webster, G., Weir, C., 2005. Emotional responses to music: Interactive effects of mode, texture, and tempo. Motivation and Emotion 29 (1), 19–39.
URL http://psych.colorado.edu/~gwebster/Music.pdf

Weisberg, S., 2005. Applied linear regression. Wiley-Blackwell.

Weiss, A., 2000. Music selection for internet radio.

Whitman, B., 2005. Learning the meaning of music. Massachusetts Institute of Technology.
URL https://dspace.mit.edu/bitstream/1721.1/32500/1/61896668.pdf

Widmer, G., Goebl, W., 2004. Computational models of expressive music performance: The state of the art. Journal of New Music Research 33 (3), 203–216.
URL http://www.cp.jku.at/research/papers/Widmer_Journal_of_New_Music_Research.pdf

Wijnalda, G., Pauws, S., Vignoli, F., Stuckenschmidt, H., 2005. A personalized music system for motivation in sport performance. Pervasive Computing, IEEE 4 (3), 26–32.
URL http://www.redant.nl/g.l.wijnalda/files/pervasivecomputing-final.pdf

Wingstedt, J., Liljedahl, M., Lindberg, S., Berg, J., 2005. Remupp: An interactive tool for investigating musical properties and relations. In: New Interfaces For Musical Expression. University of British Columbia, Vancouver, Canada, pp. 232–235.
URL http://www.nime.org/2005/proc/nime2005_232.pdf

Winter, R., 2005. Interactive music: Compositional techniques for communicating different emotional qualities. Master's thesis, University of York.
URL http://www.speech.kth.se/prod/publications/files/1701.pdf

Witten, I., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S., 1999. Weka: Practical machine learning tools and techniques with java implementations. International Conference on Neural Information Processing, 192–196.

Wu, T., Jeng, S., 2006. Extraction of segments of significant emotional expressions in music. In: Workshop on Computer Music and Audio Technology.
URL http://forum.dmc.ntnu.edu.tw/~wocmat2006/pdf/4-3.pdf

Yang, D., Lee, W., 2004. Disambiguating music emotion using software agents. In: International Symposium on Music Information Retrieval (ISMIR). Vol. 5.
URL http://www.site.uottawa.ca/~wslee/publication/ISMIR2004.pdf

Yang, Y., Lin, Y., Su, Y., Chen, H., 2008. A regression approach to music emotion recognition. Audio, Speech, and Language Processing 16 (2), 448–457.
URL http://mpac.ee.ntu.edu.tw/~yihsuan/pub/TASLP08.pdf

Yang, Y., Liu, C., Chen, H., 2006. Music emotion classification: A fuzzy approach. ACM Multimedia, 81–84.
URL http://homepage.ntu.edu.tw/~d95942025/pub/ACMMM06.pdf

Zils, A., Pachet, F., 2001. Musical mosaicing. In: Digital Audio Effects (DAFx).
URL http://www.csl.sony.fr/downloads/papers/2001/zils-dafx2001.pdf