Faculdade de Engenharia da Universidade do Porto

Automatic Recognition of Emotion for Music Recommendation

António Miguel Antunes de Oliveira

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
Major Telecomunicações

Orientador: Rui Penha (PhD.)
Co-orientador: Marcelo Caetano (PhD.)

29 July 2016

Resumo

É bastante comum associar música com emoções. O principal objetivo deste projeto é desenvolver um sistema de reconhecimento emocional na música (MER) para fins de recomendação musical. O objetivo do sistema MER é estimar automaticamente as emoções associadas com música, naquilo que é a sua perceção. Normalmente, os sistemas MER mapeiam um modelo de emoções através das características extraídas do áudio. O reconhecimento automático de emoções através do áudio é bastante desafiante porque, existem diversos fatores determinantes como a experiência pessoal e o contexto cultural do ouvinte, fatores estes que não estão compreendidos apenas nos sons musicais. Atualmente existem desafios associados com a maioria dos componentes que fazem parte dos sistemas MER, nomeadamente, com a seleção das características musicais, com o modelo de emoções aplicado, com os métodos de anotação e com as técnicas de aprendizagem automática utilizadas. Este projeto investiga a aplicação de diferentes técnicas de aprendizagem automática para associar automaticamente as características musicais calculadas a partir do áudio às anotações de emoções feitas por humanos. Uma comparação com o estado da arte dos sistemas MER é apresentada e discutida. Dois cenários principais foram testados. O primeiro, a respeito do algoritmo de aprendizagem automática, para avaliar os desempenhos dos algoritmos utilizando as mesmas características. O segundo para avaliação do impacto das diferentes características sobre o desempenho do sistema. O algoritmo *Support Vector Regression* apresentou resultados com maior precisão e robustez. A dinâmica musical, o ritmo, o timbre e a altura revelaram ter um impacto maior nos resultados. O mapeamento entre características e o modelo de emoções utilizando um modelo de aprendizagem automática pode ser usado para estimar as emoções associadas com músicas a que o sistema não tenha sido previamente exposto. Por último, o sistema desenvolvido tem o potencial de recomendar música para os ouvintes com base no seu conteúdo emocional.

Abstract

Music is widely associated with emotions. The main goal of this project is to develop a music emotion recognition system (MER) for music recommendation. The aim of MER is to automatically estimate the perceived emotions associated with music. Typically, MER maps from features extracted from audio to a model of emotions. The automatic recognition of emotions from audio is very challenging because important factors such as personal experience and cultural background are not captured by the musical sounds. Currently, there are challenges associated with most steps of music emotion recognition systems, namely feature selection, the model of emotions, annotation methods, and machine learning techniques used. This project investigates the use of different machine learning techniques to automatically associate musical features calculated from audio to annotations of emotions made by human listeners. A comparison with state of the art MER systems is presented and discussed. Two main scenarios are tested. The first one, regarding the machine learning algorithm, evaluate the different performances using the same features. The second one, evaluate the impact of categories of features on the performance of the system. The Support Vector Regression algorithm presented more accurate results and robustness. Dynamics, rhythm, timbre and pitch revealed to be the ones that had the more impact on the results. The map between the feature space and the model of emotions learned by the machine learning model can be used to estimate the emotions associated with music that the system has not been previously exposed to. Consequently, the system has the potential to recommend music to listeners based on emotional content.

Acknowledments

I am extremely grateful to all that participated in this thesis:

- Firstly, to Professor Rui Penha for creating all the conditions to the development of this innovative project;

- My supervisor at INESC Professor Marcelo Caetano which, in addition to all the guidance, I also want to show my gratitude for all the shared knowledge and determination;

- To all the members of the Sound and Music Computing Group at INESC for the constant accompaniment;

- To all my colleagues from the BEST Porto, who never cease to believe in me thank you for the friendship and support given to me throughout this year of work at FEUP;

- Wanting to avoid filling this page with names I thank all my friends who accompanied me over all these years: Bárbara, Manuel, Bruna, Pedro, Alberto, Queirós, Ivo and Hugo.

- Cristina Oliveira, my dear sister, and my dearest family, for the unconditional support and understanding shown.

It is a great pleasure to thank everyone who helped me write my dissertation.

Thank you all.

"I am part of a light, and it is the music.
The Light fills my six senses: I see it, hear, feel, smell, touch and think. Thinking of it means my sixth sense. Particles of Light are a written note. A bolt of lightning can be an entire sonata. A thousand balls of lightening is a concert."

Nikola Tesla

Table of Contents

List of Figures

List of Tables

Abbreviations

MER	Music Emotion Recognition
MIR	Music Information Retrieval
MFCCs	Mel-frequency Cepstral Coefficients
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
IDFT	Inverse Discrete Fourier Transform
MLR	Multiple Linear Regression
RMS	Root Mean Square
RMSE	Root Mean Square Error
SVM	Support Vector Machine
SVR	Support Vector Regression
K-NNR	K- Nearest Neighbor Regression
GP	Gaussian Processes
GPR	Gaussian Processes Regression
ASR	Average Silent Rate

Chapter 1

Introduction

Music is one of the fundamental expressions of human culture. It is impossible to find a civilization or a community in our world that does not express themselves by music. Some say, music came even before mankind learned how to speak. Human reaction to the musical discourse is rarely indifference. It is a universal fact that the musical experience translates an emotional experience that can be personal or shared socially in many moments of everyday life.

Recently, mankind as used technological advances to demystify the creation and reorganization of music. More and more, the proliferation of digital music and the amount of content that digital devices are able to store lead to the necessity of innovative ways of musical retrieval and organization.

1.1 - Context

In the last decade, the reorganization of musical information has become a growing area of research. This is not only due to marked development of technology and digital information, but also because there is a great ease of access to a wide range of digital musical content. More and more people have access to large databases of music data; platforms like *Spotify* and *YouTube* are examples of this; and this creates a new paradigm with the emergence of an exponential number of platforms, remaining a question to the user: How to benefit from them? How to optimize them?

So, then appeared the music recommendation systems, using tools and software that provide music suggestions to the user. Recommendation systems based on musical similarity already exist, but most of these systems do not provide reliable advice. The market of digital music listeners increases exponentially. Many of these listeners, despite having very large libraries eventually hear, most of the time, the same tracks.

The study on automated emotional recognition in music usually deals with the problem of a classification prospective, often based only on human annotations. These annotations can be made by experts such as, musicians, psychologists, and musicologists or by lay users in the field. The problem with this approach is that it limits the user database, and the user can only enjoy music that has been previously classified by such persons. Therefore, there is a

lack of solutions that automate the classification mechanisms in a way that there is no need to have direct or indirect human interference in the system.

1.2 - Goals

The primary objective of this project is to create an emotional recognition system for music recommendation. The developed system should be able to automatically associate a given song to certain emotions. In a real context, the system should be able to work with any audio file, whether this file is a complete piece or simply an excerpt. The main goal is that the system given a new entry, that is not previously noted, is capable of producing the desired output without external intervention.

Another goal is that the system should be as valid as possible, presenting a high precision and being as faithful as possible in the association between the music and the model of emotions that is used.

As a secondary goal, the recommendation system may later provide to users, including composers, musicologists and enthusiasts, a music that conveys certain emotion or even know what emotional state a song conveys with the highest accuracy level.

Therefore, this thesis is intended to provide its users an accessible and consistent solution so that they are able to, with simple steps, enjoy with satisfaction the music that they have at their disposal.

1.3 - Motivation

In the context of music recommendation systems, a Master thesis in Multimedia was developed by the student João Pedro dos Santos Figueiredo, whose approach was to base the music recommendation system on the emotions that are perceived by listeners.

Thus this project motivation is to go one step further and create an automatic recommendation system based on emotions. Currently the existing systems only work with a previously annotated database, this limits the rage of music that can be used on the recommendation system. The process of annotating a piece is quite costly and time-consuming [1], examples like annotations made by experts [2], games that implement annotations made by users (games-with-a-purpose) [3, 4] and online music services like Last.fm [5] that allow users to input free-form tags. In fact, systems like *Moodswings* exemplify that even without any cost associated, they present a lack of reliability and quality mainly due to subjective error and the noise effects created by its users [3].

Therefore, the motivation is to add a module capable of automatically annotate songs (auto-tagging) that have never been present in the system before using solutions presented in the current Music Emotion Retrieval (MER) literature.

1.4 - Dissertation Structure

The structure of this thesis is organized as follows:

- Chapter 2: State of the art

Introduces and describes the relation between music and emotion, and analyses the Music Emotion Recognition systems that already have been developed.

- Chapter 3: Characterization of the problem

Refers the project structure and its stages to accomplish the proposed solution, as well as the software available in the market that will serve to solve it.

- Chapter 4: System design

Describes the tools and components of the developed system. Presents its functionalities and interface.

- Chapter 5: Evaluation

Describes the experiments and its results and performance. Evaluation and validation of the machine learning algorithms.

- Chapter 6: Conclusions

Presents the contributions of this dissertation and the future work in the context of this project.

Chapter 2

State of the art

This review of literature has been divided into three sections. Firstly, we focus on the close relation between music and emotion, giving a context to the work on this thesis. The second section is devoted to the MER, in particular, the representation framework is deliberated that occurs in the vicinity and in what way musical aspects are associated to emotions. Lastly, on the third chapter we discuss the music classification problem, emphasizing the methods that have been utilized for acquiring audio aspects and organize music. The aim is to explain and exemplify the methods found on the literature to implement reliable classifiers, like signal deliberation for audio feature extraction, supervised learning, and machine teaching techniques.

2.1 - Music and Emotion

2.1.1. - Relation between music and emotions

To understand the relation between music and emotion one must understand the importance of music in human society. Analyzing the evolution of mankind, we see that since its beginning music fulfills an essential purpose in numerous social and cultural contexts [6]. Every human culture has developed some form of musical expression [7], so people are always exposed to music when living in society. Music affects us on our daily lives even as a group, it is experienced on events such as weddings or funerals, or even as a symbol of an ideology or a nation [6]. But its importance cannot be fully understood by such experiences since it affect us also in a very personal way, in the way people live. What is evident is that is required psychological explanation for the ability that music has to surprise, delight, energize sooth and in a direct way shape our emotional states.

Researchers in neuroscience and cognition have studied this concept and it supports the idea that music activates certain areas of the human brain. This activation happens in such a way that motivates people to listen to a certain music, only because it communicates a particular emotion [8, 9]. Its ability to profoundly affect human psychology and emotion is one of the reasons why music has such a close relation with humanity [7].

The capability that music has to express and induce emotional states has been a primary subject of scientific investigation. Music expresses emotion when a person relates emotion to

a song, whether it's a piece or an entire song, e.g. this song is happy. On the other hand, music induces emotion when a person feels a certain emotion while or after listening to a song, e.g. the listener felt happiness [10]. Research in this field is usually made by introducing different musical pieces to listeners and storing their emotional reaction. Usually, rather than recording reactions on the listeners own words, it is common to use standard ratings [11]. Based on these concepts, one study demonstrated that harken to a music that is pleasurable to the listener will activate identical brain sections that are stimulated only though euphoric stimuli like food, sex or even drugs [8]. Other authors, like Thompson [6], sustain that music have attributes such as "intensity (loudness), tempo, dissonance, and pitch height", that can interfere with emotional expressions. They are as a code that when used properly serve as an emotion communicator tool. In order to prop this affirmation the author says that "melodies that are played at a slow tempo tend to evoke emotions with low energy such as sadness, whereas melodies that are played at a fast tempo tend to evoke emotions with high energy, such as anger or joy" [6].

This means that music has a biological effect on human brain [8]. Scherer [11] also states that music can produce physiological and behavioral changes, emphasizing the fact that this changes are particularly glaring in motor expressive movements in the face, body, and voice. That is, that in the presence of an emotional music the listener tends to sing along, in a non-reflected action, or to clap or dance, as being an expressive response to the patterning of the physiological emotion.

This apparent evidence is still outward to some authors. Konecni [10], for instance, puts some doubts on this relationship sustaining that in this equation is important to make a distinction on emotion and moods, preferences, attitudes, and personality traits in order to fully determine the existence of a connection between music and human emotions.

Scherer [11], in turn expressed the opinion that emotions can be remarked by some characteristics like the physiological arousal, such as, temperature sensations, respiratory and cardiovascular accelerations and decelerations, trembling and muscle spasms, as well as feelings of constriction in internal organs. According to this author emotions can also be expressed or observed by motor expression since facial expressions or gestures or even the posture one assumes during a emotion revel. Also by speaking a human can express emotions but in terms to the special context of our study al this physiological and cognitive process of detecting emotions is not sufficient to measure them.

Monteith et al [12] advocate that if we aim to design and implement a system that produces human-like behavior, since emotions are a main human characteristic it seems primordial to incorporate that emotional awareness into the system. Furthermore, the use of emotions may produce better results by improving usability, satisfaction and performance.

"Music occurs fundamentally inside the brain" [13]. So, in order to measure an emotional response to a particular piece, we must consider all the mechanisms behind the human processing musical information occurring in the brain. It is in this aspect that personal influences and even memory have impact.

2.1.2. - Induced Emotion or Expressed Emotion

The notion that music is perceived as an expression of emotion is uncommonly deliberated [14]. However, less harmony is observed on the capability that music has to induce certain emotions in listeners. Although the existence of research that corroborate this capability [15], its veracity is still a subject of extensive discussion [11]. According to Evans

et al. [16] and Vieillard et al. [17] the processes of induction and expression of musical emotion have many similar characteristics. In the same line of thought, Zentner et al [18] recommended that stimulation from music may incur under a pattern identical to various other emotional occurrence. Certain aspects of music, like the beat and the rhythm, are recognized to influence physical movements and stimulations. Moreover, a fluctuation in respiration through listening to music, would post an influence on neurophysiological systems in a similar way to other emotion that induce physiological changes [11].

At first, we must clarify the real meaning of emotions. So, what are emotions? Sherer [11], based on other theorists, defends that emotions constitute collaborated fluctuations in motor expressions, physiological stimulations, subjective experiences, behavioral development and cognitive procedures. Other researchers call the first three changes, namely, motor expressions, physiological stimulations and subjective experiences of the emotional triad [13]. Based on this componential approach to emotion, listeners' emotional response only could be reliably measured analyzing their physiological alterations, expressions, postures and gestures along with the feeling that they have recently experienced.

We must as well consider that in our daily lives we use "mood" as a synonym of "emotion". When actually mood is define as momentary situation of feelings or mind [19]. It is normal to mistake them, but we must notice that the definition of mood includes a temporal variance, so the meaning of emotion goes a lot deeper.

The link between music and emotions is normally made by researchers that analyze the listener response to particular pieces or different genres and styles, associating then to the response to certain emotions [13].

The existing debate about the capability of music to induce or express emotion has the repercussion that when people annotate pieces rather than measuring induce emotion, they are asked to measure expressed emotions [11]. Despite this, often it is difficult to differentiate induced and expressed emotion even for professionals in the music field [11].

Music expresses emotions in such a way that listeners are capable of recognize, and perceive a certain emotion without feeling that particular emotion [20]. On the other hand, is music capable of inducing an emotion? The impact of music on people has been used in wide different areas, entertainment, education, music recommendation systems, and many others [13].In what purpose? If music didn't have the capability of influence people, would it have such a main role in areas like advertising and even therapy?

The relation between music and emotion is still under study, and there is yet to appear an unanimously model [13]. Various people are subjected to experience emotions especially when turned to same music. All emotions are subjective by nature, even in the instance that individuals listening are harmonious on the felt emotions, there is ambiguity with regard to its description [21]. Besides, multiple personal factors need to be considered as sociocultural context, personal history and psychological factors.

2.1.3. - Subjectivity

There are substantial evidence that emotional reacting to music in regards to stimulation of multiple components occurs, for example, psychophysiology, subjective emotions, brain stimulation, emotional expressions, regulation of emotions, tendencies, reflexes, in exchange, stimulates various psychological mechanisms such as evaluative conditioning, brain reflexes, visual imagery, musical expectancy, and rhythmic entertainment" [13]. The

information that is adjacent to the music does not have a direct influence on those mechanisms. Specifically it varies according to the listener and the situation in question.

There are some vital aspects to influence individuals' emotional reaction to music, those factors can be determined by the listeners' personal factors and situational influences. Individualistic aspects are inclusive of gender, age, personality, musical preferences and training and present ambience [13]. Whereas situational aspects are inclusive of physical elements like visuals, acoustics, location, social factors such as audience type and particular reasoning for the gathering [13]. Besides this, there still exist musical factors related to the musical qualities of certain pieces, such as genre, style, key, tuning, melody, and rhythm, among others.

On the other hand, several researchers [15, 20] have evidences that sustain the existence of an direct relation between music and the emotional response by listeners. Such researches came to the conclusion that listeners have frequently the same perception when it comes to the basic emotions that are suggested by a particular music, such as anger, disgust, joy, sadness and fear [15]. In such a way that cultural and personal influences have no central role has initially thought, when are considered these basic emotions. In fact this evidence corroborates the idea that are qualities and characteristics in music that produce the same emotional experiences in different listeners.

According to Eerola [1] there exist three types of emotion models, namely, "discrete, dimensional and music-specific" [22]. The discrete emotion model states that all emotions are ramifications stemming from various baseline emotions. This model is not normally used in musical contexts since there are basic emotions such as disgust that are not commonly perceived in musical pieces, in musical study it is normally substituted for tenderness [1].

Studies of music in different cultures and contexts suggest that may exist an universal power of music in terms of psychological and emotional effects, transcending the language, cultural and social barriers [23]. Fritz et al [24] found that people that had no contact with Western culture categorized musical pieces in a similar way as Westerns. So, although there is no dispute regarding the emotional impact of music, it is still discussed if its capability is universal. Thus, the similarities and differences between different listeners emotional response is still a subject of profound study.

2.2 - Music Emotion Recognition

A particular area of research that has met substantial development in previous decade is automated emotional assessments. This research area has been termed the Music Emotion Recognition (MER) [2]. The primary objective of MER is developing and implementing frameworks that automatically associate emotional responses to musical pieces [14]. The current MER systems classify emotions into a restrict number of different categories and then by applying machine learning algorithms they train the classifier. The achieved results are then compared to those obtained with human annotations [14, 25, 26]. Researchers are still trying to improve MER systems since the best results rarely are above 65% [13]. In this section, we will analyze the central barriers to the implementation of MER systems and the solutions found in the literature.

2.2.1. - Musical Features

During the last century, the capability of music to communicate emotions has received great attention by researchers. Mainly, the aim is to evaluate the importance of each musical feature on the prediction of the emotions that are transmitted by musical pieces. Since Hevner's work in 1936, musical features influence on emotions have been studied continuously [27]. It is now known that emotions that are expressed by music are more connected to musical features than personal factors or context effects [20]. As the author's explain, "music often features expressive acoustical patterns similar to those that occur in emotional speech". This capability makes music a "easy to manipulate" structure, and, therefore, there are sophisticated techniques in acoustics that enable researchers to standardize a stimulus with regard to certain acoustic features, while leaving others intact [20].

Amongst the primary issues of MER systems is considered which musical attributes influence the emotional response in listeners [11], but there is still a discussion on which features in music have a primary conduct with emotional expressions, and this hardens the association between music an emotions, making it difficult to have a straight and infallible system of MER.

When developing a MER system we must decide which musical features the system will gather. This is important because it allows the system to have enough data so that is capable of making a reliable association to emotions. It is widely accepted that features such as harmony, timbre, rhythm, tempo and even lyrics affect emotion. Also, the comprehension that the emotions expressed by a piece changes is critical, since its beginning until the last note. In fact, the "mood of a piece may change over its duration" creating a problem to MER that is also exposed by the cited author: "there may be considerable disagreement regarding the perception and interpretation of the emotions of a song or ambiguity within the piece itself" [2].

According to Eerola [1] the research to understand the importance of musical features on emotional response shares the same three central areas of cognitive science, namely, empirical observation, formalization and cognitive relevance. The first one is related to the theories elaborated that allow posterior organized observations. The appearance of widely used protocols for music representation such as MIDI, and the advance of computational systems has allowed the proliferation of empirical observations in musical context. Empirical observations have an important role on the features choice to construct a computational model. Furthermore, the test of statistical hypothesis based on methods from psychology have made an important improvement in the accuracy of the conclusions from such empirical observations [1]. The second area represents the veracity of scientific models. Computational models are favored due to their transparency, since they are more replicable and testable [1]. The third area covers the cognitive revolution, in which computational models have a main role, not only because of its consolidated development, over the last decade, but also in concern to the cognitive approach that requires "cognitive plausibility for any of the features or processes involved in the model" [1]. As Eerola holds this features might seem a contrast if we consider computational efficiency and the fact that these figures in first place when it comes to an engineering approach[1].

Research in this field accepts the existing of musical features that are more cognitively relevant in the perception of musical emotion.

The majority of MER systems tend to use a large number of musical features. The signal processing adjacent to the analysis of those features is used to have the basis to implement an optimal emotional classifier based on sophisticated algorithms [26, 28, 29]. Those algorithms will be the key to determine the performance evaluation of the regression method and the evaluation of data space and feature space, in which the best combination of data and future space is found by a summing process and the data transforming does not make significant difference to the prediction accuracy [1].

Music has a wide range of representations, such as audio, the score, lyrics, title, artist and even annotations. The features can be extracted from one or more of those representations [2]. On the literature, we can see that most of the MER systems use audio to extract those musical features [2, 5]. The principle elements that can be achieved from audio, for example, timbre, melody, rhythm and deliberating under a signal processing algorithms.

The most frequently implemented aspects are MFCCs, fundamental frequency, attack time, root mean square (RMS) energy, spectral centroid and spectral roll-off, among others [13, 30, 31]. Despite the prominent usage of MFCCs for deliberation of audio waves, discrepancies are observed commonly in regards to execution and usage of particular coefficients. One of the reason pointed for explaining that fact is that the range of the dimensions of the audio waves varies widely, and many features are correlated [32].

A summary of the main features is provided in in Table 1. The number of music features will define the dimension of the input space, some features like MFCCs have multiple dimensions. In the literature we can see that it is usual to reduce that dimension using principal component analysis (PCA), and that this reduction happens to guarantee the correlation between features [13, 33, 34]. Investigators have been analyzing the various effects of different musical feature selection algorithms in MER systems, sustaining that MER systems usually “join all the features together, which is called a bag of features approach” [33].

Table 2.1 - Musical features commonly used for emotional classification [2].

Type	Features
Dynamics	RMS energy
Timbre	MFCCs, spectral shape, spectral contrast
Harmony	Roughness, harmonic change, key clarity, majorness
Register	Chromagram, chroma centroid and deviation
Rhythm	Rhythm strength, regularity, tempo, beat histogram
Articulation	Event density, attack slope, attack time

All the musical features used in MER systems are “related with three different musical experiences, the rhythmic, the perceptual, and the former levels” [13]. This means that difference of the values of such features in a musical piece must be analyzed in order to make a connection to emotions.

According to Eerola [1] contemporary researches explore this connection of emotions to musical features [6, 15] via experimental manipulation and focus primarily on specific instance-based aspects, such as loudness, tempo, mode, pitch, dynamics, among others. This way is called causal, since the musical features are continuously varied to acknowledge their

effect on emotional responses [1]. Another way of viewing the problem is the correlational way. In this approach musical features are mapped to certain emotions using statistical models [35], this approach does not restrict the musical material that can be used, but at the same time has problems on finding reliable annotations, whether it is made by experts or laypeople [36]. Both causal and correlation studies have contributed to the study of the relation between emotion and musical features, they have restricted the features that have a close relation to emotional responses [1]. In fact, studies have been made to isolate the features that have more influence in the emotional content of a specific piece [37]. These approaches have created the main conditions to develop a MER system, since it is a primordial requirement to have main content that can be measured by computational models.

Furthermore, these conditions have allowed the expansion of the field known as Music Informatics Research (MIR). This field has been evolving and has created the bases for the automatic extraction of musical features [1]. Also, the creation of reliable tools has enabled the improvement of the efficiency when processing audio information [5]. The innovation in this field enabled MER systems to present results that previously were not possible.

There are various approaches to find the most important musical features to emotional responses. Eerola et al. [38] in order to refine the features that have more impact on the emotional responses to music have explore regression techniques such as, Principal Component Regression, Robust Regression, Partial Least Squares Regression. They have created a specific set of musical features dedicated to emotional responses [2].

Schubert et al. [35] used musical features such as “loudness, tempo, melodic contour, texture, and spectral centroid as predictors in linear regression models” [13].

Mion and De Poli elaborated a system for the selection of musical features important to the association with emotions. They selected a set of single-dimensional features, that included intensity and spectral shape [34]. The developed system relies on sequential feature selection (SFS) that was followed by principal component analysis (PCA) in order to extract central feature dimensions. The aim of their analysis was to classify monophonic instrument into nine different emotional categories. They concluded that of the 17 features that were tested the ones that carried more important information were attack time, peak sound level, roughness and notes per second [2].

Another study elaborated by Macdorman et al. [39] analysed the ability of different musical features to predict pleasure and arousal appraisal of musical pieces. They concluded that features such as spectral histogram, periodicity, sonogram, fluctuation pattern, and mel frequency ceptral coefficients (MFCCs) are better predictors of arousal ratings than pleasure. Many authors explore the goal to find the most reliable features and normally the conclusion is that the results are far better when it is used multiple features together [2],

With this goal in mind, Schmidt et al. [40] used MFCCs , chroma, octave-based spectral contrast and statistical spectrum descriptors, namely, centroid, flux and They achieved this purpose testing the features individually and combined together. They concluded that the results were more satisfactory when those feature were MFCCs and spectral contrast, but once again the highest results were achieved using multiple features [40].

Lu et al. [28] have adopted a wider approach to the musical feature extraction, compiling multiple features that result in multi-dimensional spaces and then applying dimension reduction.

Sturm [32] developed a system for music genre classification. He reproduced the system of Chang et al. [41] called SRCPC and that is based on sparse representation classification (SRC). This system implements six short-term features, namely, octave-based spectral contrast (OSC), MFCCs, spectral centroid, spectral roll-off, spectral flux, and zero-crossings and six long-term features, namely, OMSC, low-energy modulation spectral crest and flatness measure [32].

Some authors [13, 42, 43] defend that MER systems should also consider the temporal change of the musical features, how variations in a specific feature are correlated with the expression of emotion rather than only that feature value [13]. Moreover, Eerola et al. [44] listed features that can be directly and automatically estimated from the audio and are divided into three musically relevant temporal scales, this allows the creation of new relevant annotations [1, 13, 44].

Independently of the methods applied on the implementation of a MER system, the systems that produce better performances are the ones that use multiple musical features. This appends mainly due to the fact that the musical experiences related to the rhythmic, the perceptual, and the former levels are connected to different features. Having this in mind, the approach to the reality is more genuine wider the range of features considered in the MER system. All the musical features used in MER systems are "related with three different musical experiences, the rhythmic, the perceptual, and the former levels" [13]. This means that difference of the values of such features in a musical piece must be analyzed in order to make a connection to the emotions, also leading to the idea previously cited that many of the features of music are correlated, and, therefore, thorny to disentangle[32]

According to Eerola [1] contemporary researches explore this connection of emotions to musical features [6, 15] via experimental manipulation and focus primarily on specific instance-based aspects, such as loudness, tempo, mode, pitch, dynamics, among others. This way is called causal, since the musical features are continuously varied to acknowledge their effect on emotional responses [1].

Another way of viewing the problem is the correlational way. In this musical features approach is mapped to certain emotions using statistical models [35], this oncoming does not restrict the musical material that can be used, but at the same time has problems on finding reliable annotations, whether it is made by experts or laypeople [36]. This last aspect seems to be of great importance to Eerola and Vouskoski [1], because, as they stated, "the method for selecting the stimulus material is probably of great importance, especially for studies focusing on felt emotions". According to these investigators, experts, in what come to emotions, an due to the fact that know from experience the music and the study objectives, might incur on the temptation of perfection, giving composed replies in prejudice of true ones [36].

Both causal and correlation studies have contributed to the study of the relation between emotion and musical features, they have restricted the features that have a close relation to emotional responses [1]. In fact, studies have been made to isolate the features that have more influence in the emotional content of a specific piece [37]. These approaches have created the main conditions to develop a MER system, since it is a primordial requirement to have main content that can be measured by computational models.

Furthermore, these conditions have allowed the expansion of the field known as Music Informatics Research (MIR). This field has been evolving and has created the bases for the automatic extraction of musical features [1]. Also, the creation of reliable tools has enabled

the improvement of the efficiency when processing audio information [5]. The innovation in this field enabled MER systems to present results that previously were not possible.

There are various approaches to find the most important musical features to emotional responses. Eerola et al. [38] in order to refine the features that have more impact on the emotional responses to music, have explore regression techniques such as, Principal Component Regression, Robust Regression, Partial Least Squares Regression. They have created a specific set of musical features dedicated to emotional responses in order to, as they stated, "annotate each music piece with a set of emotions" [2].

Schubert et al. [35] used musical features such as "loudness, tempo, melodic contour, texture, and spectral centroid as predictors in linear regression models" [13].

Mion and De Poli [34] elaborated a system for the selection of musical features important to the association with emotions. They selected a set of single-dimensional features that included intensity and spectral shape by extracting audio features from expressive performances that were recorded and by asking the musicians to perform in order to convey different expressive intentions. These authors, by this way, demonstrate that higher recognition ratings are achieved by using a set of four features which can be specifically related to qualitative descriptions of the sound by physical metaphors and that these audio features can be used to retrieve expressive content on audio data, and to design the next generation of search engines for music information retrieval [34].

The developed system relies on sequential feature selection (SFS) that was followed by principal component analysis (PCA) in order to extract central feature dimensions. The aim of their analysis was to classify monophonic instrument into nine different emotional categories. They concluded that of the 17 features that were tested the ones that carried more important information were attack time, peak sound level, roughness and notes per second [2].

Another study elaborated by Macdorman et al. [39] analyzed the ability of different musical features to predict pleasure and arousal appraisal of musical pieces. They concluded that features such as spectral histogram, periodicity, sonogram, fluctuation pattern, and mel frequency ceptral coefficients (MFCCs) are better predictors of arousal ratings than pleasure. Many authors explore the goal to find the most reliable features and normally the conclusion is that the results are far better when it is used multiple features in conjunction [2],

With this goal in mind, Schmidt et al. [40] used MFCCs, chroma, octave-based spectral contrast and statistical spectrum descriptors, namely, centroid, flux and They achieved this purpose testing the features individually and combined together. They concluded that the results were more satisfactory when those features were MFCCs and spectral contrast, but once again the highest results were achieved using multiple features. "Combining MLR in multiple stages produces results comparable to more computationally complex methods" [40].

Lu et al. [28] have adopted a wider approach to the musical feature extraction, compiling multiple features that result in multi-dimensional spaces and then applying dimension reduction.

Sturm [32], conversely, developed a system for music genre classification. He reproduced the system of Chang et al. [41] called SRCPC and that is based on sparse representation classification (SRC). This system implements six short-term features, namely, octave-based spectral contrast (OSC), MFCCs, spectral centroid, spectral roll-off, spectral flux, and zero-

crossings and six long-term features, namely, OMSC, low-energy modulation spectral crest and flatness measure [32].

Some authors [13, 42, 43] defend that MER systems should also consider the temporal change of the musical features, how variations in a specific feature are correlated with the expression of emotion rather than only that feature value. Caetano (2014) sustains this position by affirming that "music is intrinsically temporal, and time is experienced through memory. Studies suggest that the temporal evolution of the musical features is intrinsically linked to listeners' emotional response to music" [13].

Moreover, Eerola et al. [44] listed features that can be directly and automatically estimated from the audio and are divided into three musically relevant temporal scales, this allows the creation of new relevant annotations [1, 13, 44].

Independently of the methods applied on the implementation of MER systems, the systems that produce better performances are the ones that use multiple musical features. This appends mainly due to the fact that the musical experiences related to the rhythmic, the perceptual, and the former levels are connected to different features. Having this in mind, the approach to the reality is more genuine wider the range of features considered in the MER system.

2.2.2. - Model of Emotions

When developing a MER system we must decide how to measure and represent the emotions. The adapted solution must express listeners' emotional reaction in a reliable way. As we foremost stated, emotions are subjective and change according to personal and situational factors, so the MER systems have to use the model most similar to that reality.

According to Scherer [11] there are "three major schools of thought: basic emotions, the emotional dimension, and the eclectic approach". The basic emotion and the emotional dimension are linked with the concept of discrete emotion. This concept defends the assessment of a trivial, established volume of rudimentary emotions namely, anger, fear, joy, sadness and disgust [11]. The dimensional theory defends the measurement of a valence dimension, positive or negative feeling, and an activation dimension, aroused or calm feeling. The eclectic approach states that the emotional labels should be selected according to the aims of a specific study [11] . Exemplifying, the emotional labels of a music study may have labels such as "bright", "thrilling", "contemplative", or "sad". Scherer advocates the use of the eclectic approach since it would fit on the study that it is being developed.

MER systems when it comes to the classification of an emotion either use categorical or parametric models. Categorical models involve the search and organization of a set of emotional labels. This organization of different emotion into categories is manly based on their relevance to some analyzed music. Parametric models suggest that emotion can be measured in a continuum way, i.e. multi-dimensional metrics.

Through the literature reviewed, we can say that there are two main models, which are most frequently used in MER systems. The categorical model proposed by Hevner [27], consisting of a model that separates emotions based on adjective categories, and Russell's model [45], considered a parametric and circumflex model, which acknowledges the existence of two dimensions on emotion.

Hevner's initial study [27] listed emotions in 66 adjectives, which were divided into 8 main categories as seen in **Figure 2.1**. Many studies conducted since [9, 46], have indicated

that this sort of labeling is consistent and very intuitive, even appropriated for listeners without musical training.

Hevner's work has been the base of many studies conducted ever since. Zent et al. [18] produced a list of 801 emotional adjectives into a minor metric of 146 more specific terms for emotional association with music. Their study showed that the interpretation of these adjectives varies between different music genres. Recently the Music Information Retrieval (MIR) Evaluation exchange (MIREX[1]) community divided the adjectives into 5 critical emotional categories, which form the emotional model applied for the validation and evaluation of the different mood classifiers presented [47].

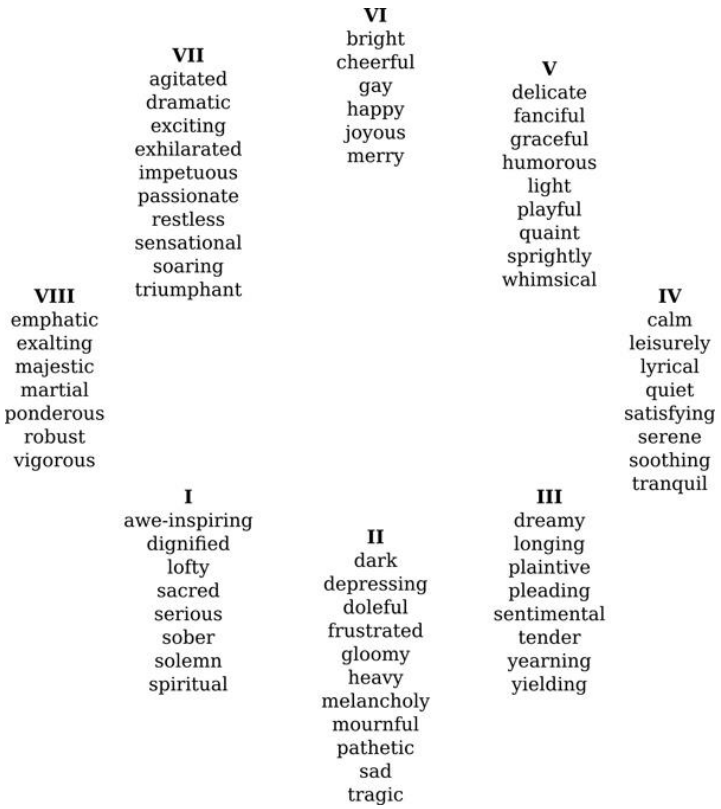


Figure 2.1 - Hevner's classification model [27].

Yang et al. (2011, s/p) state that categorical representation of emotions have the problem of having a restricted number of categories that do not reflect the wide range of emotions that may be perceived by listeners. "Typical categorical approaches that simply assigning one emotion class to each song in a deterministic manner does not perform well in practice" [48].

In this model the listener can only express the emotional states that are present in the inventory, and when that emotional state is not present it is integrated on the emotional category that is most identical. The solution does not reside on increasing the number of categories, mainly because the language is subjective [15].

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

Scherer et al. [11] also defend that while the fixed format of the categorical model can have advantages when analyzing a reduced number of emotions, it is fallible when analyzing emotions related to music. This happens mainly due to the subjectivity of music, and since it is important to ensure that the measurement is valid, restrict the number of emotional reactions may seem not like a reliable solution. The explanation for this intricacy is simple: "this diversity in the ways in which musical expressiveness or affective reactions to music are measured makes it very difficult to compare findings from studies using a different theoretical approach and different conceptualizations and measures of the affective phenomena under investigation." [11]

Thus, many researchers have adopted Russell's model [2, 48], also known as the circumplex model of affect (CMA). The benefit of adopting this model is the possibility to measure a specific emotion, locating it in a two-dimensional space as seen in **Figure 2.2**. This model decomposes emotion in music according to two dimensions, valence and arousal. Valence ranges from positive to negative semantic meaning (e.g happy vs. sad) and arousal varies between high and low energy (e.g aroused vs. sleepy). MER systems based on this model are trained to compute the valence and arousal values and associate each piece with a specific point in the emotional space [48]. One of the problems with the adoption of this model is the asymmetry of the representation since emotions that are quite despair in terms of semantic value can be near from each other [13]. This model has a close relation with

musical features, and divides the emotional space into four quadrants associated with four categories of main emotions.

Figure 2.2 - Circumplex Model of Affect, purposed by Russel [45].

For instance, music that has high pitch, greater variability, faster tempo, heavy sound level, high energy and fast tone attacks correspond mainly to anger that is located on high arousal and negative valence has seen in **Figure 2.2** [1]. Medium sound level, fast tempo, high pitch level, and fast tone attacks, major mode and bright timbre awakes the joy emotion that is located on high arousal and positive valence [1]. Slow tempo and attacks, low pitch level, minor mode and dark timbre corresponds to sadness that has low arousal and negative valence the opposite of joy [1]. Lastly, calmness is located on negative arousal and positive valence, it is similar to sadness but uses major key [1].

Studies have also expanded the dimensional approach to develop a three-dimensional spatial emotional metrics, thought the third dimension's semantics are still debated [49]. Many studies on the dimensional model suggest another dimension rather than valence [50]. Other authors suggest that positive and negative feeling should be treated as separated categories. The Positive and Negative Affect Schedule (PANAS[2]) tool, defines that all discrete emotions exist as an incidences of positive or negative effect, in an equally form as valence [51].

It is important to state that studies that use different models of emotion, consequently do not apply the same measurements. This fact difficult the comparison between different studies because "eclectic lists of emotions generated by researchers to suit the needs of a particular study may lack validity and reliability and render a comparison of research results difficult [11].

So in order to validate results of a MER system one must compare the results to systems that apply the same model of emotions.

The categorical and the dimensional model have been implemented in different MER systems, whether for the classification of music into different emotional categories [52] or to predict emotional dimensions using regression models [38]. Eerola et al sustains that, for instance regression is a technique that "is less influenced by collinearity due to the inclusion of constant variance parameter" [38].

Human musical annotations are made by asking listeners to choose a category or a space where a particular piece belongs. The performance of a MER system may be affected by the choice of that emotional response by listeners. It is known that factors like musical preference affect this choice. However, one way to decrease this subjectivity is to define correctly the applied model of emotion. Therefore, the chosen model of emotion must take into account the particular aims of the project in order to optimize the MER system developed.

2.2.3. - Mapping the emotion

In order to perform, currently MER systems assume that music is simply an audio signal and do not consider the adjacent musical experience [13]. Machine learning techniques perform a mapping from musical features to emotional states. Therefore it is assumed that music causes a particular emotional response. These systems ignore the personal and situational influences that are known to condition human emotional reactions [13, 53-55].

[2] https://www.statisticssolutions.com/positive-and-negative-affect-schedule-panas/

Emotion can be represented as a "multi-dimensional vector or a time-series of vectors over a semantic space of emotions" [2]. This refers to understand that its dimensions yield singular emotion or a pair of negative-positive emotions. Dimension value assesses the semantic correlation amidst a particular emotion and music excerpt. This can be represented in a binary way, acknowledging the existence of a particular emotion, and it can be also represented as a real value score of a Likert scale or a probability estimated value [2].

There are various methods for estimating values of the multi-dimensional vector. One of the forms is to ask human listeners to annotate particular pieces normally according to a set of emotions. Such can be achieved, for instance, using surveys, social tagging mechanisms, games that implement annotations made by users (games-with-purpose) [3]. Another way to obtain emotional classification information for music is through web data, for example, analyzing with text-mining web-documents, large collections of social tags and musical lyrics. Another method is to analyze the audio content, for example, using signal processing and automatically annotate emotional reactions to music using supervised machine learning techniques. Another method is based on multimedia content such as, music videos, photographs, and movies. Furthermore, there are still a multiple data sources, like lyrics and audio, that may be combined to determine the emotional content of music [2]. Currently the development of technologies such as search engines and computational algorithms had a main role on the development of these methods, but they are still matter of wide investigation. Hence, there exists many ways to determine the emotional content of music. This allows different approaches when it comes to have a truthful basis to develop a specific system.

In 2008, Turnbull et al. [56] gathered the CAL500 dataset of annotated music mainly due to the fact that many annotated databases are rarely shared, such as All Music Guide database. The CAL500 is a collection of 500 songs from different artists. Each song has been annotated by a minimum of three laypeople, using a vocabulary of 174 tags, of which 18 are related to different emotions. Other publicly available database was created by Trohidis et al. [57] gathers 593 songs each of them annotated by 3 expert listeners using 6 basic emotions.

Social tagging is another approach to collect human emotion annotations. Last.fm2 is a known example of this practice, the website collects social tags that can be made by user easily on their interface. Last.fm differs from All Music Guide mainly because they make their data publicly available. Social tagging may be considered very useful for the music information retrieval (MIR) community, but it has many flaws like malicious tagging and sparsity due to the cold-start problem and popularity bias, ad-hoc labeling techniques and multiple spellings of tags [2, 58].

One of the ways to contour the annotation problem is to develop collaborative online games for the collection of music data. There are several examples of "Games With A Purpose", such as *Moodswings* [3], *TagATune* [59], *MajorMiner* [60], *Listen Game* [61] and *Hear it* [62]. This proves that has been an effort to create the most reliable dataset in this field of research. This happens mainly due to its importance on the results of a system of this nature.

Since the intention is to develop an automatic recommendation system, we will focus on the machine learning methods and their results on MER systems.

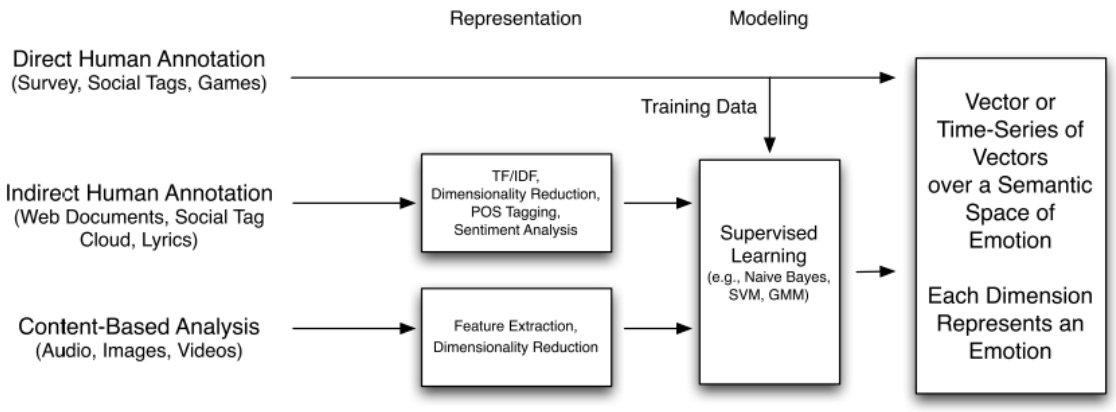


Figure 2.3 - Overall model of emotion classification systems [2].

Automatic annotation of music requires advanced computational techniques. MER systems currently depend on human annotations and they use mainly supervised learning to train statistical models thus map musical features into the different regions of a previously defined emotion space as seen in **Figure 2.3** [14].

The MER systems may work using classification [28] or regression [26], this depends on the model of emotions that is used. If a system uses a categorical model it would work using classification. On the other hand, if it uses a parametric model it would need a regression technique. The MER system is trained to perform with a database that has been previously annotated by humans. After this phase, the framework can anticipate emotional reaction to a certain music that did not exist in the previous annotated database. The performance of the system is evaluated by the existing differences between the emotional association made by the MER system and the human annotation made to that specific piece [13].

MER systems apply machine learning algorithms like support vector machines, which are classifiers for training and testing data and are used to classify music mood categories (SVM) [2, 33].

The machine learning techniques assume that the music features used by the system are reliable predictors of musical emotions. The association from the feature to the emotional space is a one-to-one relationship, since the system assumes that it is captured all the psychological meaning and subsequent emotional reaction of the listener [13].

There are algorithms that tend to be more suitable to the specify aim of mapping emotions and their applicability is still subject of discussion. There are various authors that have experimented different techniques in order to optimize MER systems.

Support Vector Machines (SVM) – Classification

The research carried out by Li et al. [63] classified music into thirteen mood categories using musical features associated to rhythm, pitch and timbre to teach Support Vector Machines (SVMs). They achieved an accuracy of approximately 45% using a database of 499 pieces, each were 30 seconds long. By the time it was considered to be a huge development on the field and increased the interest on this approach.

Years later, Mandel et al. [64] developed a scheme for music recommendation using active learning techniques, this system can yield suggestions centered on musical context provided by users. The user would insert a number of songs that represented the desired

playlist. The created system would then use this data and verification information from the user to construct a binary SVM classifier using MFCCs features. The system was then tested using 72 distinct moods from All Music Guide labels, the results achieved had a peak of 45.2% [2].

Many of the MER systems were firstly implemented to the purpose of genre classification and then adapted to the purpose of emotion classification [47]. In 2007, the MIREX included the task of audio music mood classification. The submitted systems at that time classified the pieces into 5 mood categories. The system that presented the best performance, achieved an accuracy of 61,5% using an annotated database of 600 songs, of 30 seconds each [65]. This system was developed by Tzanetakis using only MFCCs and spectral shape, centroid and roll-off features within an SVM classifier [66].

Cao and Li provided a system that performed above expectations on various groups in 2009, inclusive of mood classification (65.7%) [67]. This framework utilizes a "super vector" of low-level acoustic features, and utilizes a Gaussian Super Vector trailed by Support Vector Machine (GSV-SVM). It is viable to observe that top performers in each multi-year assessment were frameworks based for performing on MIREX objectives.

Speck et al. [25] used outlier detection, teaching a supervised machine learning system. Their system used the one-class SVM implementation from the SVM-KM toolbox[3]. The performance of their system in emotion prediction had an average of 20.52% accuracy.

Support Vector Regression

Another authors such as Schmidt et al. and Han et al. also implemented SVMs algorithms for classification [40, 68]. However, facing unexpected results, all below 50%, they changed their investigation to a regression approach. Their emotional model was quite different, Schmidt et al. [40] used a valence-arousal space focusing on the four main quadrants, Han et al. [68] divided the emotions into 11 critical categories. The system that Han implemented mapped 95% of the emotions into the annotated original categories. It is important to state that they used Support Vector Regression along with Gaussian Mixture Model (GMMs) algorithms and having such a short number of possibilities enabled the system to present a high performance.

Yang et al. applied the Support Vector Regression (SVR) [69] for mapping high-dimensional acoustic features to the bi-dimensional space [70], targeting the prediction of one Valence-Arousal label for each of the 195 music clips. In their work, to extract features, they used *Marsyas* [71], a publicly available extraction tool, and labeled a total of 114 feature dimensions. Before the regression, they needed to use a Principal Component Analysis (PCA), to reduce the data to a tractable number of dimensions. Their results determine a R2 (coefficient of determination) score of 0.58 for arousal and 0.28 for valence.

[3] http://asi.insa-rouen.fr/enseignants/~arakoto/toolbox/

Gaussian Processes Regression

Markov et al. [72] developed two systems using Gaussian Processes (GP) models. Being one of them for music emotion recognition. In this project, it was used both SVM and GP models in order to compare their performances. Their results showed that GP performed better than SVM. In fact, GP achieved a total 11% increase of the coefficient of determination, regarding the emotion recognition.

Viewing the problem of emotion recognition emotions as a classification problem from the perspective of regression, Chen et al [73] divided the emotional space in different areas, associated with 9 different emotions. Their proposed system is based on a deep Gaussian Process. The experimental results demonstrate that the proposed system performs well in emotion recognition, having a better performance than SVM and standard GP.

Gaussian Mixture Model (GMM) - Classification

The investigation developed by Lu et al. [28] used a Gaussian Mixture Model (GMMs) classifier for mood detection and tracking that was based on musical features such as timbre, intensity and rhythm. The musical pieces were then classified using Thayer's model [50], dividing them into the four main quadrants in the valence-arouse dimensions. The system was trained using a set of 250 classical musical pieces, with 20 seconds each, and achieved an accuracy of 85% when trained on 75% of the database. This high performance is mainly due to the division of the emotional space into four quadrants.

Peeters work [47] introduced a larger set of musical features such as, MFCCs and chroma features. Using a GMM approach to classification, the system achieved a performance of 63,7%. It is important to notice that for the selection of the best features that could be used for the association between emotion and music, he employed IRMFSP. Then he reduced the dimension of the data performing Discriminant Analysis (LDA). He employed a categorical model of emotion more wide than most of the systems used until that date.

Neural Networks – Classification

Monteith et al [37] applied neural networks techniques to the automatic generation of music for inducing emotional responses on the musical features previously extracted by the "Phrase Analysis" component of the open software provided by *jMusic*[4]. The resulting system presented average results of 54%. The authors defend the incorporation of more specific musical features into the neural network evaluators, since it would increase the capability of selecting melodies according to the target emotional response.

Partial Least Squares – Regression

Eerola et al. [38] assessed numerous regression methods inclusive of Partial Least-Squares (PLS) regression, a method that evaluates the regression amidst label dimensions. Utilizing a three-dimensional emotion model, R2 regression productivity of 0.72 was acquired for valence, 0.85 for activity, and 0.79 for tension. PLS utilization also concluded peak R2

[4] http://explodingart.com/jmusic/applications.html

prediction rates for 5 basic emotion classes (sad, happy, angry, tender and scary) fluctuating in 0.58 to 0.74.

Multiple Linear Regression

Schmidt et al. [40] noted that dividing the Valence-Arousal space into 4 quadrants was inconsistent and did not took advantage of the parametric model. He applied regression using multiple algorithms, namely, SVR and Multiple Linear Regression (MLR). Assessing the span between anticipated values and actual coordinates on the valence-arousal space, they achieved a performance with only 13,7% of average error.

Bynary Classifiers – Classification

Skowronek et al. established binary classifiers for all twelve non-exclusive mood groups utilizing data from 1059 track excerpts. Utilizing elements dependent on tempo and rhythm, temporal, chroma and key information, and instances of percussive music instances they established quadratic discriminant features for all individual moods, with accuracy levels fluctuating from 77% (carefree-playful) to 91% (calming-soothing), based on each individual group [29].

K-Nearest Neighbor – Classification

Regarding classification problems, this model has been used frequently. In this approach for each input, a search is conducted to find the instance on the training data with the minimum distance to the actual value. Euclidean distance is usually used as the measure [26].

N-Nearest neighbor as been widely applied on genre classification problems. Tsanetakis and Cook [74] propose a set of feature for musical genre classification using the K-Nearest Neighbor. Deshpande et al. [75] also used Nearest Neighbors to classify the music into different genres.

Chapter 3

Characterization of the Problem

3.1 - Definition to the problem

Determining automatically the emotional reaction to a certain musical piece requires knowledge about many distinct fields, such as psychology, musicology, and mainly engineering knowledge about signal processing and machine learning algorithms. The recognition of emotion remains a challenging problem, due to its complexity that presupposes the interaction between different modules. Furthermore the problem expands in different levels, such as, feature selection, the model of emotions, annotation methods and machine learning techniques used to map the emotions.

If the aim is to recognize automatically emotion in music it is primordial to define a set of main components that will define its scheme, as seen in **Figure 3.1**.

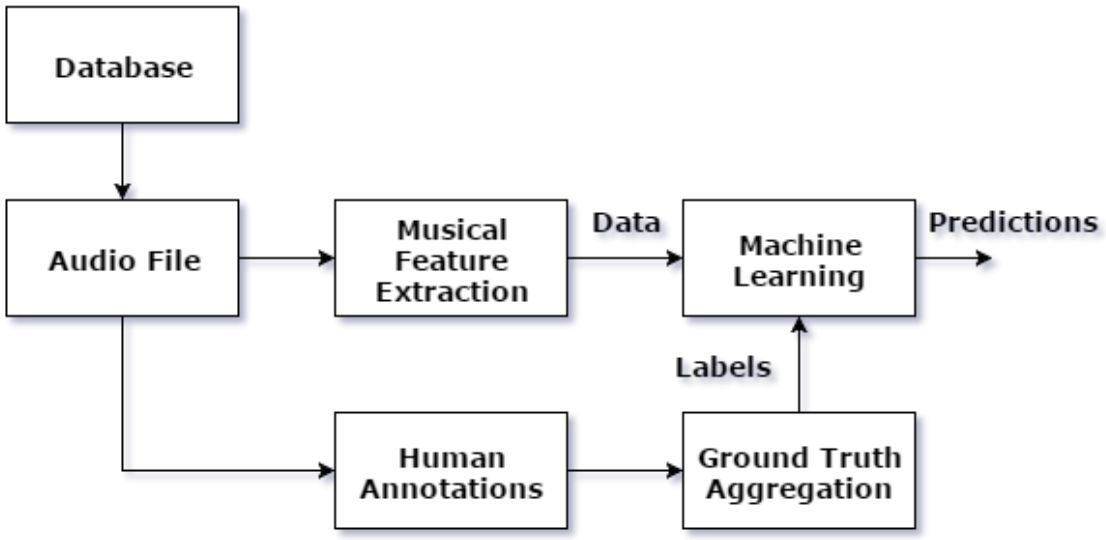


Figure 3.1 - Overview of the system to be implemented.

When developing a MER system there are four main fields that must be taken into account: 1) Database; 2) Model of Emotion; 3) Musical Feature Extraction; 4) Machine Learning Algorithm.

Firstly, when defining the database one most assume that it is a reliable source of information, so it could create the ground truth opposed to the predictions provided by the machine learning algorithms. Secondly the model of emotions must be selected according to the database information.

Regardless of the subjectivity of emotions and the discussion that remains in the field, there are difficulties for machines to solve. Considering the automatic part of the system, we have to process the audio signal and implement an extraction method that performs on the most reliable way. It is necessary to extract the data from the dataset as well. This extraction must be completed having in mind what was viewed on the literature and that if these steps do not occur correctly the automation of the system is compromised. Then the problem is mapping the emotion from those extracted features using a machine learning algorithm. The desire is to make the system "learn" the general rules that map features into emotion, so it is primordial to train it. It is expected to face problems such as the bias-variance tradeoff, problems related to the complexity, the heterogeneity and amount of data, the dimension of the input space and the noise expected in the output values. The system will have to be designed to overcome these challenges.

Thus one must consider this challenges when developing a MER system and consider as well that all of the system is interconnected and every part of it must be implemented according to the previous components. The main problem of MER systems is their accuracy, due mainly to the difficulty of mapping the features into measurable values of emotion.

3.2 - Solution to the problem

This is a complex problem and it is certainly unrealistic to expect to develop a MER that will perform correctly all of the times. However, what is expectable is the creation of a system that will predict the emotions with the minimum possible error. Thus, analyzing the different challenges of this problem, one must consider a solution to every main question and expect to produce the best outcomes.

3.2.1. - Database

The first requirement to implement the intended music emotion recognition system is to identify an audio ground-truth database that relates audio content with emotional annotations. Then we must define the database that would fit with the adopted emotional model, so we considered a list of databases available to the public. We include the most relevant ones on **Table 3.1**, listing projects where the databases were applied. A set of free databases[5] is available online.

It is primordial to choose the most reliable annotated database. Through the literature reviewed we were presented with an amount of possibilities that fit this requirement. Then, we must analyze each and select the ones that have more consistent annotations and that present heterogeneous music in the same way as the databases presented in the market.

One of the problems in MER systems has been the choice of musical pieces used as basis to this systems purpose. This choice varies in the quality and the number of musical examples. There is a tendency to use Western music, recognizable to the people responsible

[5] http://www.audiocontentanalysis.org/data-sets/

for annotating the pieces. However, most of these musical pieces are chosen arbitrarily, and the authors try to include a wide range of emotions into the databases.

Table 3.1 – Database most commonly used on MIR systems.

Authors	Year	Title	Database
Kim et al. [3]	2008	"MoodSwings: A Collaborative Game for Music Mood Label Collection"	Large database labelled manually by a team of experts
Speck et al. [25]	2011	"A contrasting research of collaborative vs. traditional musical mood annotation"	Database obtained through MTurk available for MIR community (Music Information Retrieval)[6]
Pesek et al. [76]	2014	"Introducing a Dataset of Emotional and Color Responses to Music"	200 annotated songs made by 952 users
Eerola et al. [22]	2012	" A contrast of the distinct and dimensional emotional aspect under music "	110 annotated songs annotated by 116 laypeople
Saari et al. [5]	2015	"Genre-Adaptive Semantic Computing and Audio-based Modelling for Music Mood Annotation"	Last.fm[7] - dataset of the online radio with 960 thousands annotations
Figueiredo et al. [77]	2015	"Music Recommendation System Based on Emotions"	744 annotated songs from Free Music Archive (FMA)[8]
Sturm et al. [32]	2013	"On Music Genre Classification Via Compressive Sampling"	GTZAN[9]
Aucouturier et al. [78]	2002	"Finding songs that sound the same"	Database with 17,075 containing information like genre, title and artist.

[6] http://mturk.com
[7] http://www.last.fm
[8] http://freemusicarchive.org
[9] http://marsyasweb.appspot.com/download/data_sets/

Kim et al. [3] used a wide music information database with 179 different mood labels from All Music Database[10]. The advantages of this database are that it has multiple genres and a vast collection, which is constantly updated.

The Pandora[11] service has a large database that has been labeled manually by a team of experts, including musicians. The main disadvantage of this service is that it is only available in the United States, Australia and New Zealand or on request.

It is important to notice that in the market there exists also tools like Moody plug-in for iTunes that allows user to tag their own collections using a quantized 4x4 valence-arousal matrix, but due to their uncertainty this options will not be considered [3].

Speck et al. [25] based on their previous designed game *MoodSwings* [3], made the dataset used initially on the game available online. This dataset was created based on the valence-arousal space, using a labeling task for the Mechanical Turk website. The *MoodSwings* Turk Dataset[12] consists of 240 pieces, with 15 seconds each that were expanded to thirty seconds in the annotation objective to provide audience further practice. The database has labels on each second of an excerpt, made by a group of people that were paid $0.25 per hit on 11 pieces. If the annotations were unsatisfactory they were discarded, this way they would have the most reliable annotation on the valence-arousal space.

This database is available in a MATLAB structure and contains data for each song, such as artist, album, song, song id, the user id, valence value, arousal value, and time (in seconds) at which the value and arousal correspond to. The Moodswings Turk dataset also contains the musical features of each of the 240 songs, namely, MFCCs, octave-based spectral contrast, chronogram, statistical spectrum descriptors (spectral centroid, spectral flux, spectral roll-off and spectral flatness) and EchoNest[13] audio features (extracted using EchoNest Python API, which allow the aggregation of features, such as, timbre, pitches and loudness). This is a consistent database with a large set of musical features that could be used in the further association between emotions and features.

The Moodo Dataset[14] was developed by Pesek et al. [76] and contains users emotional state before taking the annotations, the emotions that were induced by the songs, the expressed musical emotions and also the colors the users associated with a specific piece. Their perception of color in a song is a symbol of the relation to emotion. The dataset contains 200 songs and over 37 different annotations each piece, inclusive of positioning in valence-arousal space. 952 individuals, yielding 6609 emotion/color stimulations, made the annotations of this dataset. According to the authors no present musical emotion dataset has this volume of annotations based on musical pieces [76]. Their conclusion regarding the perception of color and valence-arousal space, despite inconsistences, was that less active emotional states correspond to darker hues (e.g blue, violet), and more active emotional states correspond to lighter hues (e.g red, yellow). To consider this database we should have to discard the colors association data due to its uncertainty and irrelevance to the aims of our work.

[10] http://www.allmusic.com
[11] http://www.pandora.com
[12] http://music.ece.drexel.edu/research/emotion/moodswingsturk
[13] http://developer.echonest.com/client_libraries.html
[14] http://mood.musiclab.si/en/dataset

Another option is the dataset developed by Eerola et al. [22], Soundtracks[15], that gathered film soundtracks. The dataset is constituted by 110 film music excerpts of 15 seconds each. They were annotated by 116 laypeople that listened to each excerpt that corresponds to a film soundtrack. The set is divided in two: one half is an example of five basic emotions; the second half is an example of extremes in the valence-arousal-tension space [22].

Soleymani et al. [79] developed an annotated database of 1000 songs that was selected from Free Music Archive (FMA) called Emotion in Music[16]. The database was then reduced to 744 songs due to the detection of some redundancies. These songs are divided in two main sets: a development set constituted by 619 songs and an evaluation set constituted by 125 songs. Each song is available in full length, and also in 45-second excerpts that were extracted from random starting point in a specific song. The pieces have a sampling frequency of 44100Hz. There is also information about each song, such as, artist, title, duration and genre. The annotations consist of continuous valence-arousal space values for each song. These values were generated based on the averaged and standard deviation annotations with 2Hz sampling rate. The continuous annotations ignore the first 15 seconds of each piece, due to the verified instability of the annotations at the beginning of the song.

Considering all of the above-described databases, the database *Emotion in Music* was used on the music emotion recognition system. This choice is mainly due to the fact that this options is free available to the public, the high reliability of their annotations, the audio file formats that they provide, and lastly because they present heterogeneous music set, constituted of various artists and different genre, such us, Rock, Pop, Reggae, R&B, Electronic, Blues, Country, Folk, Jazz, Classical, among others [76, 79].

3.2.2. - Model of Emotion

Analysing the review articles on the topic [2, 26] it is perceived that the categorical model and parametric model are the ones that are preferably used in recommender systems or musical reorganization.

However, to serve the purpose of this project it was used the parametric model that was firstly proposed by Russell [45]. This is due to its use on the reviewed literature and mainly to the fact that representing emotions according to the emotional represents a greater challenge to the system. Moreover, it allows the exploration of a wide range of emotional states. Besides, this model features the majority of emotional reactions to a certain music. Focusing on aspects of arousal and valence, is the most non-redundant way of representation emotions in music [22].

Having this in mind, the system will now have to extract the musical features from the audio signal of the songs that constitute the applied database.

3.2.3. - Musical Features

In order to design and implement the planned system, one of the main requirements is to identify a framework for acquiring musical elements. This will then allow the system to

[15] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/emotion/soundtracks
[16] http://cvml.unige.ch/databases/emoMusic/

perform a mapping from musical features to emotional states through machine learning techniques.

So that the system that will be created present the most accurate results, the MER system must incorporate the most relevant musical features to the expression of emotions. The extracted features must have a meaning for listeners in such a way that the annotations that are created have a reliable meaning.

Through the literature review, it can be assumed that the attributes that are more commonly used are the MFCCs [25, 37, 77, 78] and the spectral shape and contrast [25] that are associated to the music timbre, harmonic changes [12, 25] e chromagrams [25]. One must have in mind all those features. However the selection of the extraction tool will have a major influence on the features that can be use posteriorly to the mapping of emotions, reducing or increasing the available set.

The remaining task to create the module that extracts the musical features, is to find a capable open-source tool and that fits the analyzed problem. So in order to fulfill this requirement, it was considered the most prominent software available.

In this area *Essentia*[17] is one of the open-source tools available to musical features extraction. *Essentia* is a C++ library designed for audio analysis and MIR purposes, provisioned by Affero GPLv3 license[18]. It is constituted by a comprehensive set of algorithms that can be used in our project. The available algorithms are audio file descriptors input/output, standard signal processing blocks, Filters (FIR & IIR), statistical, spectral, tonal, and rhythm descriptors. It implements the extraction of all the main musical features studied and even implements other functions, such as, SVM classifiers.

Developed by Müller et al. [80] another available option is the *Chroma Toolbox*[19] . This toll is released under the terms of GPL[20]. *Chroma Toolbox* consists of MATLAB executions for acquiring multiple audio; mainly pitch, timbre and chroma features. This tool has been used mainly in tasks such as structure analyses and music synchronization, due to the fact that it has not been experimented in MER systems before and it has a short range of musical feature, its choice seems unlikely.

Marsyas[21] implemented by Tzanetakis et al. [71] is an open-source software for music analysis and synthesis. It was specifically designed for MIR projects. They provide a wide range of components that allow the audio processing. It allows the adaptation of command-lines for the extraction of information from the music and provides a C++ library that has a wide range of audio processing functions. It processes and extracts the main musical features required and its flexibility makes this toolbox appealing.

Released in 2010, *Yaafe*[22] an audio features extraction toolbox developed by Mathieu et al. [81] that allows the user to extract features and its parameters by declaring them in a text file. The user is able to extract the musical features using Python or MATLAB. *Yafee* provides a C++ API to integrate a project; it is capable of reading different audio file formats, automatic sample rate conversion and the flexibility that allows the user to create its own

[17] http://essentia.upf.edu/
[18] http://www.gnu.org/licenses/agpl-3.0.html
[19] http://resources.mpi-inf.mpg.de/MIR/chromatoolbox/
[20] http://www.gnu.org/licenses/gpl.html
[21] http://marsyas.info/
[22] http://yaafe.sourceforge.net/

musical feature library. Among the musical features available are MFCCs, spectral features, energy, and loudness, among others.

Sonic Annotator[23] is a tool for extraction of audio features from multiple audio files and annotation, published under GNU. *Sonic Annotator* is developed entirely in C++ and requires *Vamp plugins*[24] for feature extraction and audio analysis [82]. It is intended for publishing features data about an audio database. Typical features might include tempo, key, and fundamental frequency. The set of available features is not depending on *Sonic Annotator* itself, but on the Vamp plugins. On this project the target plugins are the analysis and extraction extensions. It is possible to use these plugins in Python and it is also possible to adapt them to use in a Java application using *jVamp*. Furthermore, due to the need to implement different components, such as the Vamp plugins, *Sonic Annotator* seems time-consuming and for these reasons was discarded.

Through the analysis of the most used tools, we can say that some stand out from the others, mainly due to their capability to extract multiple musical features. All these tools had in common the fact that they are free available to the public. The choice will have to consider also the machine learning algorithm that will be applied and their compatibility.

3.2.4. - Machine Learning

Focusing on the way the system will be able to accomplish the desired task that is mapping the emotion automatically. The answer lies in a platform capable to use machine learning models in a very specific and reliable way. An important part of the work to be done comes from the approach to the association made between the model of emotions and music.

Having this in mind, one must choose between the different options that are available.

Developed by Lin et al. [83], *LIBSVM*[25] is a software for support vector classification (SVM) and regression (SVR) implemented in C++ and Java. This software supports multi-class classification and is able to estimate distributions. It allows the use of SVM as a tool and has efficient multi-class classification, various SVM formulas, cross validation for model selection and viability estimations.

Another option is to use *LIBLINEAR*[26] that is a linear classifier for data. It has interfaces in MATLAB, Java and Python. It includes same capabilities as *LIBSVM* plus and automatic parameter selection. In our system as the databases may be considered not to be very large it is advisable to use *LIBSVM* due to its fast ability.

3.2.5. - Musical Features and Machine Learning

There are software solutions that provide both musical feature retrieval and machine learning capability.

Bob[27] is a toolbox for signal processing and machine learning developed by the Biometrics group at Idap Research Institute[28], in Switzerland.

---

[23] http://omras2.org/SonicAnnotator
[24] http://www.vamp-plugins.org/
[25] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[26] https://www.csie.ntu.edu.tw/~cjlin/liblinear/
[27] http://idiap.github.io/bob/
[28] http://www.idiap.ch/

Figure 3.2 - Overview of Bob concepts.

Bob provides a Python environment and C++ library for processing data. The library is extensible and can be easily adapted. The *Bob* concept is to provide a trainer that uses data from the dataset to train a specific machine, as seen in **Figure 3.2**. The major Bob's features that could be implemented in our system are signal processing and machine learning. The first one has signal processing that relies on a sequence of mathematical operations. The second one, can perform dimensionality reduction using mainly Principal Component Analysis. Also, *Bob* can perform classification, using methods such as SVM and GMM.

Another option is *Weka*[29] a framework under the GPL. It provides an assortment of teaching sequences for data assessments. The software can be implemented and adapted to a previous developed Java code. *Weka* also contains tools that can be applied to our system such as, audio signal processing, classification and regression [84].

Another widely used toolbox on the area of MIR [2] is *MIRtoolbox*[30] designed by Eerola et al. [85] under the GNU license. It is written in MATLAB and requires Matlab version 7 and Mathworks Signal Processing toolbox. This toolbox offers a large set of functions able to extract musical features from audio files. *MIRtoolbox* includes statistical descriptors and 50 audio and musical features extraction tools. It allows the user the selection and combination of the different features. Its distribution includes other toolboxes that can be useful to this project such as *The Auditory Toolbox*, which includes MFCC computations, and the *Netlab toolbox*, that includes Gaussian Mixture Modeling (GMM) routines used for classification. Considering the aims of our project this option is a useful tool to be considered. This is mainly due to its capability to extract a huge number of musical features and the appearance of other toolboxes that could be implemented together serving the aim of automation to this system.

Lastly, MacKay [86] developed a free software called *jMIR*[31]. This framework has been written in Java and is widely implemented in MIR research, particularly in automatic music classification [86]. GNU license and has the advantage to facilitate the modification of the software according to a specific project aim. What makes *jMIR* a useful tool to our system is that it holds various utilities for retrieval of implementing machine teaching sequences and feature retrieval [86]. It has two flexible components that are capable of pattern recognition and perform tasks related to automatic music classification, such as feature extraction. The author points out a number of reasons why this a reliable software, including: the ability to assign more than one class to a particular instance, useful in music emotional classification; the capability to label or structure any instance; the ability to logically group features; and the capability to specify and impose a structure on class labels [86].

[29] http://www.cs.waikato.ac.nz/ml/weka/
[30] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox
[31] http://jmir.sourceforge.net/

Considering all of the mentioned options, the model of emotions used to serve the aims of this project was the parametric one. Since it was more challenging and not so widely used on the found systems, thus being one of the must consensual models. When it came to choose a database to serve the aims of this project it was decided to implement one that had the valence and arousal numerical values annotated and being sufficiently diverse to train the machine learning algorithm with different values. The aim is to mirror what is found in every person musical collection.

When it came to choose the tool to extract musical features and the software used to map the emotions it was considered the capability to perform regression since the parametric model was used. Also, the choice considered if the software presented a wide set of musical features and assured the compatibility between modules. It was important that the software had been previously integrated in MER systems, thus validating its capabilities. One of the steps of our project after integrating the database was to delineate which of the above mentioned tools would be actually suitable to the purpose of the system. Thereby, the feature extraction tool that best fit this project was considered to be *MIRtoolbox,* and due to the interoperability that the MER system need the machine learning tool used was *Weka*.

The overall structure of the proposed solution will be explained and the reasons to this selection will be described in the following chapter of this dissertation.

Chapter 4

System design

4.1 - Overview

This chapter describes the creation of a music emotion recognition system. It uses emotion annotations that are associated with each audio piece to train different machine learning algorithms and predict the associated emotion, through the axes points represented on the valence-arousal space.

The system was developed so that an unexperienced user can use it. In terms of software, *MIRtoolbox* is used to extract the musical features more relevant to the prediction of the emotion. This is a toolbox that requires MATLAB environment and also requires the *Signal Processing Toolbox*, which is a sub-package available with MATLAB and is used mainly to correctly analyze the audio presented in each track. Using *Weka*, software in Java that has a collection of different machine learning algorithms the system is able to predict the numerical values associated with each plan of the emotion. To constitute the ground-truth of the system, the machine learning algorithms were trained with the annotated database developed by Solemany et al. [87].

Furthermore, the **Figure 4.6** describes each part of the proposed system model. What follows is a description of the development tools used during this dissertation in order to implement the proposed system.

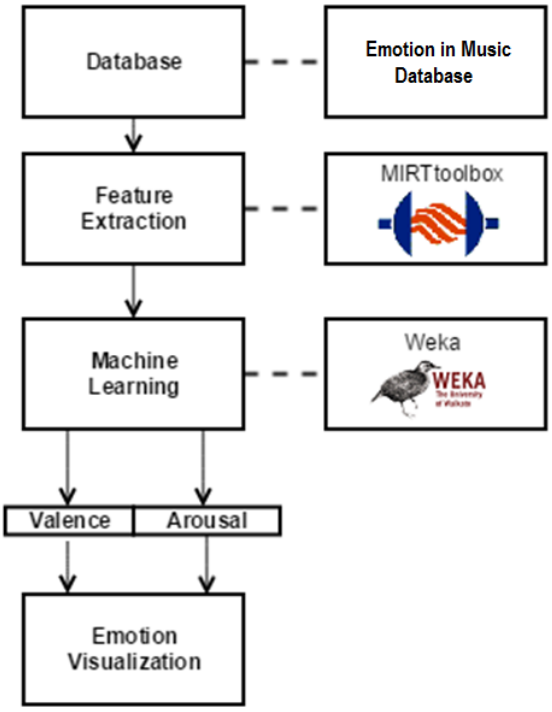


Figure 4.1 - Proposed system model.

4.2 - System Components

As presented in chapter 2 and 3, throughout the literature that reference this matter, there exist several different tools that could be used as a part of the proposed system. The model of emotion selected was the circumplex model proposed by Russel [45], because it represents a wide range of emotions thought mathematical values in the emotional space, and still represents a challenge for the MER systems. Due to the fact that the main goal is to predict the numerical value presented on the valence and arousal plane, it is a regression problem. So, the created system predicts through different regression techniques the emotion annotated. Considering, the importance for this system to be user-friendly, compatible between models and interoperational it was implemented as a MATLAB application.

4.2.1. - Database collection

In order to implement a music recommendation system that takes into account the emotional state perceived by users, it must integrate an annotated database that constitutes the ground-truth to the system. To achieve this aim we selected the existing dataset "Emotion in Music Database" was selected. Besides the fact that it is annotated in the two-dimensional space of emotions it has plenty of diversity represented in eight different main genres. Furthermore, after detecting its redundancies, the creators of this database reduced it to 744 different songs [79]. This will help the system to predict the emotions in a more valid way since the entries that could confuse the machine learning technique are discarded.

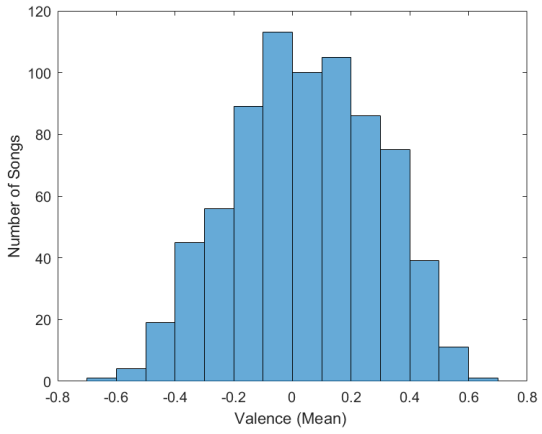
To serve the aims of this project the last 30-seconds excerpts of each 45-seconds song were used, ignoring the first 15-seconds. They are not presented due to the instability of the annotations during the beginning of each track, as Soleymani et al. described [79]. Each song

is in MPEG Audio Layer III format (MP3) and all have a sampling frequency of 44100 Hz. The songs have emotional annotations with a sampling rate of 2Hz. This means that it has continuous annotations every 0.5 seconds of the 30-seconds excerpt. The approach to implement this database on this project was to calculate the average value of all the 60 annotations for every valence and arousal value on the emotional space and the standard deviation to have a notion of the associated error. Each one of the valence and arousal annotations is represented between -1 and +1. **Table 4.1** displays the range of the emotional space that is framed on the dataset.

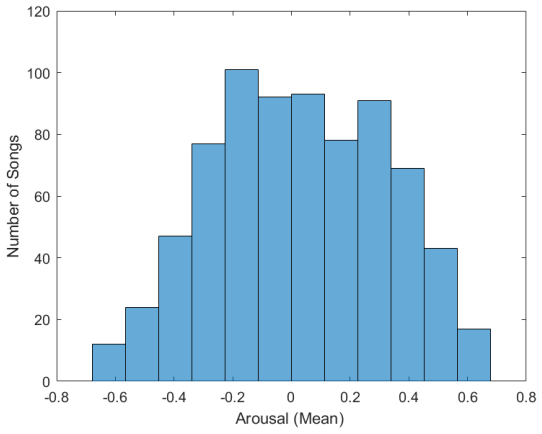
Table 4.1 - Dataset values on the two dimensions.

Value	Range	Mean
Arousal	[-0,673; 0,669]	0,031
Standard Deviation Arousal	[0,038; 0,160]	0,032
Valence	[-0,637; 0,615]	0,044
Standard Deviation Valence	[0,004; 0,162]	0,031

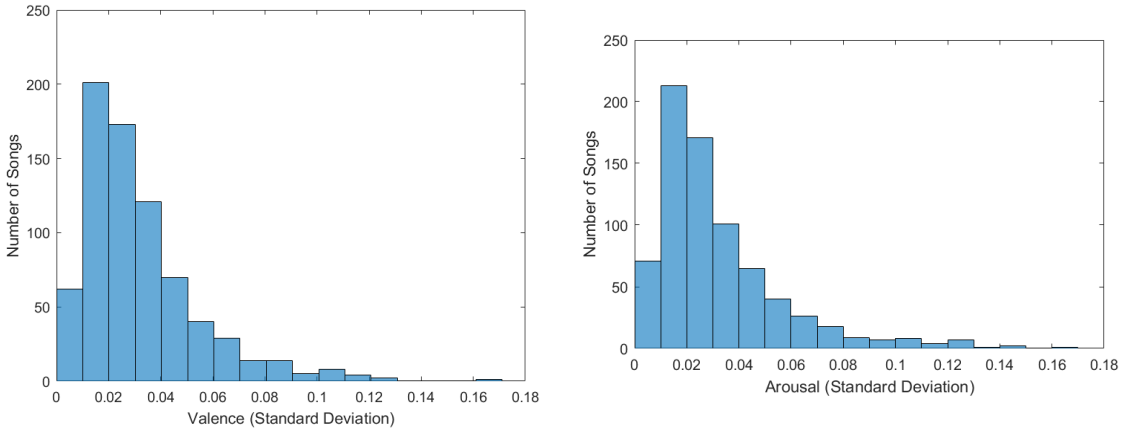
The arousal and valence variables follow a normal distribution (**Figure 4.2 (a)** and **(b)**), with the mean value near zero. It is important to evaluate the consistency of the annotations and since the values are obtained according to the average of the listeners' annotations the larger the standard deviation is, the less representative the ground truth is. **Figure 4.2 (c)** and **(d)** display the histogram of the standard deviation for arousal and valence. In a range of -1 to 1 the majority of the values are inferior to 5% of the range of possible values, it is assumed that does not represent a large margin of error. Yet this value reflects the expected subjectivity mentioned in chapter 2.



(a)



(b)



(c) **(d)**

Figure 4.2 – Histogram of the average of Valence **(a)** and Arousal **(b)** and standard deviation of Valence **(c)** and Arousal **(d)**.

Figure 4.3 illustrates all 744 songs representation within the dimensional plane of valence and arousal dimensions. There is predominance on the first and third quadrant. This means there exists a certain lack of emotions in the positive valence-low arousal associated serene emotions and negative valence-high arousal associated with anxiety/angry emotions.

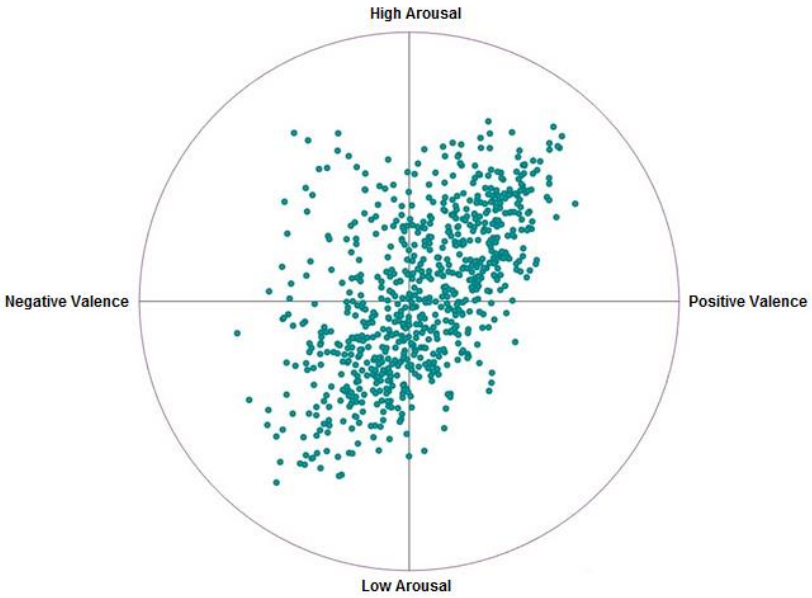


Figure 4.3 - Database songs on the valence-arousal dimensions.

4.2.2. - Feature Extraction Tool

For the extraction of musical features of each song we integrated in the system the *MIRtoolbox*[32] tool designed by Eerola et al. [85]. This toolbox was selected due to its use in for recognition in other MIR systems and its capability to extract a huge number of musical features plus the personal experience in MATLAB.

The extraction of features gives the MER system the capability to analyze the content of the music. Using *MIRtoolbox* allows the system to ideally simulate the perception the music based on specific musical features of each track. This way, the system is able to produce the most accurate results within the operational environment.

Since the aim is to produce a system that can classify a song emotionally it must focus in features that are directly relate to the perception of an emotion in an audio file. Therefore, along the 50 features provided by *MIRtoolbox*, only 22 were considered.

Hence, to produce the best results this module of the system extracts each feature in every 0.5 seconds then calculating the average and standard deviation of each song. This happens so that the extraction has consistency with the selected database.

The musical features extractors within *MIRtoolbox* are organized by 5 main musical dimensions, namely, dynamics, rhythm, timbre, pitch and tonality. The following paragraphs describe the used features that belong to each one of these categories.

Dynamics

The dynamic features are related to the loudness and intensity of a song. It express the physical component of a track.

- Root-mean-square energy: it is defined as the global energy of a specific signal and is computed by calculating the root average of the square of the amplitude. This feature is widely defined as the volume of the waveform. To a certain point of the project it was used also the root-median-square that replaces the mean by the median, but after testing it was found irrelevant the purpose of this dissertation. This feature is define as follows – Equation 2.1

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} - \sqrt{\frac{\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)}{n}}, \tag{2.1}$$

  where x_i are the different values of amplitude throughout time and n is the number of frames with a length of 50 ms.

- Low-energy: In order evaluate the contrast within a song, this feature is used to find the temporal distribution of energy. It is very useful, mainly because it finds if the energy is constant throughout the audio signal. It is define also as short-time energy (STE), and is widely used to the detection of voice [88]. *MIRtoolbox*

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

also computes the Average Silence Ratio (ASR) [89] through the low-energy feature, and it is related to RMS without the root-square.

Rhythm

The rhythm features are directly related with the structure and organization of a musical piece. Also it expresses the time relation within a song. Calculating the rhythm features is of paramount importance when predicting the emotional perception to a musical piece as seen in chapter 2.

- Onset: This feature is used so the system can detect the tempo of a song. It calculates the pulses based on the onset detection curve. Onset estimates the positions of notes by detecting attack phases.

- Event Density: Using the number of notes per second, this feature estimates the frequency of events within an audio excerpt.

- Tempo: Based on the onset detection curve it calculates the musical tempo by detecting periodicities. This is basically the value that corresponds to the speed or pace of a musical piece.

- Pulse clarity: Used for the estimation of rhythmic clarity, indicating the strength of the beats [90]. This means that it measures how easy listeners can capture the rhythmic pulsation of a song. This project used several outputs from this function, based on the autocorrelation curve estimated for tempo calculation. These approaches are based on the analysis of the periodicity of the onset curve via autocorrelation, resonance functions and entropy. This feature has also been widely employed in psychological studies [88].

Timbre

The term timbre covers many perceptual parameters that are not accounted for by pitch, loudness, spatial position, duration, and various environmental characteristics such as room reverberation [91]. In a nutshell, timbre research investigates the ways in which sounds are perceived to differ. This category of features is of crucial importance to this project, due to its relation to emotional states as presented in chapter 2.

- Spectral roll-off point: It is a measure of the skewness of the spectral shape. It estimates the frequency where a percentage of the total power spectral distribution is restricted below. There are mainly two threshold values, 85% proposed by Tzanekis et al. [92] and 95% proposed by Pohl et al. [93]. For this project it was followed the threshold proposed by Tzanekis et al. [92]. This feature is calculate through the equation that follows – Equation 2.2

$$\sum_1^R M[f] = 0{,}85 \sum_1^N M[f], \tag{2.2}$$

Where $M[f]$ is the magnitude of the FFT at frequency bin f and N the Number of frequency bins. R is the roll-off value.

- Brightness: It is a musical feature that calculates the high and low frequencies of a sound, expressing the balance of the signal energy. "A sound is bright when it has more high than low frequencies" [88]. The method used, consists of using the cut-off frequency and measures the amount of energy above it. The cut-off frequency was selected, this project followed the value proposed by Laukka et al. [94] of 1000 Hz.

Signal statistics: We used in the context of timbral features, statistical descriptions of the spectral distribution were used, such as:

- Centroid: Returns the first moment of the signal frame (frequency position of the mean), which is the geometric center of the distribution of the signal measuring its central tendency. The centroid is defined by the following – Equation 2.3

$$\mu_1 = \int x f(x) dx, \tag{2.3}$$

where x represents the value of the signal.

- Spread: Constitutes the second moment of a signal being the standard deviation of a signal spectrum. It measures the variance of the input, meaning the spread of its distribution. This feature is defined as follows – Equation 2.4

$$\sigma^2 = \mu_2 = \int (x - \mu_1)^2 f(x) dx, \tag{2.4}$$

where μ_1 is the centroid value and x the signal value.

- Skewness: Is the third order moment of the spectral distribution. The coefficient of skewness measures the asymmetry of the distribution around the mean value. The returned value can be positive or negative. When positive it means that the distribution is positively skewed with values much larger than the mean. On the contrary, when negative means that the distribution is negatively skewed. Skewness is defined as the following equation – Equation 2.5

$$\mu_3 = \int (x - \mu_1)^3 f(x) dx, \tag{2.5}$$

where μ_1 is the centroid value and x the signal value.

- Kurtosis: This feature calculates the excess of the data. It is the fourth order moment of the spectral distribution. It returns the sum of random variables. This feature is calculated using the following equation – Equation 2.6

$$K = \frac{\mu_4}{\sigma^4}, \tag{2.6}$$

where μ_4 is the fourth cumulant and σ^4 is the square of the square of the variance of the probability distribution. K is the kurtosis value. The fourth is defined by the following equation:

$$\mu_4 = \int (x - \mu_1)^4 f(x) dx, \tag{2.7}$$

where μ_1 is the centroid value and x the signal value.

- Flatness: Indicates if the distribution is uniform in its frequency distribution of the power spectrum, and its calculated dividing the geometric mean by the arithmetic mean. This feature is defined as follows – Equation 2.8

$$\frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\left(\frac{\sum_{n=0}^{N-1} x(n)}{N}\right)}, \tag{2.8}$$

where x the signal value, N the total number of frames and $x(n)$ the value of the signal on the frame n.

- Entropy: Describes the spectrum uniformity. In this project it is followed the Shannon approach. This is widely used in information theory and indicates if the data contains predominant peaks estimating a general description of the input curve.

- MFCCs: As mentioned in chapter 2 this is a feature that is widely used in MIR systems. It offers a description of the spectral shape of a sound. Its coefficients are a representation of the power spectrum. These coefficients are calculating using the Fourier Transform of a signal, then mapping the powers of the spectrum onto the mel scale, then taking the logarithms values of the mel frequencies they are decorrelated using the Discrete Cosine Transform (DCT). *MIRtoolbox* returns 13 coefficients. In order to calculate the MFCCs the total energy in each critical band is used as described on the following equation - Equation 2.9.

$$Y(i) = \sum_{k=0}^{N/2} log|s(n)| \cdot H_i\left(k \cdot \frac{2\pi}{N}\right), \tag{2.9}$$

where $Y(i)$ is the total energy in the critical band, N is the framelength, $s(n)$ is DFT signal for which the MFCC's is calculated, H_i is the critical band filter at the i[th] coefficient and n is the number of points used in the short term DFT (with zero padding).

Afterward it computes the actual IDTF to get the coefficients. For this goal it handles each critical band individually, which is expressed on the Equation 2.10

$$\tilde{Y}(k) = \begin{cases} Y(i), & k = k_i \\ 0, & other\ k \end{cases} \in [0, N-1]\ , \tag{2.10}$$

where $\tilde{Y}(k)$ is the total energy each critical band.

The final cepstrum can be derived by the following equation – Equation 2.11

$$c_s(n) = \frac{1}{N} \cdot \sum_{k=0}^{N-1} \tilde{Y}(k)\, e^{jk\left(\frac{2\pi}{N}\right)n}, \tag{2.11}$$

If the real cepstrum is used, the sequence $\tilde{Y}(k)$ is symmetrical (even) about the critical band center frequency. The previous equation can therefore be reduced to the equation 2.12.

$$c_s(n) = \frac{2}{N} \cdot \sum_{i=1,2,\ldots,N_{cb}} \tilde{Y}(k_i) \cdot \cos\left(k_i \cdot \frac{2\pi}{N} n\right), \tag{2.12}$$

where N_{cb} is the number of critical bands. $c_s(n)$ is the calculated coefficients.

- Irregularity: This feature calculates the successive variation of peaks of a spectrum. *MIRtoolbox* allows the implementation of two different theoretical approaches. In the first one, irregularity is the sum of the squares of the different in amplitude between adjoining partials and was proposed by Jensen [95]. In the second one, irregularity is the sum of the amplitude minus the mean of the preceding, same and next amplitude, this approach was proposed by Krimphoff [96]. The both were used on this project. Irregularity is calculated as follows – Equation 2.13

$$\sum_{k=2}^{N-1} \left| a_k - \frac{a_{k-1} + a_k + a_{k+1}}{3} \right|, \tag{2.13}$$

where a_k is the amplitude of the signal.

Pitch

Pitch is the subjective perceptual attribute that allows listeners to rank sounds from low to high [97]. The perception of pitch forms the basis of musical melody and harmony. It consists of one of the most commonly used features in MER.

- Fundamental frequency: Also commonly called fo, correspond to the lowest frequency of a harmonic stationary audio signal, being the peak of the autocorrelation factor. This feature is defined but the following equation - Equation 2.14

$$f_0 = \frac{1}{T}, \tag{2.14}$$

  where T is the period of the waveform. f_0 is the fundamental frequency.

- Inharmonicity: Measures the extent of partials that are not multiples of the fo. In *MIRtoolbox* inharmonicity calculates the amount of energy outside the ideal harmonic series.

Tonality

It is a characteristic that organizes the notes of a musical scale according to a specific musical criteria. Furthermore, tonality is directly related to harmony, which describes the structure of sounds constituted by a series of harmonically related frequencies.

- Key: Returns the estimation of tonal center positions and its clarity.

- Key Strength: Calculates in a score between -1 and +1, the key strength of each possible candidate, based on a cross-correlation of the chromagram. Returns two arrays of values for each major and minor tonality.

- Mode: Calculates the modality of a song, returning a value between -1 and +1. If the value is closer to +1, it means the more major the given excerpt is predicted to be. If is closer to -1 the more minor the excerpt is.

- Tonal Centroid: Estimates the tonal centroid vector from the chromagram, returning an array that corresponds to 6-dimensions. It is based on the work of Harte et al. [95], which projects the chords along circles of fifth, minor thirds and major thirds.

All of the above features were used in the system to test which combination would produce the most accurate results

4.2.3. - Machine Learning

To implement machine learning techniques and develop a trained model *Weka* was used. As mentioned in chapter 3, *Weka* is a platform that provides several machine learning algorithms that can be used on regression contexts. Besides, *Weka* provides a graphical user interface (GUI), illustrated in **Figure 4.4** that is very intuitive to use. *Weka* is a software implemented in *Java* but MATLAB has a wrapper that allows communicating with Java. To serve the flexibility of this project and based on the need to perform analyses on MATLAB it was used an interface *matlab2weka*[33] developed by Sunghoon Lee. The main advantage of this interface is that it converts the file with the features extracted into an interface object of *Weka*, namely an ARFF file.

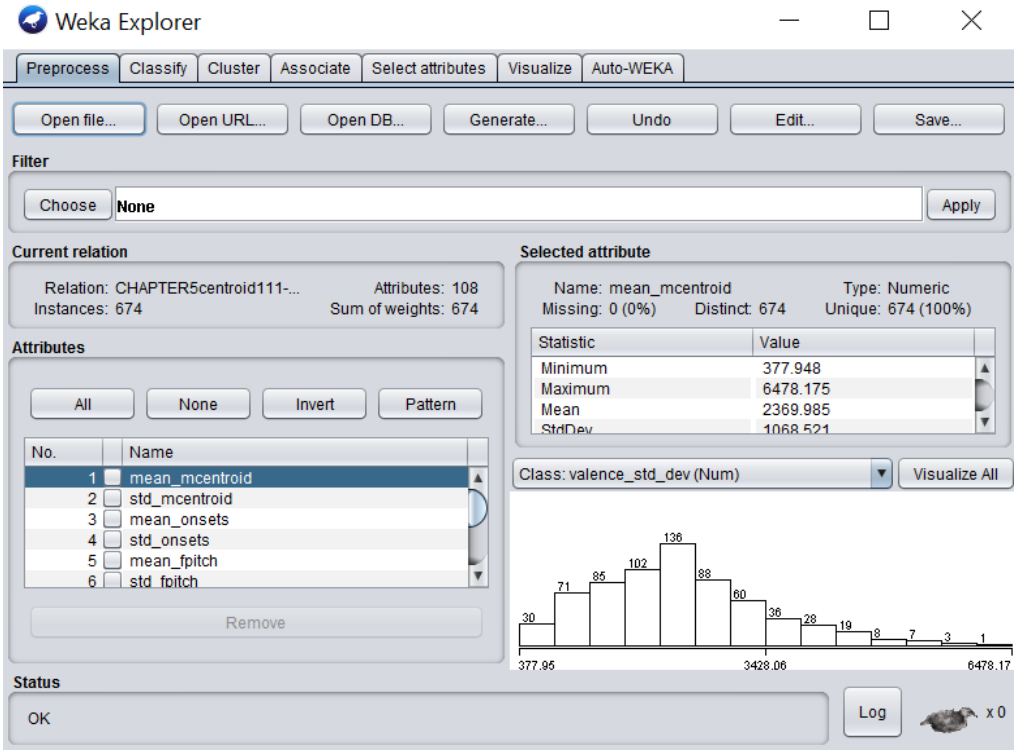


Figure 4.4 - Screenshot of Weka work environment

[33] http://www.sunghoonivanlee.com/matlab2weka.html

Machine Learning Algorithms

Regarding the subject of this dissertation, a machine learning algorithm is mainly used to predict the values of valence and arousal. Due to the aim of predicting numerical values in [-1, 1], the problem of mapping the emotion was formulated as a regression problem. During the development of the system there were used 5 main different algorithms.

Firstly, *Auto-Weka* was used on the dataset [98]. This is a software implemented in *Weka* that is capable of considering the database, selecting the machine learning algorithms that best fits. The output of this application was Multiple Linear Regression.

Furthermore, Simple Linear Regression was used using the implementation of *Weka* to compare to the values predicted and due to its simple implementation in the system. This regression method as well as MLR uses linear classifiers to predict.

Considering this and due to the fact that Support Vector Regression has been widely implemented on regression problems, was decided to implement SVR as the predictor in this system as well [69].

Gaussian Processes Regression were used using the *Weka* implementation on the system due to the results obtained on recent work regarding MER systems [72]. Recently, this method has increasingly been used on problems with big data.

To have a different approach and since this algorithm have not been widely used on regression problems regarding MER systems, K-Nearest Neighbor Regression were used as well. The K-Nearest Neighbor algorithm has been mainly applied on classification problems [26]. Each one of these algorithms will be described on the following paragraphs.

- Simple Linear Regression

This algorithm learns a simple linear regression model and it was used using the *Weka* implementation [84]. A linear regression model attempts to explain the relationship between attributes, in this system features, and emotions using a straight line ($f(x) = mx + c$), in which $f(x)$ is the relation between the attributes and the emotion. The SLR algorithm, selects the attribute that results in the lowest squared error prediction. Due to the fact that the prediction is related to a single predictor, also called explanatory variable, it is called simple. It is important to use this algorithm on this project to have a better notion of the results if it was only used the attribute that leads to the best overall error.

- Multiple Linear Regression

This approach calculates the relation between different explanatory variables and the value to predict. Since it uses more than one explanatory variable it is called Multiple Linear Regression.

In linear regression, all the relations are modelling by using functions called linear predictors. These linear models parameters are calculated from the data. This MLR implementation uses the Akaike criterion for this model selection [84]. This criterion measures the quality of different models for a specific data. This will enable the prediction to be more accurate since this criterion optimizes the system.

- Support Vector Regression

As stated in chapter 2 Support Vector Machine (SVM) can be applied to regressions problems. The method implemented in this project is based on the iterative sequential minimal optimization algorithm (SMO) that solves the regression problem using SVM [70]. The

SMO algorithm contains many optimizations that were designed to speed up and improve its strength. In this approach the algorithm is selected by setting the *RegOptimizer*, a popular algorithm designed by Shevade et al [99]. This approach uses RBF Kernel to make the SVR algorithm nonlinear.

- Nearest Neighbour Regression

K-nearest neighbors stores all available cases and predict the numerical target based on a similarity measure, such as distance functions. Its implementation is based on the work of Aha et al. [100] and it can select automatically the appropriate value of K based on cross-validation. A simple implementation of K-NNR is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. This implementation can also do distance weighting.

- Gaussian Processes Regression

Implements Gaussian processes for regression based on the work of Mackay [101]. Gaussian processes extend multivariate Gaussian distributions to infinite dimensionality. GPR is a probabilistic model which produce Gaussian distributions as their output. It is a Bayesian nonparametric models that capture highly nonlinear data relationships. Rather than assuming that f(x) relates to some specific models, such as linear regression does, a Gaussian process represent f(x) obliquely. This means that this algorithm analyses the data before any assumption. GPR is still a form of supervised learning, but the training data is not used in an abrupt way. To make choosing an appropriate noise level easier, this implementation applies normalization to the target attribute as well as the other attributes. Moreover, in this implementation polynomial kernel was used and the prediction of values occurs without hyper parameter-tuning.

4.3 - Music Emotion Recognition System

This section describes the perspective to an automatic music recommendation system based on emotion. As described before it results in a MATLAB application oriented to the recognition of emotions in audio files. The GUI of the proposed system is illustrated in the **Figure 4.5**.

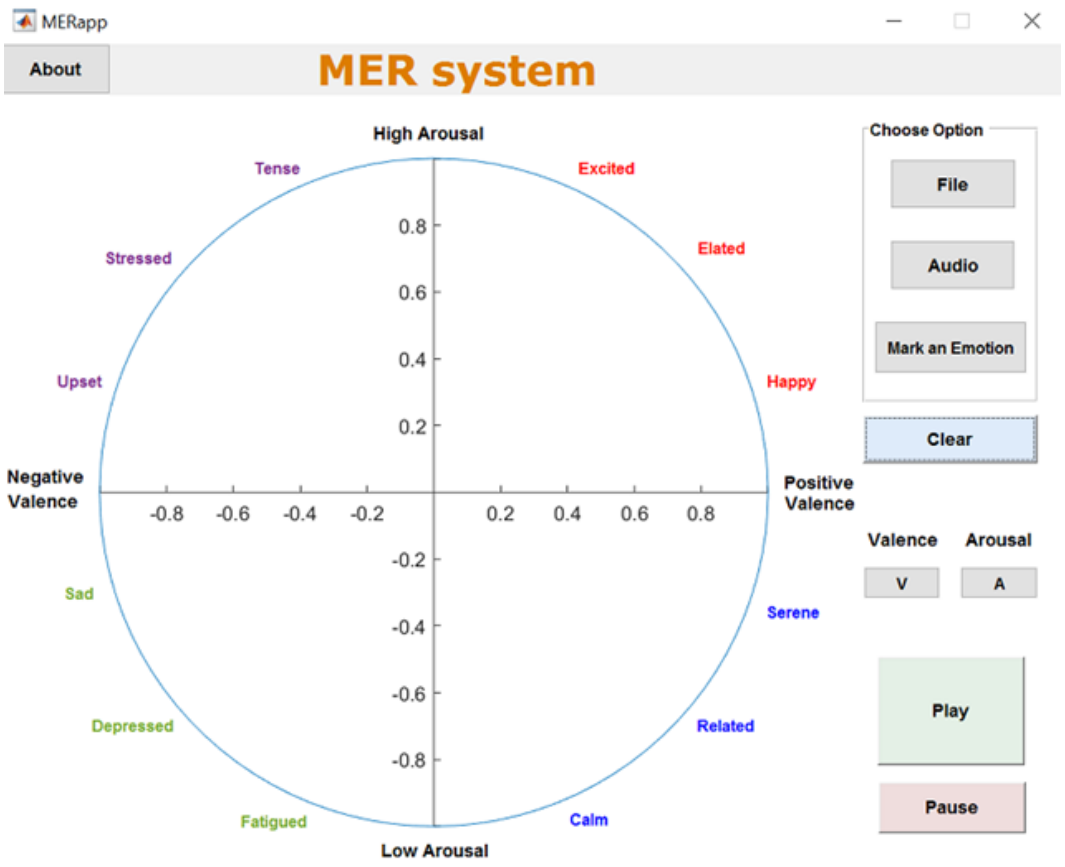


Figure 4.5 - Screenshot of the Graphical User Interface (GUI) of the proposed MER system.

The system is composed of one main screen where the valence-arousal space appears. This facilitates the user interaction allowing the visualization of the emotion.

There is an “About Panel” that introduces the system to its users, providing an explanation and instructions to its use. **Figure 4.6** displays the about screen.

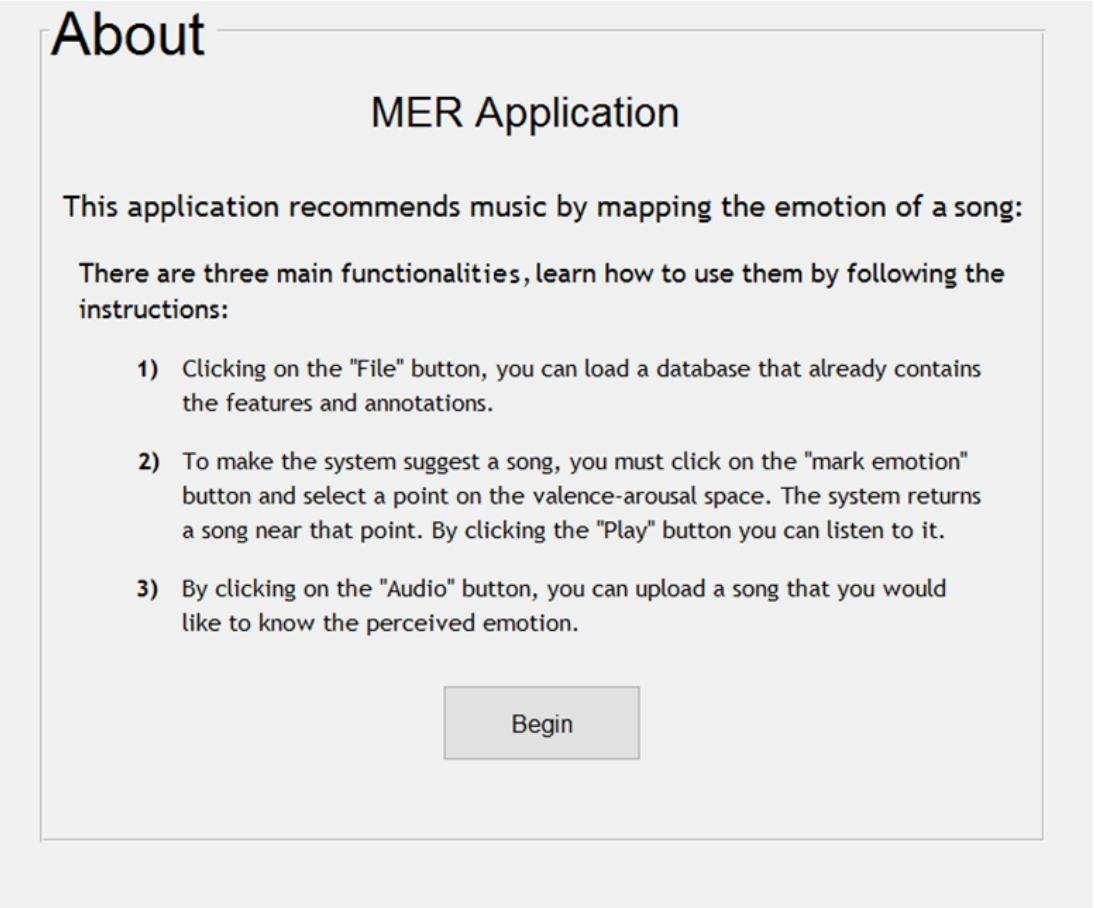


Figure 4.6 - System's about screen panel.

The developed MER system has three main usability's, training the system using a specific annotated database, marking the emotion that the listener wants to perceive and mapping the emotion of a certain song. These functionalities were created based on different potentialities that this system would be able to perform. The further sections will explain each one of the system's components.

4.3.1. - Loading a database

Considering that the main goal of this dissertation was to map an emotion through machine learning techniques, it was necessary to implement in the system a functionality that allowed the user to load a dataset already annotated. Ideally this dataset would have features extracted by using the *MIRtoolbox* so that the predictions can be made by the MER system. This functionality enables the system to train the machine learning algorithm, considering the loaded database. Furthermore, it allows the system to rapidly map the emotions and visualize the outcomes in the valence-arousal space.

Thus, this was the first module implemented on the system due to the necessity to implement the machine learning algorithm and train the selected dataset. Since the features were already extracted from the database, there was no need to repeat the process systematically. The respective process is described in **Figure 4.7**.

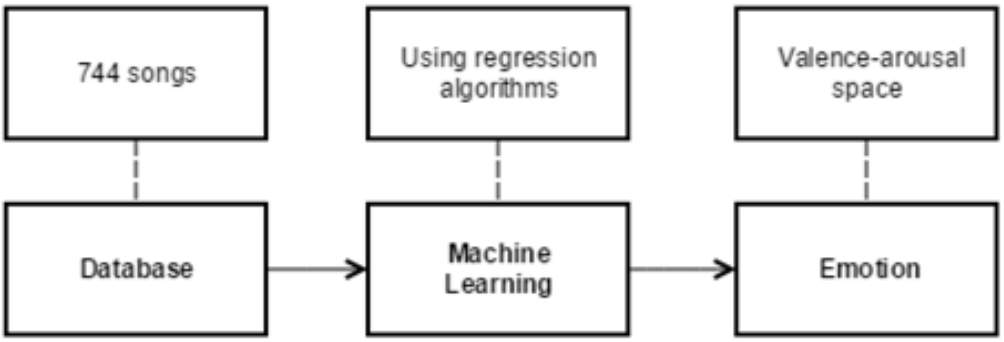


Figure 4.7 - Training the system using the annotated database.

Basically, as shown in **Figure 4.7**, the loading process does not need the computational time to extract the features. This allows the system to more rapidly predict the emotional states. It is also important that the loaded file must be in Attribute-Relation File Format (ARFF), a specific use with the *Weka* software or CSV (comma-separated values) in order to correctly evaluate the attributes and instances.

4.3.2. - Marking the emotion

Regarding, the perspective to a future usability of the MER system, when entering the main screen of the application, the user is able to mark an emotion in the valence arousal space. This option will allow the user to hear a song of the loaded database. The song that would be played is the one whose value is near the marked point in the space. Basically, the system searches through the dataset the difference between each valence and arousal point of each song and returns the one which the minimum euclidean distance. The process is illustrated in **Figure 4.8**.

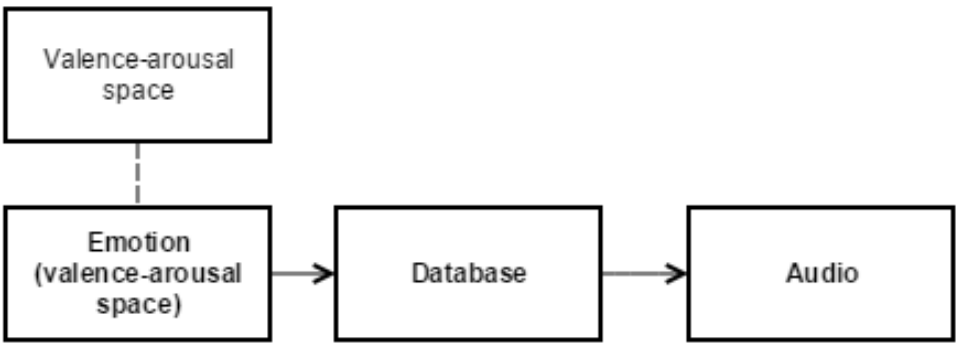


Figure 4.8 - Marking the emotion that listener wants to perceive.

This functionality was developed to serve the aim of a recommendation system. The user marks the emotion that he wants to perceive and the system recommends a song based on the user interaction as seen in **Figure 4.9**.

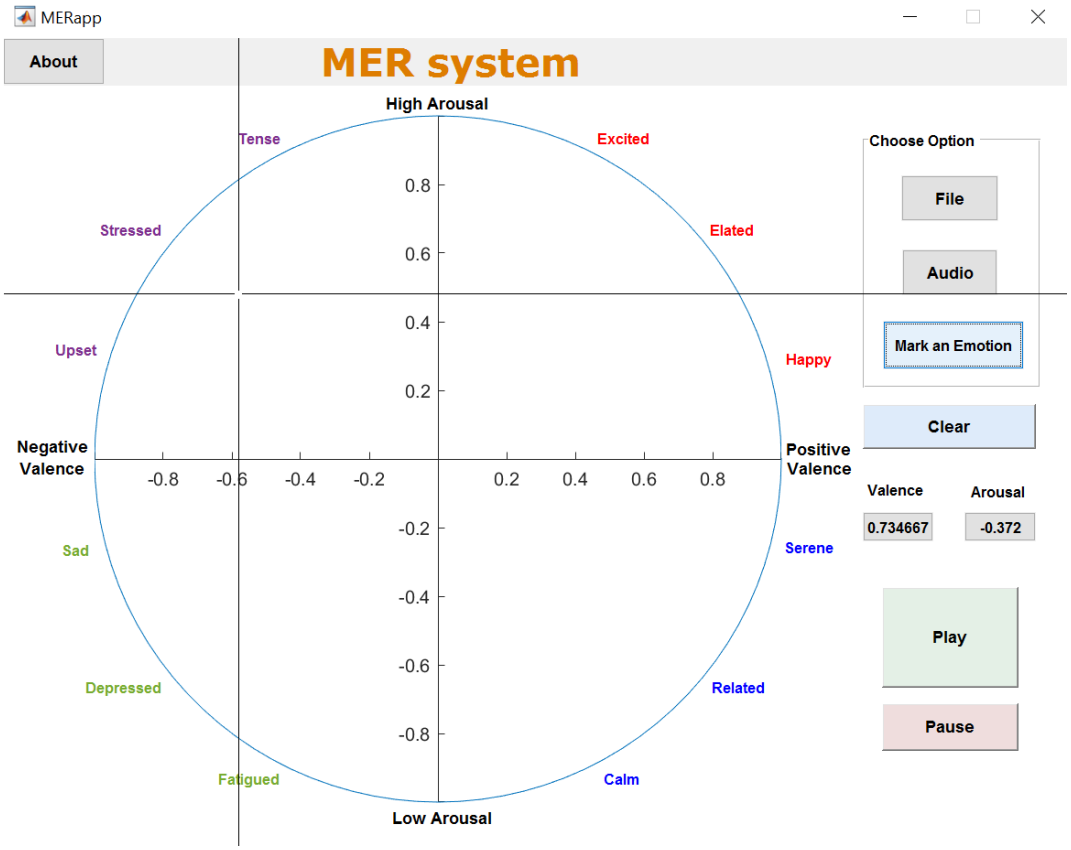


Figure 4.9 - Marking an emotion on the emotional plane.

4.3.3. - Mapping the emotion

This is the functionality where the system predicts the emotion of an audio file that has not been previous annotated on the valence-arousal space. When opening the MER system in the main panel, the user through manual interaction can load an audio file. This audio file can be in MPEG-1 or MPEG-2 Audio Layer III (MP3) format or Waveform Audio File Format (WAVE). The "Audio Button" allows the user to load an audio file so that the system through machine learning techniques can predict the emotion associated. The process is explained in **Figure 4.10**.

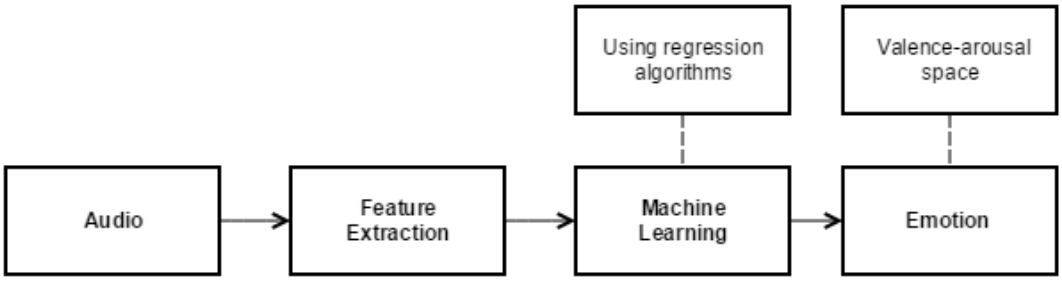


Figure 4.10 - Predicting an emotion associated with a certain audio.

Firstly, it extracts the features using *MIRtoolbox* and stores it in a CSV file, then this file enter in the machine learning module, and then by using machine learning algorithms it predicts the associated emotion. The system through machine learning models predicted the emotion of the song. The system then returns the valence and arousal value displaying it on the plane, as seen in **Figure 4.11**.

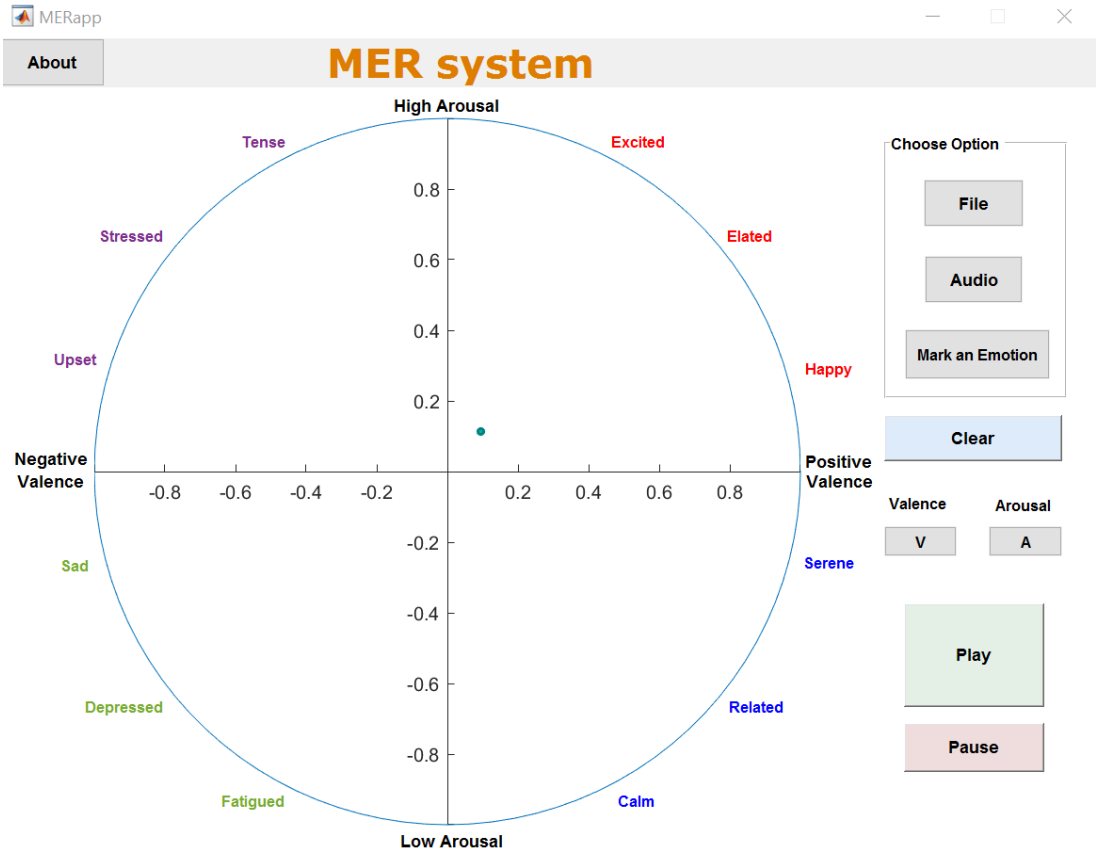


Figure 4.11 - Visualizing the predicted emotional value on the plane.

This capability of the MER system is also used when loading a database that has the features extracted, entering on the machine learning function that predicts the value associated. This allows the user to visualize the emotional space within a certain collection of music.

Chapter 5

Evaluation

5.1 - Results

5.1.1. - Overview

In this chapter, it is described different tests performed on the proposed system described in chapter 4. The results are then explained and interpreted in this chapter as well.

The main purpose is to evaluate the system performance using different machine learning algorithms and different combinations of features in order to achieve the most accurate system possible. Thereby, there exist two main scenarios regarding the different modules of the MER system. Based on the state of the art [26] [31], the perspective were that the Support Vector Regression (SVR) model would have the most accurate results. Although the work of Markov et al. [72] proved that he Gaussian Process Regression (GPR) is also a valuable solution.

The two different main test scenarios are described and the results were collected and then evaluated. Then these results were compared to the existing solutions found on the literature.

Having in mind the main purpose of the recommendation system, after the analyses of the results, the system was validated with the most accurate bag of features and machine learning algorithm.

5.1.2. - Experiments

As presented in section 3.2, the machine learning techniques and the features have a direct influence on the results but they are independent from each other. Hence, the aim is to capture the influence of different features and machine learning algorithms on the accuracy of the proposed system independently.

Having this in mind, it was decided that there would be two paramount scenarios to compare:

- Scenario 1 - Machine Learning Algorithm: same features using different machine learning algorithms;

- Scenario 2 - Feature Selection: different features using the same machine learning algorithm;

In all the different experiments, the performance of the regression was evaluated by 10-fold cross validation technique due to its use on the literature [26]. Moreover, this methodology will help to compare the different projects. This technique divides the database into 10 random parts where 9 of them are used to train the algorithm and the other one to test and validate it, as illustrated in **Figure 5.1**. This procedure is repeated 20 time and the average results are calculated.

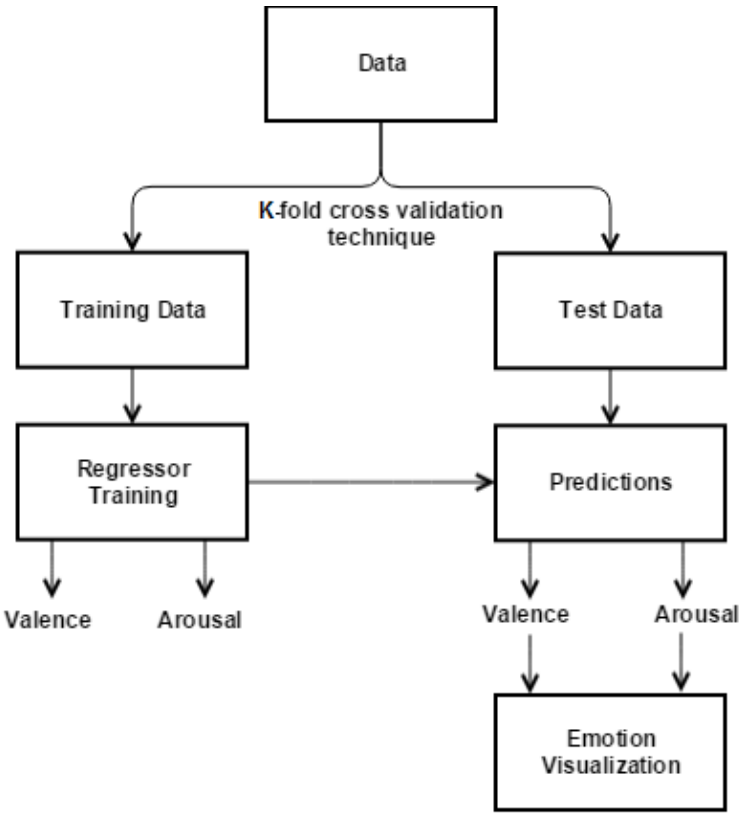


Figure 5.1 – Procedures for system training.

The results will be compared in terms of RSME (Root Mean Squared Error) and R^2 statistics; this is a standard way of measuring the reliability of the regression models, that it is defined as follows – Equation 5.1:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad (5.1)$$

where y_i is the i^{th} value of the attribute to be predicted, $f(x_i)$ is the predicted valu, $\bar{y}$ is the mean of the ground-truth and n the total number of observations. The residual sum of squares makes R^2 comparable between different scenarios.

Machine Learning Algorithm

Firstly, all the features provided by *MIRToolbox* related to loudness, level, dissonance, tonality and pitch were used since they are directly related to emotional perception [102]. Tis approach is based on the previous work developed by Yang et al. [69], comparing directly the results with the ones obtained in this project. The features regarding this scenario are displayed on **Table 5.1**.

Table 5.1 - Features used to train each model

Features	Type
RMS	Level
Low-Energy	Level
Centroid	Loudness
Skewness	Loudness
Kurtosis	Loudness
Entropy	Dissonance
Irregularity	Dissonance
fo	Pitch
Roll-off	Pitch
Key	Tonality
Mode	Tonality
KeySound	Tonality
Tonal Centroid	Tonality

The accuracy of each prediction on the valence-arousal space is measured in terms of R^2, as stated before. This value when approximately 1 means that the prediction values are similar to the database values, when negative it means that the model is worse than simply taking the database mean into account. The values are displayed in percentage. As mentioned in chapter 4 there are five algorithms implemented on the system and each one is based on the *Weka* software implementation. Every prediction of the value of valence or arousal is computed separately. This mean that, the algorithm does not have the valence value when prediction the arousal, and vice-versa.

As mentioned before, the higher R^2 values mean that the prediction was closer to the actual value. On the other hand when analyzing the RMSE results, lower results mean that the system is more accurate.

The results when using this bag of features on the Simple Linear Regression (SLR) algorithm for each dimension are presented on **Table 5.2**.

Table 5.2 - Results when using SLR algorithm

Model	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
SLR	39.7%	14.5%	0.225	0.223

To have a better notion of what this method is capable of, it is shown in **Figure 5.2** the distribution of the actual values of the database and the predicted values by the SLR algorithm.

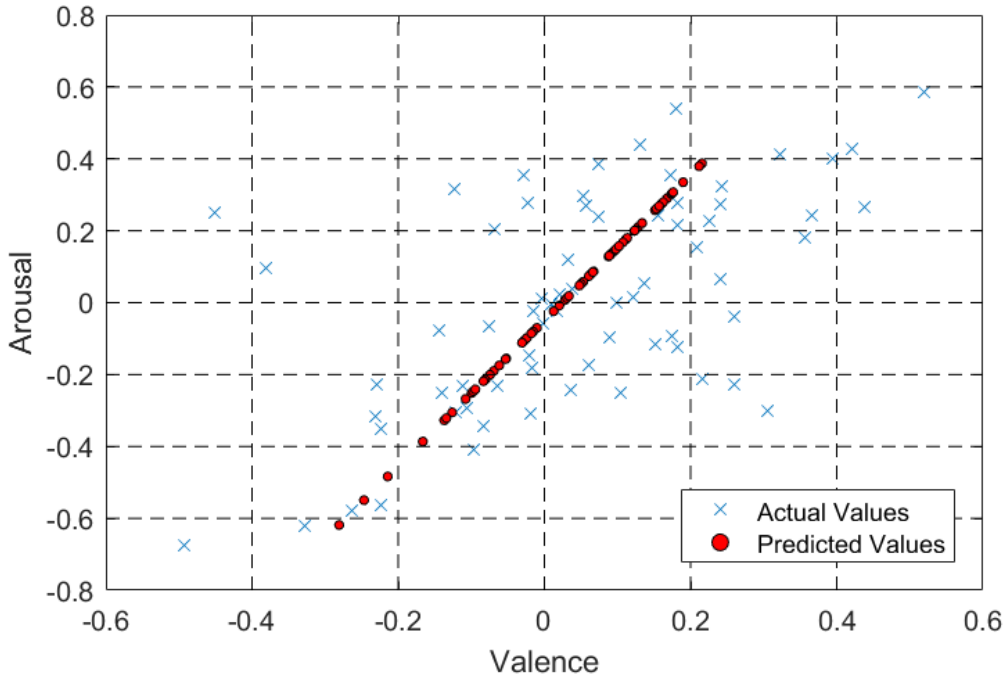


Figure 5.2 - Distribution of the ground-truth (blue cross) and the prediction values (red points) using the SLR algorithm.

As shown in the previous **Figure 5.2** and has expected, the predictions within this module follow a linear function, it is rational since as seen before it uses a straight line to predict the values of arousal and valence.

From now on, the figures displayed will connect the ground-truth values to the predicted ones for a better visualization of the differences between each method. Thus, the following **Figure 5.3** illustrates the results on this manner.

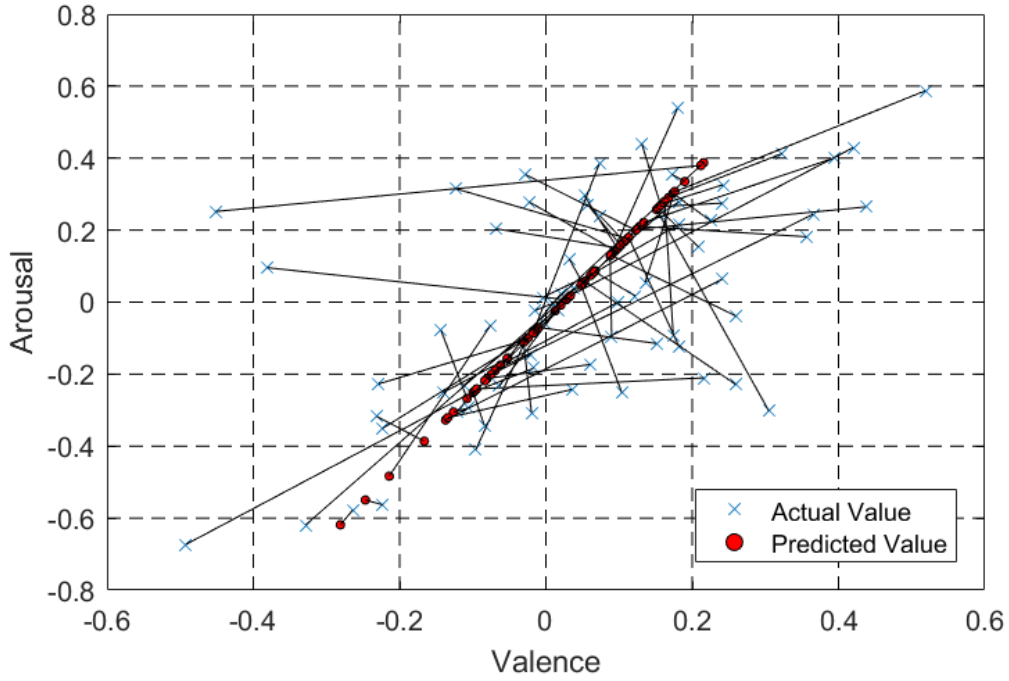


Figure 5.3 - Distribution of actual values (blue cross) with lines connecting to the predicted values (red points) when using the SLR algorithm.

It is important to state that each one of the tests proceeded divide the database into 10 random parts for cross-validation, so the values displayed in the figures will not be the same for all the algorithms. Testing the annotated database with the Multiple Linear Regression (MLR) algorithm produces the results presented on **Table 5.3**.

Table 5.3 - Results using MLR.

Model	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
MLR	55.6%	29.6%	0.198	0.205

Analysing the values displayed in **Figure 5.4**, it is reasonable to state that this method produces results more approximated to the actual value. This happens due to the use of more than a single explanatory variable that calculates a linear relationship with the predicted value. The distribution of the values is widely spread on the valence-arousal space.

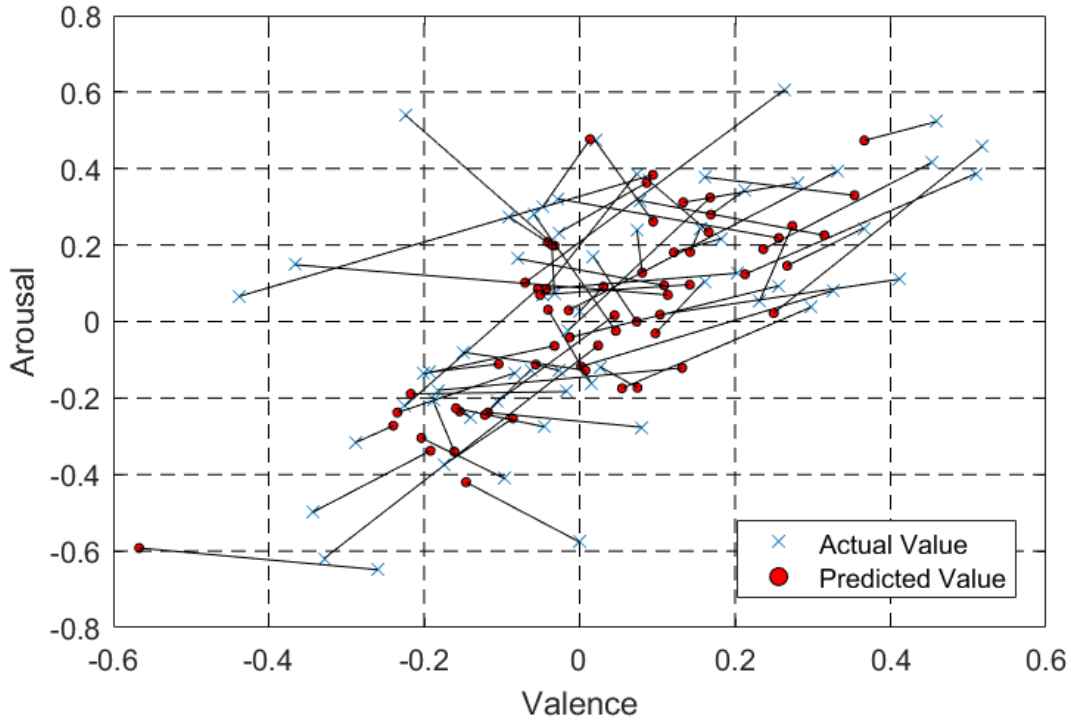


Figure 5.4 - Distribution of actual values (blue cross) with lines connecting to the predicted values (red points) when using MLR algorithm.

Using the Support Vector Regression (SVR) algorithm, the results obtained with the testing dataset are presented on **Table 5.4**.

Table 5.4 - Results using SVR.

Model	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
SVR	55.4%	29.8%	0.209	0.201

Furthermore, the **Figure 5.5** illustrates the difference between the actual values and predicted using the algorithm. It can be observed, that combined distribution are quite identical.

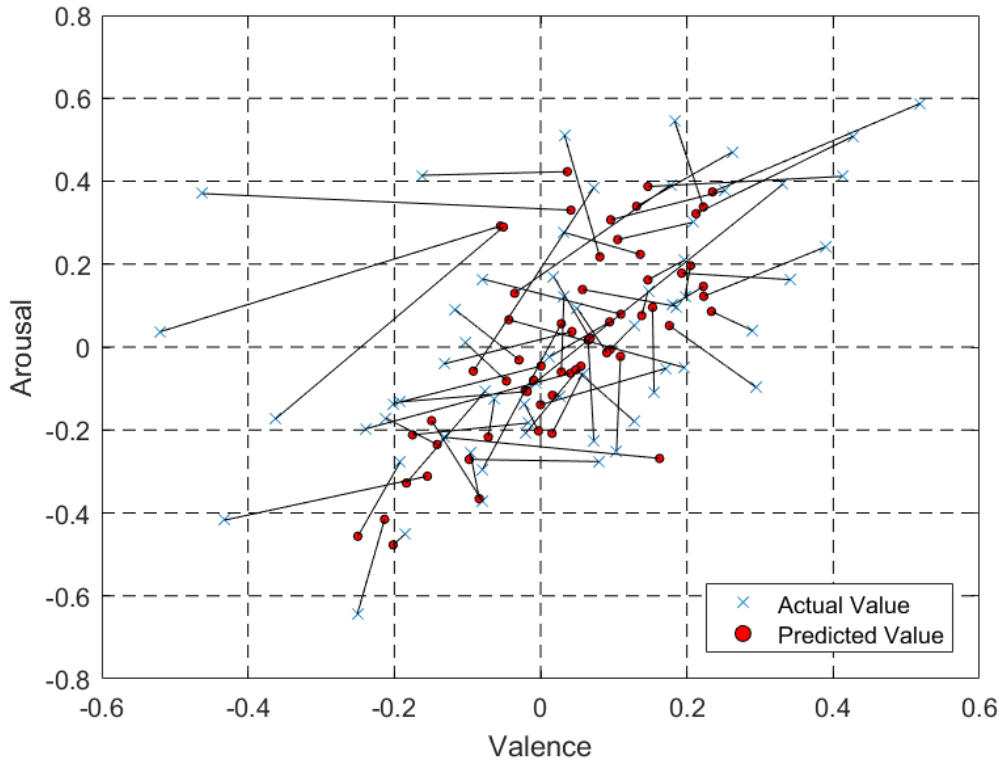


Figure 5.5 - Distribution of actual values (blue cross) with lines connecting predicted values (red points) when using the SVR algorithm.

The R^2 statistics and RMSE achieved values when testing the K-NNR algorithm are presented on **Table 5.5**.

Table 5.5 - Results using K-NNR.

Model	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
K-NNR	43%	21.4%	0.234	0.223

The visualization of the same results is shown in **Figure 5.6**. It can be perceived that the predicted values distribution is quite centralized on the valence-arousal plane, and almost following a linear distribution. This probably happens due to the applied method, which is based on a similarity measure, calculating the average of the numerical target of the K nearest neighbors.

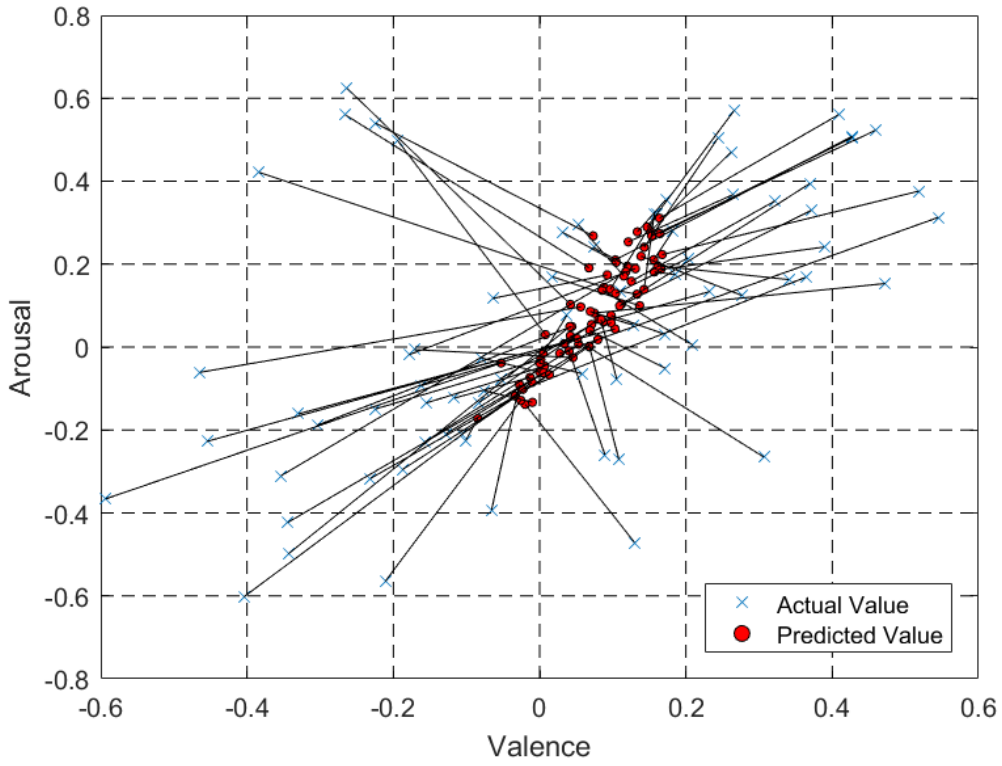


Figure 5.6 - Distribution of actual values (blue cross) with lines connecting to the predicted values (red points) when using the K-NNR algorithm.

Lastly, the same test was applied on the Gaussian Processes Regression (GPR) algorithm and the results are presented on **Table 5.6**.

Table 5.6 - Results using GPR.

Model	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
GPR	53.8%	33.5%	0.208	0.203

Figure 5.7 illustrates the results when using the GPR algorithm. It can be observed that the predictions are more spread on the valence-arousal plane approximating its value from the real one.

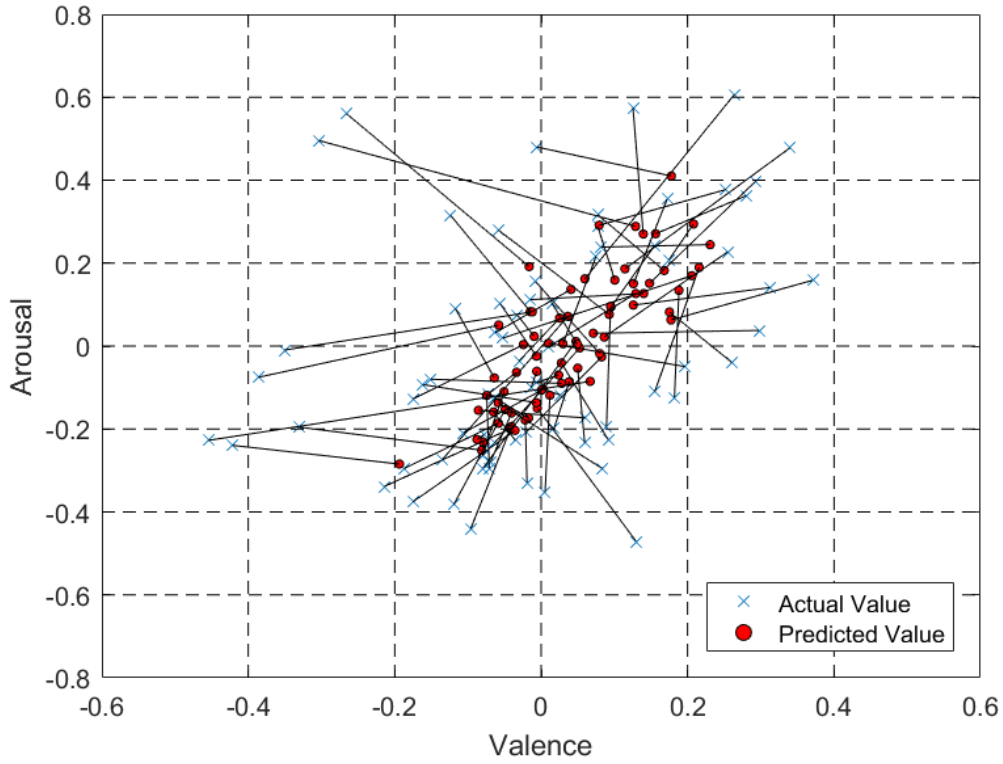


Figure 5.7 - Distribution of actual values (blue cross) with lines connecting to the predicted values (red points) when using the GPR algorithm.

On the following subsection it will be described the second test scenario regarding the feature selection.

Musical Features Selection

The musical features have always received great attention regarding the MER systems. The association of emotion and music with acoustic proprieties facilitates the aim of any of these systems. As described in Chapter 2, specific features are directly related with the valence and arousal perception.

Since *MIRtoolbox* is divided into five main categories of features, namely pitch, rhythm, timbre and tonality, it was tested which one of those would be more determinant on mapping the emotion. Also as seen in Chapter 2 it is usual to divide the feature space into different main categories. This approach was also considered by Chen et al. [73] that also used *MIRtoolbox*. This test introduced features that were not tested in the previous section, more importantly features related to the rhythm.

Considering the SLR algorithm, the obtained results are displayed in **Table 5.7**. Analysing the results it is clear that the most relevant features to this algorithm are the dynamics and timbre being the R^2 results respectively 53.8% and 45.2%. The tonality and rhythm produce the most inaccurate results.

Table 5.7 - Results using different categories of features with SLR.

Category	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
Dynamics	53.8%	33.5%	0.208	0.203
Rhythm	10.4%	17.2%	0.28	0.225
Timbre	45.2%	17.7%	0.223	0.221
Pitch	35.3%	14.8%	0.235	0.220
Tonality	9.3%	1.8%	0.284	0.242

The results produced by the MLR algorithm are illustrated in **Table 5.8**. The most relevant features to this algorithm are the same as the SLR, namely the *dynamics* and *timbre*. In this case the R^2 results are above 55%. The tonality and rhythm produce the most inaccurate results. Although it is important to state that as the previous algorithm, the rhythm produces more accurate results in the valence value, this happens mainly to the fact that rhythm features are directly related to the valence perception [2].

Table 5.8 - Results using different categories of features with MLR.

Category	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
Dynamics	62.3%	35.2%	0.163	0.239
Rhythm	16.7%	23.4%	0.269	0.212
Timbre	56.7%	22.3%	0.196	0.219
Pitch	34.5%	12.9%	0.231	0.230
Tonality	26.9%	7.0%	0.261	0.249

Using the same test scenario, the results produced by the SVR algorithm are illustrated in **Table 5.9**. The most relevant features are the same as in the previous algorithms. In this case the R^2 results are above 50% only in the timbre case. Although is important to consider that the results are similar to the MLR algorithm excluding the dynamic related features.

Table 5.9 - Results using different categories of features with SVR.

Category	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
Dynamics	46.4%	27.7%	0.261	0.249
Rhythm	16.6%	25.8%	0.272	0.217
Timbre	50.4%	19.8%	0.201	0.219
Pitch	34.9%	12.3%	0.243	0.224
Tonality	25.3%	5.7%	0.257	0.235

The results obtained with the K-NNR algorithm are illustrated in **Table 5.10**. The most relevant features are the same as in the previous algorithms. In this case the R^2 results are above 50% only in the dynamics case. So comparing with scenario of the SVR algorithm, the

opposite happened. It is importance to note the significant decreasing values regarding the tonality features, being the RMSE higher than any other previous case.

Table 5.10 - Results using different categories of features with using K-NNR.

Category	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
Dynamics	52.6%	32.6%	0.171	0.241
Rhythm	12.6%	23.9%	0.274	0.216
Timbre	46.2%	23.1%	0.219	0.221
Pitch	31.4%	11.2%	0.237	0.233
Tonality	11.4%	2.9%	0.287	0.240

Considering the GPR algorithm the results are present in **Table 5.11**. The arousal values are higher on the dynamics and timbre cases. Concerning the valence values, these results were unexpected, especially the dynamics and pitch cases. Considering the work of Markov et al. [72] that stated R^2 statistics results around 30% when using features related to those categories. It is important to observe that the RMSE values are higher in this scenario than in any previous test conducted before.

Table 5.11 - Results using different categories of features with GPR.

Category	R^2 Statistics		RMSE	
	Arousal	Valence	Arousal	Valence
Dynamics	61.3%	10.4%	0.166	0.458
Rhythm	15.1%	18.7%	0.279	0.232
Timbre	48.5%	20.0%	0.216	0.228
Pitch	11.9%	2.5%	0.436	0.435
Tonality	25.0%	6.5%	0.254	0.232

5.2 - Comparison

In order to compare the results it is crucial to identify the importance of the musical feature space. From the point of view of the machine learning algorithm, features do not always have equal importance. Moreover, features that are irrelevant may lead to inaccurate results. As pointed previously, this project is based on previous works and knowledge about music and emotion in order to identify the paramount features.

The results for each method are summarized in **Table 5.12**.

Table 5.12 - Comparing the results with different machine learning algorithms – Scenario 1.

Model	R^2 Statistics	
	Arousal	Valence
SLR	39.7%	14.5%
MLR	55.6%	29.6%
SVR	55.4%	29.8%
K-NNR	43.0%	21.4%
GPR	53.8%	33.5%

When using the same features as Yang et al. [31], the results show the predominance of 3 algorithms being them SVR, MLR, GPR. This occurs due to the prediction methods performed by SLR, K-NNR. The results were very similar to those found on the literature [31] that states 57.0% for arousal and 22.2% for valence in terms of R^2 statistics.

SLR models the relation between the values assuming that the distribution is linear, but considering the target values to predict and their wide disposal on the plane, the algorithm presents a considerable error.

The K-NNR regression presents such results because this algorithm is severely dependent on other valence and arousal values presented on this dataset. One major disadvantage of K-NN regression is the need to have a large number of training data for the prediction to be more accurate. This also results in the need of a larger amount of computations.

The differentiation between SLR and MLR also occurs due to the difference of explanatory values. Moreover the results present a gap between the valence and the arousal predicted values. In fact the arousal reach a 55.6% in terms of R^2 statistics, opposing to the 33.5% on valence. This values were expect, since it was reported identical results in many previous works regarding MER systems [26] [103]. It is known that normally the arousal value is easier to predict when compared with valence. As Yang et al. stated there are 2 fundamental motives for this to occur. One concerns the number of relevant features to arousal such as loudness and pitch. While there are few predominant features associated with valence. The second reason concerns the consistency of the human annotations. Mainly because of the perception of valence is widely subjective. It is normal for 2 different people perceive opposite values regarding the same excerpt. Regarding the variance that occurs in every 20 testing experiences and the approximated result values using SVR, MLR and GPR it is impossible to state which one has the best performance.

As observed on the **Table 5.12** the GPR and SVR have very similar results. This happens due to the many common characteristics. Both models are non-parametric, they are kernel base, and their implementation and usage is identical. The results obtained with GPR were similar to those found in the literature, mainly in the work developed by Markov et al. [72]. On their work, it is stated that the best results were 69.2% for arousal and 47.3% for valence in terms of R^2 statistics, but in this case it was used a total of 388 different musical feature dimensions and different kernel combinations.

Concerning the second scenario and the results obtain in the subsection Musical Features Selection in terms of arousal the best features are dynamics, timbre and pitch. Although, the results using the Pitch feature are significantly below the other two. The dynamics features

are intrinsically related to the energy of the sound. This is directly related to the arousal perception [15]. The pitch and timbre influence the arousal perception by a listener [26].

In terms of valence the features that present better results are dynamics, timbre and rhythm. This happened for all machine learning algorithms. As stated in subsection 2.2.1. - Musical Features, rhythm and timbre affect deeply the emotion perceived. When using Tonality feature the results regarding the valence were very low. This was not expected due to, as stated by Gabrielsson et al. [15] the musical mode and harmony has a direct correlation with valence.

Table 5.13 - Results with features on scenario 1 plus rhythm.

Model	R^2 Statistics	
	Arousal	Valence
SLR	11.3%	27.1%
MLR	8.8%	31.9%
SVR	63.3%	51.5%
K-NNR	47.5%	27.8%
GPR	47.3%	37.0%

Considering the results obtained on the valence space when using the rhythm features it was decided to implement those in the scenario 1 and compare the achieved results. The R^2 statistics are presented in **Table 5.13**. Observing these results it is clear that all the valence predictions were incremented significantly. On the other hand, the arousal values were all reduced. This happens mainly on the SLR and MLR. This could be explained that, as stated before, the increasing features values not always indicate the best results. These algorithms are strictly calculated on linear predictors that use explanatory variables. The inclusion of more features may have an impact the algorithm sometimes. Also, this happens due to statistical restrictions imposed by linear models. Selecting the optimal amount of features is more challenging in these cases.

Table 5.14 - Results with features on scenario 1 plus rhythm and MFCCs.

Model	R^2 Statistics	
	Arousal	Valence
SLR	37.8%	14.1%
MLR	53.1 %	28.2 %
SVR	57.7%	55.6%
K-NNR	42.8%	22.6%
GPR	49.5%	31.4%

Regarding the importance of the MFCCs on MER systems [2, 13, 30, 31] consider to be one of the main Timbre feature that influence the arousal perception it was implemented on this system. The results presented on **Table 5.14**. are very similar to the previous test except in what concerns MLR and SLR predictions. Although the valence values were significantly reduced except on SVR, confirming the robustness of the SVR algorithm.

It was test on this system the use of all the features available. The obtained results are explicit on **Table 5.15**. The results are quite identical to the previous scenario mainly due to the fact that only more 4 features were inserted on the system.

Table 5.15 - Results with all implemented features.

Model	R^2 Statistics	
	Arousal	Valence
SLR	45.7%	13.2%
MLR	57.9%	27.1%
SVR	55,7%	31.2%
K-NNR	44.2%	24.4%
GPR	45.7%	30.9%

Considering the above table, the worst results are found in the case of SLR, K-NRR and GPR; the SLR mainly due to its simplicity; the K-NRR and GPR case, due to the fact that they are lazy classifiers, being more useful with fewer attributes on the database.

Overall, considering the results, the algorithm that presents more consistency throughout the evaluation tests is the SVR. This happens mainly due to the optimization algorithm implemented within. Besides this, it is a non-linear algorithm capable to adapt to different scenarios giving the conditions to resist to any kind of error. Another advantage is the possibility of parameter learning from the training data. On the other hand, SVM provide sparse solution, i.e. only support vectors are used for the inference, which can be a plus when working with large amount of data.

The features that influence more the results were, without doubt the dynamics and timbre. This features influence both valance and arousal. Regarding arousal, pitch has a greater influence, in accordance with the literature [26]. Rhythm has a relation with the predicted valence values. Considering all this, the validation of the system will be presented on the next section.

5.3 - Validation to a Recommendation System

Choosing the machine learning algorithm creates the basic conditions to map an emotion within a musical piece. Regardless of any associated error and considering the efficiency presented in the previous section, the algorithm SVR can be considered sufficiently accurate to be implemented in an automatic recommendation system.

Since one of the aims of this project is to create a recommendation system that is user-friendly it is important to consider the speed of the proposed system, when loading a new song. The main problem is that it takes the system a great amount of time to extract the features within a musical excerpt. And since the proposed extracts in every 0.5 second the loop is considerable. Considering this fact, and giving use to the results presented in the previous section. The aim was to reduce the used features to the most important ones, and validating this result considering the previous tests performed on the system. The results are presented in **Table 5.16**.

Table 5.16 – Results with reduced implemented features.

Model	R^2 Statistics	
	Arousal	Valence
SVR	57.7%	31.3%
SVR [31]	57.0%	22.2%

As observed in the table the results are very similar to those obtained in previous tests. Furthermore, the results in the valence value are higher to those presented by Yang et al. [31]. The features validated concern those who had more influence on the MER system results, being them, dynamics, timbre, rhythm and pitch. This features are presented in **Table 5.17.** Discarding features related to the tonality and features such as MFCCs that proved its irrelevance to this system main goal.

Table 5.17 - Features used on the system

Features	Category
RMS	Dynamics
Low-Energy	Dynamics
Event Density	Rhythm
Tempo	Rhythm
Pulse Clarity	Rhythm
Spectral roll-off point	Timbre
Brightness	Timbre
Signal statistics	Timbre
Irregularity	Timbre
fo	Pitch
Roll-off	Pitch

Chapter 6

Conclusions

This chapter presents the main contributions and considerations about this project. Furthermore, there are presented possible improvements and future work on the MER system.

5.4 - Contributions

The aim of the Music Emotion Recognition system was to automatically predict the emotional state associated with a musical piece. This goal was achieved and the results of this project show that this function is effective.

By comparing different models of regression this work contributes to the study of an area that has gained great attention in recent years. One of the findings of this project were that among the tested machine learning algorithms, the most accurate was the Support Vector Regression.

The contribution to the field resides on the creation of an automatic system capable of associating an emotion to a song, represented on the valence arousal space. This is possible due to the use of Support Vector Regression algorithm, this regression model was trained to predict the emotion values that represent the emotional content of a musical piece. This content is represented as a point in the emotional space. So, after the training phase the system it is capable of performing predictions.

Possibly the most relevant main finding for this work was the relevance of the features when training a specific regression algorithm. The features that were found to be more determinant were dynamics, timbre, rhythm and pitch.

Furthermore, the system described in this dissertation creates the main conditions to the development of an automatic recommendation system that is based on emotion.

Giving computers the same capabilities as a human to recognize an emotion within a sound has always been a main goal to music information retrieval, due to the possibilities that emerge. People's criteria for selecting a song are often related to their emotional state. This work approximates these goals and contributes to the innovation of MER systems, giving a good perspective towards what it can become.

Moreover, implementing a machine learning modules within the proposed system does not create all the conditions to serve the goal of a usable recommendation system. Specifically

due to the speed of this system regarding the features extraction. This constitutes a barrier to its usability.

As the final conclusion it is fundamental to underline the implementation and evaluation of different algorithms such as K-Nearest Neighbour Regression and Simple Linear Regression as well as its evaluation on the system. This project confirmed Support Vector Regression as the current state-of-the-art model for MER tasks.

5.5 - Future Work

The main recommendations that comes directly from the results obtained in this project is that a deeper training, meaning the use of an extensive annotated database would certainly have an impact on the accuracy of the system predictions. Training the proposed machine learning model with more reliable datasets would increase its validity allowing it to serve more efficiently its purpose as a recommendation system.

The first improvement should be considering the regression as a classification problem. A possible approach would be dividing the emotion space into main areas, thus increasing the reliability of the MER system and reducing the error. This happens mainly due to the fact that the prediction would be an area on the emotional space and not an actual numerical value.

Focusing on the MER system developed, the MATLAB application could be improved in terms of the GUI. Improving its functionalities, specifications and being more user-friendly.

While working on this project and its development, different approaches to this problem emerged. Concerning the aim of its usability, its implementation in a different language or even software could spread its use. This happens mainly due to the fact that MATLAB is not used by every person.

An important aspect to be improved is the speed of the feature extraction, when thinking of large collections found on every person computer or even smartphone, it seems unlikely that this system could handle such amount of data. Decreasing the time needed to do this procedure is a paramount goal concerning the system as a capable recommendation system to be used by the community. An ideal approach would be to implement a feature selection algorithm to identify good features.

Extending the goals of this project, instead of automatically predicting a specific emotional value representing an all song, implement the temporal variation of emotions within an audio. This way the user could see the temporal evolution and associated emotion within a musical piece.

In addition, as an extension of this work, the automatic music recommendation system based on emotion could even serve commercial goals as an android application or iOS.

References

[1] T. Eerola, "Modeling listeners' emotional response to music," *Topics in cognitive science,* vol. 4, pp. 607-624, 2012.

[2] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, *et al.*, "Music emotion recognition: A state of the art review," 2010, pp. 255-266.

[3] Y. E. Kim, E. M. Schmidt, and L. Emelle, "MoodSwings: A Collaborative Game for Music Mood Label Collection," 2008, pp. 231-236.

[4] E. Law and L. Von Ahn, "Input-agreement: a new mechanism for collecting data using human computation games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 1197-1206.

[5] P. Saari, G. Fazekas, T. Eerola, M. Barthet, O. Lartillot, and M. Sandler, "Genre-adaptive Semantic Computing and Audio-based Modelling for Music Mood Annotation," 2015.

[6] W. F. Thompson and L. Quinto, "Music and emotion: psychological considerations," *The Aesthetic Mind: Philosophy and Psychology,* pp. 357-75, 2011.

[7] G. A. Wiggins, "A preliminary framework for description, analysis and comparison of creative systems," *Knowledge-Based Systems,* vol. 19, pp. 449-458, 2006.

[8] A. J. Blood and R. J. Zatorre, "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 11818-11823, 2001.

[9] P. N. Juslin and J. A. Sloboda, *Music and emotion: Theory and research*: Oxford University Press, 2001.

[10] V. J. Konecni, "The influence of affect on music choice," in *Handbook of Music and Emotion: Theory, Research, Applications*, P. N. Juslin and J. Sloboda, Eds., ed: Oup Oxford, 2011.

[11] K. R. Scherer, "Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them?," *Journal of new music research,* vol. 33, pp. 239-251, 2004.

[12] K. Monteith, T. Martinez, and D. Ventura, "Automatic generation of music for inducing emotive response," 2010, pp. 140-149.

[13] M. Caetano and F. Wiering, "Theoretical Framework of A Computational Model of Auditory Memory for Music Emotion Recognition," 2014, pp. 331-336.

[14] M. Caetano, A. Mouchtaris, and F. Wiering, "The Role of Time in Music Emotion Recognition: Modeling Musical Emotions from Time-Varying Music Features," in *From Sounds to Music and Emotions*. vol. 7900, M. Aramaki, M. Barthet, R. Kronland-Martinet, and S. Ystad, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 171-196.

[15] A. Gabrielsson and E. Lindström, "The role of structure in the musical expression of emotions," *Handbook of music and emotion: Theory, research, applications,* pp. 367-400, 2010.

[16] P. Evans and E. Schubert, "Relationships between expressed and felt emotions in music," *Musicae Scientiae,* vol. 12, pp. 75-99, 2008.

[17] S. Vieillard, I. Peretz, N. Gosselin, S. Khalfa, L. Gagnon, and B. Bouchard, "Happy, sad, scary and peaceful musical excerpts for research on emotions," *Cognition & Emotion,* vol. 22, pp. 720-752, 2008.

[18] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: characterization, classification, and measurement," *Emotion,* vol. 8, p. 494, 2008.
[19] C. Laurier, *Automatic Classification of Musical Mood by Content Based Analysis*: Universitat Pompeu Fabra, 2011.
[20] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and brain sciences,* vol. 31, pp. 559-575, 2008.
[21] R. Panda and R. P. Paiva, "Using support vector machines for automatic mood tracking in audio music," 2011.
[22] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music,* 2010.
[23] C. McKay, "Emotion and music: Inherent responses and the importance of empirical cross-cultural research," *Course Paper. McGill University,* 2002.
[24] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, *et al.*, "Universal recognition of three basic emotions in music," *Current biology,* vol. 19, pp. 573-576, 2009.
[25] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation," 2011, pp. 549-554.
[26] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 3, p. 40, 2012.
[27] K. Hevner, "Experimental studies of the elements of expression in music," *The American Journal of Psychology,* vol. 48, pp. 246-268, 1936.
[28] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 5-18, 2006.
[29] J. Skowronek, M. F. McKinney, and S. Van De Par, "A Demonstrator for Automatic Music Mood Estimation," 2007, pp. 345-346.
[30] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *Selected Topics in Signal Processing, IEEE Journal of,* vol. 5, pp. 1088-1110, 2011.
[31] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 448-457, 2008.
[32] B. L. Sturm, "On music genre classification via compressive sampling," 2013, pp. 1-6.
[33] A. Huq, J. P. Bello, and R. Rowe, "Automated music emotion recognition: A systematic evaluation," *Journal of New Music Research,* vol. 39, pp. 227-244, 2010.
[34] L. Mion and G. D. Poli, "Score-independent audio features for description of music expression," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 458-466, 2008.
[35] E. Schubert, "Analysis of emotional dimensions in music using time series techniques," *Context,* vol. 31, p. 65, 2006.
[36] M. Zentner and T. Eerola, "Self-report measures and models," *Handbook of music and emotion,* pp. 187-221, 2010.
[37] K. P. Monteith, "Automatic Generation of Music for Inducing Emotive and Physiological Responses," 2012.
[38] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models," in *ISMIR*, 2009, pp. 621-626.
[39] S. O. C.-C. H. MacDorman, Karl F, "Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison," *Journal of New Music Research,* vol. 36, pp. 281-299, 2007.
[40] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 267-274.
[41] K. K. Chang, J.-S. R. Jang, and C. S. Iliopoulos, "Music Genre Classification via Compressive Sampling," in *ISMIR*, 2010, pp. 387-392.
[42] B. Snyder, *Music and memory: An introduction*: MIT press, 2000.

[43] E. Coutinho and A. Cangelosi, "Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements," *Emotion,* vol. 11, p. 921, 2011.
[44] T. Eerola, "Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journal of New Music Research,* vol. 40, pp. 349-366, 2011.
[45] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology,* vol. 39, p. 1161, 1980.
[46] E. Schubert, "Update of the Hevner adjective checklist," *Perceptual and motor skills,* vol. 96, pp. 1117-1122, 2003.
[47] G. Peeters, "A generic training and classification system for MIREX08 classification tasks: Audio music mood, audio genre, audio artist and audio tag," 2008.
[48] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 762-774, 2011.
[49] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet, "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts," *Cognition & Emotion,* vol. 19, pp. 1113-1139, 2005.
[50] R. E. Thayer, *The biopsychology of mood and arousal*: Oxford University Press, 1989.
[51] D. Watson and L. A. Clark, "The PANAS-X: Manual for the positive and negative affect schedule-expanded form," 1999.
[52] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1802-1812, 2011.
[53] D. B. Huron, *Sweet anticipation: Music and the psychology of expectation*: MIT press, 2006.
[54] C. L. Krumhansl, "Music: A link between cognition and emotion," *Current directions in psychological science,* vol. 11, pp. 45-50, 2002.
[55] E. Schubert, "Modeling perceived emotion with continuous musical features," *Music Perception: An Interdisciplinary Journal,* vol. 21, pp. 561-585, 2004.
[56] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 467-476, 2008.
[57] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-Label Classification of Music into Emotions," in *ISMIR*, 2008, pp. 325-330.
[58] P. Lamere and O. Celma, "Music recommendation tutorial notes," *ISMIR Tutorial, September,* 2007.
[59] E. L. Law, L. Von Ahn, R. B. Dannenberg, and M. Crawford, "TagATune: A Game for Music and Sound Annotation," in *ISMIR*, 2007, p. 2.
[60] M. I. Mandel and D. P. Ellis, "A web-based game for collecting music metadata," *Journal of New Music Research,* vol. 37, pp. 151-165, 2008.
[61] D. Turnbull, R. Liu, L. Barrington, and G. R. Lanckriet, "A Game-Based Approach for Collecting Semantic Annotations of Music," in *ISMIR*, 2007, pp. 535-538.
[62] L. Barrington, D. O'Malley, D. Turnbull, and G. Lanckriet, "User-centered design of a social game to tag music," in *Proceedings of the acm sigkdd workshop on human computation,* 2009, pp. 7-10.
[63] T. Li and M. Ogihara, "Detecting emotion in music," 2003, pp. 239-240.
[64] M. I. Mandel, G. E. Poliner, and D. P. Ellis, "Support vector machine active learning for music retrieval," *Multimedia systems,* vol. 12, pp. 3-13, 2006.
[65] X. Downie, C. Laurier, and M. Ehmann, "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, 2008, pp. 462-467.
[66] G. Tzanetakis, "Marsyas submissions to MIREX 2009," *Music Information Retrieval Evaluation eXchange (MIREX),* 2009.
[67] C. Cao and M. Li, "Thinkit's submissions for MIREX2009 audio music classification and similarity tasks," 2009.
[68] B.-j. Han, S. Ho, R. B. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression," 2009.

[69] Y. Yang, Y. Lin, H. Cheng, I. Liao, Y. Ho, and H. Chen, "Advances in Multimedia Information Processing-PCM 2008, ser," *Lecture Notes in Computer Science. Springer Berlin/Heidelberg,* pp. 70-79, 2008.
[70] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing,* vol. 14, pp. 199-222, 2004.
[71] G. Tzanetakis, "Marsyas submissions to MIREX 2012," ed: Citeseer, 2012.
[72] K. Markov, M. Iwata, and T. Matsui, "Music Emotion Recognition using Gaussian Processes," in *MediaEval*, 2013.
[73] S.-H. Chen, Y.-S. Lee, W.-C. Hsieh, and J.-C. Wang, "Music emotion recognition using deep Gaussian process," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 495-498.
[74] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing,* vol. 10, pp. 293-302, 2002.
[75] H. Deshpande, R. Singh, and U. Nam, "Classification of music signals in the visual domain," in *Proceedings of the COST-G6 Conference on Digital Audio Effects*, 2001, pp. 1-4.
[76] M. Pesek, P. Godec, M. Poredos, G. Strle, J. Guna, E. Stojmenova*, et al.*, "Introducing a Dataset of Emotional and Color Responses to Music," in *ISMIR*, 2014, pp. 355-360.
[77] J. P. d. S. Figueiredo, "Music Reccomendation System Based on Emotions," University of Porto, 2015.
[78] J.-J. Aucouturier and F. Pachet, "Finding songs that sound the same," 2002, pp. 1-8.
[79] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 Songs Database," ed, 2014.
[80] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011. hal-00727791, version 2-22 Oct 2012*, 2011.
[81] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software," in *ISMIR*, 2010, pp. 441-446.
[82] C. Cannam, M. Sandler, M. O. Jewell, C. Rhodes, and M. d'Inverno, "Linked data and you: Bringing music research software into the semantic web," *Journal of New Music Research,* vol. 39, pp. 313-325, 2010.
[83] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *The Journal of Machine Learning Research,* vol. 6, pp. 1889-1918, 2005.
[84] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.
[85] O. Lartillot, P. Toiviainen, and T. Eerola, "MIRtoolbox v1. 5," ed: Finnish C. of Exc. in Interdiscipl. Music Res, 2013.
[86] C. McKay, "Automatic music classification with jMIR," Citeseer, 2010.
[87] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," 2013, pp. 1-6.
[88] F. Alías, J. C. Socoró, and X. Sevillano, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds," *Applied Sciences,* vol. 6, p. 143, 2016.
[89] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 375-376.
[90] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimization," in *ISMIR*, 2008, pp. 521-526.
[91] T. as a Multidimensional, "The Perception of Musical Timbre," *The Oxford Handbook of Music Psychology,* p. 113, 2016.
[92] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organised sound,* vol. 4, pp. 169-175, 2000.
[93] P. Knees, T. Pohle, M. Schedl, and G. Widmer, "A music search engine built upon audio-based and web-based similarity measures," in *Proceedings of the 30th annual*

international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 447-454.

[94] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cognition & Emotion,* vol. 19, pp. 633-653, 2005.

[95] K. Jensen, *Timbre models of musical sounds*: Department of Computer Science, University of Copenhagen, 1999.

[96] J. Krimphoff, S. McAdams, and S. Winsberg, "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique," *Le Journal de Physique IV,* vol. 4, pp. C5-625-C5-628, 1994.

[97] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *The Journal of the Acoustical Society of America,* vol. 100, pp. 3491-3502, 1996.

[98] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 847-855.

[99] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the SMO algorithm for SVM regression," *IEEE transactions on neural networks,* vol. 11, pp. 1188-1193, 2000.

[100] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning,* vol. 6, pp. 37-66, 1991.

[101] D. J. MacKay, "Introduction to Gaussian processes," *NATO ASI Series F Computer and Systems Sciences,* vol. 168, pp. 133-166, 1998.

[102] Y.-H. Yang, Chia-Chu Liu, and Homer H. Chen. , "Music emotion classification: a fuzzy approach," *Proceedings of the 14th ACM international conference on Multimedia.,* 2006.

[103] E. Schubert, "Measurement and time series analysis of emotion in music," 1999.