POLITECNICO
MILANO 1863

Dipartimento di Elettronica, Informazione e Bioingegneria

Master Degree in Computer Science and Engineering

Thesis Title

and its subtitle

by:
Gioele Pozzi

matr.:
10454628

Supervisor:

Co-supervisor:
Clara Borrelli

Academic Year
2019-2020

POLITECNICO
MILANO 1863

Dipartimento di Elettronica, Informazione e Bioingegneria

Master Degree in Computer Science and Engineering

Titolo Tesi

e sottotitolo

Candidato:
Gioele Pozzi

matricola:
10454628

Relatore:

Co-relatore:
Clara Borrelli

Anno Accademico
2019-2020

Abstract

One of the most attractive functions of music is that it can convey emotion and modulate a listener's mood [1]. Music can bring to tears, console us when we are grieving and drive us to love.
Most important thing is that music information behavior studies have identified emotion as an important criterion used by people in music searching and organization. Now become important the field of music emotion recognition.

Sommario

Piacere, so Mario

Acknowledgements

This thesis is the result of almost a year of work at the Image and Sound Processing Lab. First I would thank my supervisor...

Thanks to friends.

Thanks to family.

N.S.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Music has an important role in human life. More important, is that music in capable to evoke different emotions for people, but how is structured the relationship between music and emotion? We don't know yet. It's a hard problem, which have very different fields of background, from computer science, machine learning and psychology.

Emotion-aware Music Information Retrieval has been difficult due to the subjectivity and temporal of emotion responses to music. The role of physiological signals related to emotions could potentially be exploited in emotion-aware music discovery.

Music is the vehicle for emotions, feelings, passion and actions. With the music the composer create a narration which is purely emotional.

Can we measure emotions related to music?

1.2 Outline of the thesis

This thesis is organized as follows:

After a brief introduction about the objective of the thesis, in chapter 2 and 3 is presented a complete overview about the main arguments in chapter 2, as Music Information Retrieval (MIR) and Music Emotion Recognition (MER), Electrodermal Activity (EDA) and other physiological data using on-body sensors.

Chapter 4 is devoted to a complete overview of the state of the art about the main aspects related to chapters 2 and 3 of this thesis, in order

to have a general idea about what has been done in the past and which results they have achieved.

In chapter 5 is presented how the dataset we have considered is structured and what results they have reached. Is also shown our implementation of the problem.

Chapter 6 is about the results we have achieved and the comparison between the PMEmo performances.

Finally Chapter 7, draws the conclusions and outlines possible future research directions.

1.3 Application fields

The work proposed in this thesis finds potential application in several fields. Thanks to the work of PMEmo that created a large dataset containing emotion annotations and electrodermal activity signal, we have the possibility to study the relationship between music emotion and physiological signals.
Music Browsing can be an important field of application, because it helps in general in finding, generally in large datasets, what music user are looking for. For example one application could be to create a playlist based on the emotion that songs produce in each of us. Another important application is given by understanding the relationship between music and emotion, which is a well known relationship but hard to find structural connection between the two.

2

Theoretical Background on MIR and MER

This chapter introduces the readers to the main basics about Music Information Retrieval and Music Emotion Recognition.

2.1 Music Information Retrieval

Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music. MIR is a small but growing field of research with many real-world applications. Those involved in MIR may have a background in musicology, psychoacoustics, psychology, academic music study, signal processing, informatics, machine learning, optical music recognition, computational intelligence or some combination of these.

MIR is being used by businesses and academics to categorize, manipulate and even create music.
A few application to MIR can be:

- Recommended systems: several already exist, but few are based upon MIR techniques, instead making use of similarity between users or laborious data compilation as in Pandora[1].
- Intelligent and adaptive digital audio effects: aim of design a system that determine the settings of audio effects based on the audio content.

[1] https://www.pandora.com

- Track separation and instrument recognition: like extracting the original tracks as recorded, which could have more than one instrument played per track. Instrument recognition is about identifying the instruments involved into one track.
- Automatic music transcription: process of converting an audio recording into symbolic, such score or a MIDI file.
- Automatic categorization: common task of MIR is musical genre categorization and is the usual task for the yearly Music Information Retrieval Evaluation eXchange (MIREX).

2.2 Music Emotion Recognition

Music Emotion Recognition (MER) aim to research on modeling humans emotion perception of music [4], a research topic that emerges in the face of the explosive growth of digital music. Automatic MER allows users to retrive and organize their music collections in a fashion that is more content-centric than conventional metadata-based methods.
The main challenge is based on the human perception of emotions, their subjective nature of emotion perception. Building such a music emotion recognition system, however, is challenging be- cause of the subjective nature of emotion perception. One needs to deal with issues such as the reliability of ground truth data and the difficulty in evaluating the prediction result, which do not exist in other pattern recognition problems such as face recognition and speech recognition.
MER methods developed try to address the issues related to the ambiguity and granularity of emotion description, the heavy cognitive load of emotion annotation, subjectivity of emotion perception, and the semantic gap between low-level audio signal and high-level emotion perception.

2.2.1 Importance of Music Emotion Recognition

Music plays an important role in human life, even more in the digital age. Never before has such a large collection of music been created and accessed daily by people. Before with the use of compact audio formats with near CD quality such as MP3 and now on with the various streaming services, have greatly contributed to the tremendous growth of digital music libraries.

Conventionally, the management of music collections is based on catalog metadata, such as artist name, album name, and song title. As the amount of content continues to explode, this conventional approach may be no longer sufficient. The way that music information is organized and retrieved has to evolve to meet the ever increasing demand for easy and effective information access.

Music, is a complex acoustic and temporal structure, it is rich in

content and expressivity. When an individual engages with music as a composer, performer or listener, a very board range of mental processes is involved, including *representational* and *evaluative*. The representational process includes the perception of meter, rhythm, tonality, harmony, melody, form, and style, whereas the evaluative process includes the perception of preference, aesthetic experience, mood, and emotion. The term evaluative is used because such processes are typically both valences and subjective. Both the representational and the evaluative processes of music listening can be leveraged to enhance music retrieval. According to a study of Last.fm[2], emotion tagging is the third most frequent type of tags (first is genre and second locale) assigned to music pieces by online users.
Even if emotion-based music retrieval was a new idea, a survey conducted in 2004 from [5] showed that about 28.2% of the participants identified emotion as an important criterion in music seeking and organization.
The table 2.1 represent the responses of 427 subjects to the question *"When you search for music or music information, how likely are you to use the following search/browse options?"* [5].

Search/Browse by	Positive rate
Singer/Performer	96.2%
Title of work(s)	91.6%
Some words of the lyrics	74.0%
Music style/genre	62.7%
Reccomendations	62.2%
Similar artist(s)	59.3%
Similar music	54.2%
Associated usage	41.9%
Singing	34.8%
Theme(main subject)	33.4%
Popularity	31.0%
Mood/emotional state	**28.2%**
Time period	23.8%
Occasions to use	23.6%
Instrument(s)	20.8%
Place/event where heard	20.7%
Storyline of music	17.9%
Tempo	14.2%
Record label	11.7%
Publisher	6.0%

Table 2.1: Responses of 427 subjects to the question *"When you search for music or music information, how likely are you to use the following search/browse options?"*

[2]https://www.last.fm/home

Into another survey [6], they present findings from an exploratory questionnaire study featuring 141 music listeners (between 17 and 74 years of age) that offers some novel insights.
One of the most exciting but difficult endeavors in research on music is to understand how listeners respond to music. It has often been suggested that a great deal of the attraction of music comes from its "emotional powers". That is, people tend to value music because it expresses and induces emotions. The table 2.2 tries to resume the motivations to the answer *"Why do we listen to music?"*

Motive	Ratio
"To express, release and influence emotions"	47%
"To relax and settle down"	33%
"For enjoyment, fun, and pleasure"	22%
"As company and background sound"	16%
"Because it makes me feel good"	13%
"Because it's a basic need, I can't live without it"	12%
"Because I like, love music"	11%
"To get energized"	9%
"To evoke memories"	4%

Table 2.2: Responses of 141 subjects to the question *"Why do you listen to music?"*

Some music companies, like Allmusic.com[3], gives the possibility to search music by emotion labels. With these, the user can retrive and browse artists or albums by emotion.

Making computers capable of recognizing the emotion of music also enhances the way humans and computers interact. It is possible to play back music that matches the users mood detected from physiological, prosodic, or facial cues. A cellular phone equipped with automatic music emotion recognition (MER) function can then play a song best suited to the emotional state of the user; a smart space (e.g., restaurant, conference room, residence) can play background music best suited the people inside it.

2.2.2 Recognizing the perceived emotion of music

There is a relationship between music and emotions, that has been the subject of much discussion and research in many different disciplines, like philosophy, musicology, sociology.
In psychological studies, emotion are often divided into three categories:

- *Expressed emotion*: the ones the performer tries to communicate with the listener.

[3]https://www.allmusic.com/moods

- ***Perceived emotion***: represented by music and perceived by the listener.
- *Felt or Evoked emotion*: induced by music and felt by the listener.

MER focus on perceived emotions because they are less subjective than felt emotions and are often easier to conceptualize. This because felt emotions depends on personal factors and the situation in which the listener processes the song. From an engineering point of view, one of the main interests is to develop a computational model of music emotion and to facilitate emotion-based music retrieval and organization. MIR community has made many efforts for automatic recognition of the perceived emotion of music, various implementations will be presented further in chapter 4.
A typical approach to MER categorizes emotions into a number of classes and applies Machine Learning (ML) techniques to train a classifies. Usually are extracted some features of music to represent the acoustic property of a music piece. Typically, a subjective test is conducted to collect the ground truth needed for training the computational model of emotion prediction. Subjects are asked to report their emotion perceptions of the music pieces.
To learn the relationship between music features and emotion labels have been applied, such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), Neural Networks (NN) and k-nearest neighbor.
After training, the automatic model can be applied to classify the emotion of an input music piece, for example a schematic diagram of the *categorical approach* to MER can be seen in figure 2.1.

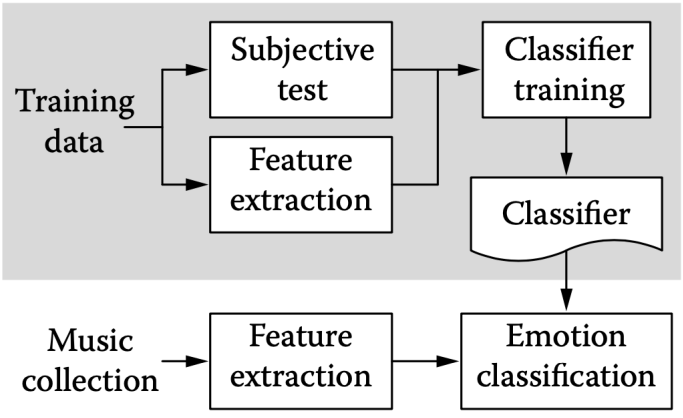

Figure 2.1: Schematic diagram of the categorical approach to MER

2.2.3 Open issues of Music Emotion Recognition

As MER is a quite new domain, there are some elements that have no clear answer. Four of these issues are:

1. Ambiguity and Granularity of emotion description: issue related to the relationship between emotions and the affective terms that denote emotions and the problem of choosing which and how many affective terms to be included in the taxonomy. Emotions are fuzzy concepts, there are main synonyms and similarities between different terms. In general, classification accuracy of an automatic model is inversely proportional to the number of classes considered [7].

2. Heavy cognitive load of emotion annotation: to collect data for training an automatic model, is typically conducted a subjective test by inviting human subjects to annotate the emotion of music pieces. The problem is that to reduce administrative effort, each music piece is annotated by two or three musical *experts* to gain consensus of the annotation result. Everyday contexts in which musical experts experience is so different from those non-experts require separate treatment. Since MER system is expected to be used in the everyday context, the emotion annotation should be carried out by *ordinary people*.

3. Subjectivity of emotional perception: music perception is intrinsically subjective and is under the influence of many factors such as cultural background, age, gender, personality and so forth. Therefore conventional categorical approaches that simply assign one emotion class to each music piece in a deterministic manner do not perform very well in practice.

4. Semantic gap between Low-Level (LL) and audio signal and High Level (HL) Human perception: it is difficult to accurately compute emotion values, and what intrinsic element of music causes a listener to create a specific emotional perception is still far from well understood.

2.2.4 Emotion description

Many researchers have suggested that music is an excellent medium for studying emotion, because people tend to make judgments about music and their affective responses to music.
Music represent emotions that are perceived by the listener or induced emotions that are felt by the listener. Now we will focus on the emotion conceptualization alone, since it's central to have a theoretical background to apply then to MER.
The celebrated paper of Hevner [8] , studied the relationship between music and emotions though experiments where subjects are asked to report some adjectives that came to their mind as the most representative part of a music played. From this have been proposed a large variety of emotion models, like the one presented and used in this thesis.

The idea of emotion conceptualization is to divide in two different approaches, the **Categorical approach** and the **Dimensional approach**.

Categorical approach

The first assumption of this emotion conceptualization is that emotions are categorized and categories are distinct from each other. For this approach, there is the idea that there are a limited number of innate and universal emotion categories such as:

- Happiness
- Sadness
- Anger
- Fear
- Disgust
- Surprise

All other emotions can be derived from these "*basic emotions*".
In psychological studies, different researchers have come up with different sets of basic emotions.
For example, another famous categorical approach to emotion conceptualization is Hevner's adjective checklist. He found eight clusters positioned in circle as in figure 2.2. The adjective within a cluster are similar, neighbor clusters varies in a cumulative way until reaching the opposite position where there is the contrast cluster. Hevner's checklist

Figure 2.2: Eight clusters proposed by Hevner

proposed in 1935 was suddenly updated and regrouped into ten groups by Fansworth and into nine groups in 2003 by Schubert.

Drawbacks of categorical approach is that the number of primary emotion classes is very small in comparison with the richness of music

emotion perceived by humans. The problem is in the sense that using a finer granularity, does not necessarily solve the problem because the language for describing emotions is inherently ambiguous and varies from person to person. Using a large number of emotion classes could submerge the subject and is impractical for psychological studies falsing results.

Dimensional approach

Categorical approach focuses mainly on the characteristics that distinguish emotions from one another, dimensional approach focuses on identifying emotions based on their position on a small number of emotion "dimensions" called axes, intended to correspond to internal human representation of emotion. These internal emotion dimensions are found by analyzing the correlation between affective terms.

There are several different names from past researchers gave very similar interpretations of the resulting factors like tension/energy, intensity/-softness, tension/relaxation for example. Most of the factors correspond to the two dimensions of emotion the *valence* (positive and negative affective states) and *arousal* (energy and stimulation level).
Russel, proposed a circumplex model of emotion in [9] which consist in a two-dimensional, circular structure as in figure 2.3 involving the dimensions of valence and arousal. In this structure, emotions that are inversely correlated, are placed across the circle from one another.

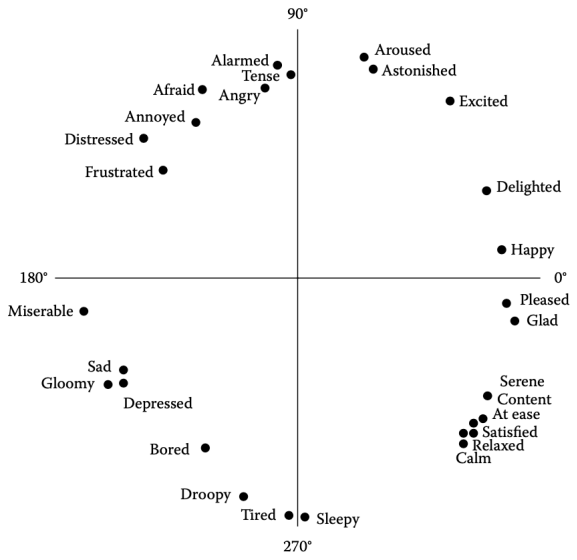

Figure 2.3: Russel's circumplex model of affect

Emotions that are easy to be confused, such as calm and sadness, appear to have similar valence and arousal values. This result implies that valence and arousal may be the most fundamental and most clearly communicated emotion dimensions among others. Also dimensional approach have its throwbacks, it is argued that dimensional approach blurs

important psychological distinctions and consequently obscure important aspects of the emotion process. One example in support of this argumentation is that anger and fear are placed close in the valence-arousal plane but they have very different implications for the organism. Also, it has been argued that using only a few emotion dimension cannot describe all the emotions without residuum.

Some researches, to overcome to these problems, tries to add a third dimension, called *potency* as dominant/submissive, to obtain a more complete picture of emotion. However, this would increase the cognitive load on the subjects at the same time, requires a more complex interface and makes hard to annotate the process. The third dimension problem is still in discussion.

Music Emotion Variation Detection

An important aspect that is not addressed in the previous two paragraphs is the temporal dynamics. Most researches has focused on music piece that are homogeneous with respect to the emotional plane. However, music can change its emotional expression during the song, becomes important to investigate the time-varying relationship between music and emotion. Here is more useful the dimensional approach to capture the continuous changes of emotional expression. Usually subjects are asked to rate valence and arousal in response of the stimulus every second.
For example, songs can be described by valence and arousal curves as in the following figure:

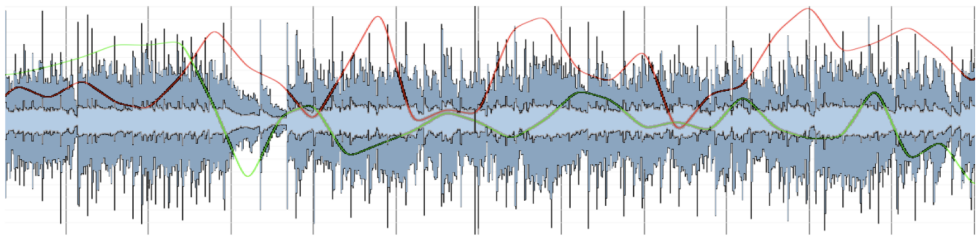

Figure 2.4: Valence and arousal curves for MEVD

2.2.5 Emotion recognition

MIR researches have been made to automate MER tasks, and the type of music under study has gradually shifted over the past few years from symbolic music to raw audio signal, from Western classical music to popular music. The purpose of MER is to facilitate music retrieval and management in the everyday music listening.

Nowdays are applied several machine learning techniques to recognize emotion from the music, and the training and automatic recognition model typically consists of the following steps:

1. Extract a certain number of features from audio signals to represent the music signal.

2. Collect from human annotators the ground truth emotion labels or emotion values.

3. Apply a learning algorithm between music features and emotion labels/values.

4. Predict emotion of an input song from the resulting computational model.

The music emotion recognition process can be schematized in the figure from [10]:

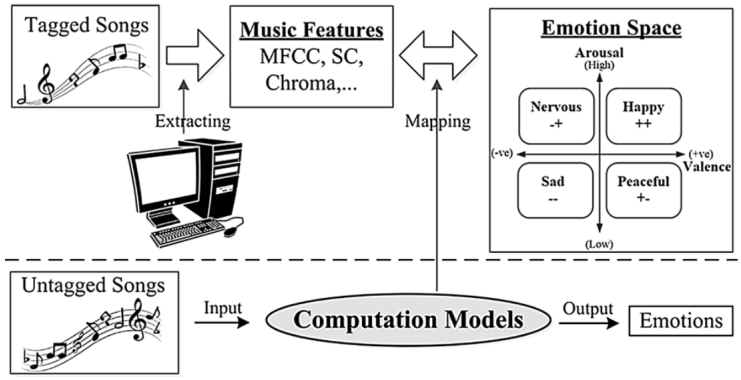

Figure 2.5: MER process

Researches that work on MER can be classified into three approaches.

The **categorical approach** that categorizes emotions into a number of discrete classes and applies machine learning techniques to train a classifier. The predicted emotion labels can be incorporated into a text-based or metadata-based music retrieval system.

The **dimensional approach** to MER defines emotions as numerical values over a number of emotion dimensions (valence and arousal). A regression model is trained to predict the emotion values that represent the affective content of a song, thereby representing the song as a point in an emotion space. Users can then organize, browse, and retrieve music pieces in the emotion space, which provides a simple means for user interface.

Categorical approach

Advantage of categorical approach is that it is easy to be incorporated into a text-based or metadata-based retrieval system. Emotion labels provide an atomic description of music that allows users to retrieve music through a few keywords. Here are present the issues discussed in chapter 2.2.3. The commonly adopted methods follows these points:

1. Data collection: nowadays there are several large-scale dataset covering all sort of music types and genres. Otherwise is desirable to collect data of the different types, getting rid of the effects called "*album effect*" or "*artist effect*" and collect a variety of music pieces. One problem is that there is no consensus on which emotion model or how many emotion categories should be used. Comparing systems that use different emotion categories and different dataset is impossible. However the issue concerning how many and which emotion classes should be used seem to remain open.

2. Data preprocessing: to compare music pieces fairly, music pieces are normally converted to a standard format, and since a complete music piece can contain sections with different emotions,a 20 to 30 second segment is often selected, which is representative of the song (like the chorus part). A good remark of the segment length can be found in [11].

3. Subjective test: emotion is a subjective matter, so the collection of the ground truth data should be conducted carefully. Annotation methods can be grouped into two categories:

   - Expert-based method: which employs a few musical experts to annotate emotions.
   - Subject-based method: employs a large number of untrained subjects to annotate emotions.

   The ground truth is set by averaging the opinion of all subjects (typically more than 10 subjects per song).
   It became important to not make a long test, in order to not compromise the reliability of the emotion annotations. Nowadays is introduced the use of listening games.

4. Features extraction: a certain number of features are extracted from the music signal to represent the different dimension of music listening like melody, timbre and rhythm.
   After features extraction, is applied feature normalization, in order to

5. Model training: the following step is to train a Machine Learning (ML) model to learn the relationship between emotion and music. Music emotion classification is carried out with classification ML algorithms, such as Neural Network, k-nearest neighbor (kNN), decision tree, Support Vector Machine (SVM) and Support Vector Classification (SVC).

Dimensional approach

The attractive part of dimensional approach is the valence-arousal

plane and the associated emotion-based retrieval methods. Due to the fact that the emotion plane contain an infinite number of emotion descriptions, the granularity and ambiguity issues are relieved.
Dimensional perspective is adopted to track the emotion variation of a classical song. The idea of representing the overall emotion of a popular song as a point in the emotion plane for music retrieval, under the assumption that the dominant emotion of a popular song undergoes less changes than a classical song. MER problem became a regression problem, and two independent models, called regressors, are trained to predict the valence-arousal values.
The dimensional approach requires the subjects to annotate the numerical valence-arousal values. This requirement impose an high cognitive load on the subjects.

Pros and cons of categorical and dimensional approach are schematized in the following table:

	Pros	Cons
Categorical	Intuitive Natural language Atomic description	Lack a unifying model Ambiguous Subjective Difficult to offer fine-grained differentiation
Dimensional	Focus on a few dimensions Good user interface	Less intuitive Semantic loss in projection Difficult to obtain ground truth

Table 2.3: Pros and cons of categorical and dimensional approaches

2.2.6 Valence and Arousal

As already mentioned before, the valence-arousal plane is the most used dimension plane to represent emotion.
In general, emotional experiences can be described by these two terms, *valence* (positive or negative affectivity) and *arousal* (calming or exciting). Some studies found that valence as well as intensity, is triggered by the amygdala, while the arousal by the reptilian brain.
The common framework for dealing with emotional experience is characterized in a two-dimensional space. Valence ranges from highly negative to highly positive, and arousal ranges from calming/soothing to exciting/agitating: High arousal emotional events are encoded better that non-arousing events. Instead of increasing overall attention to an event, an emotionally arousing stimulus decreased attentional resources available for information processing and focused attention only on the

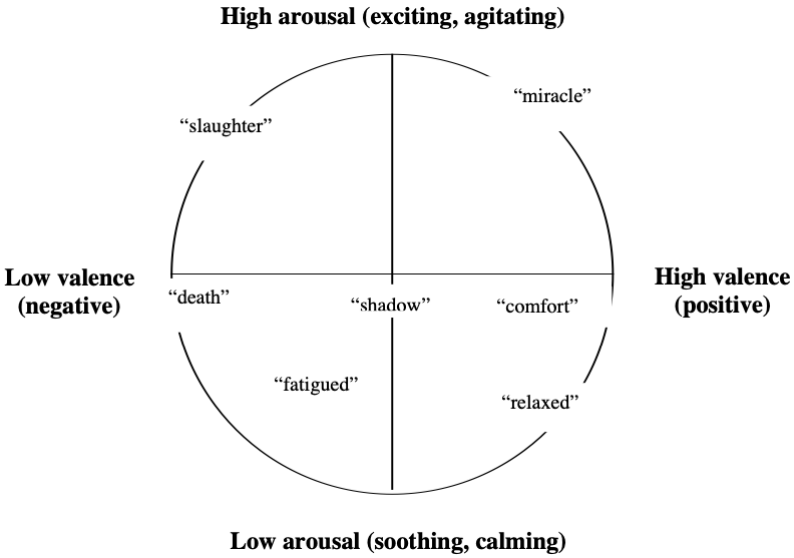

Figure 2.6: Valence and arousal plane, described in [2]

arousal-eliciting stimulus.

The experience of music listening is multidimensional. Different emotions are associated with different music patterns. For example, arousal is associated to:

- tempo (fast/slow)
- pitch (high/low)
- loudness (high/low)
- timbre (bright/soft)

while valence is associated to:

- mode (major/minor)
- harmony (consonant/dissonant)

as expressed in [12]. Emotion perception is correlated to the combination of music factor, rarely from just one of them. For example, loud chords and high-pitched chords tends to be feel as more positive valence than soft chords and low-pitched chords.

2.2.7 Music features

In MER analysis, an important step to define audio signals is to extract audio features and than apply a feature selection method.
There are several features that can be extracted from audio signal in order to represent five of the most useful perceptual dimensions of music listening:

- Energy: dynamic loudness, audio power, total loudness, specific loudness sensation coefficients.
- Rhythm: beat histogram, rhythm pattern, rhythm regularity, rhythm clarity, average onset frequency, average tempo.
- Temporal: zero-crossing, temporal centroid, log-attack-time.
- Spectrum: spectral centroid, spectral rolloff, spectral flux, spectral flatness.
- Harmony: salient pitch, chromagram centroid, harmonic change, pitch histogram.

These features are just an example of an infinite series of features that can be extracted from audio signals.

Gabrielsson et al. [12] noted that there are corresponding relations between the dimensional models and music features. Among these features, intensity is a basic feature, which is highly correlated with arousal and is used to classify the arousal dimension [13].
In [3] is shown a table summary of musical characteristics relevant to emotion, reported in 2.4. Despite the identification of these relations,

Features	Examples
Timing	Tempo, variation, duration, contrast
Dynamics	Overall level, crescendo/diminuendo, accents
Articulation	Overall staccato, legato, variability
Timbre	Spectral richness, harmonic richness
Pitch	High or low
Interval	Small or large
Melody	Range, direction
Tonality	Chromatic-atonal, key-oriented
Rhythm	Regular, irregular, smooth, firm, flowing, rough
Mode	Major or minor
Loudness	High or low
Musical form	Complexity, repetition, disruption
Vibrato	Extent, range, speed

Table 2.4: Musical features relevant to MER for [3]

many of them are not fully understood, still requiring further musicological and psychological studies, while others are difficult to extract from audio signals. Nevertheless, several computational audio features have been proposed over the years. While the number of existent audio features is high, many were developed to solve other problems (e.g., Mel-frequency cepstral coefficients (MFCCs) for speech recognition) and may not be directly relevant to MER.

Nowadays is not really clear the relationship between low-level and

mid-level features and mood. In order to capture different aspects is extracted a large set of features. This create a feature matrix that is then normalized in order to map them on the same range of values.

After the feature matrix is created is applied a feature selection or feature reduction algorithm to select the best set of features. Feature selection algorithms are based on two different ideas:

- High-level point of view: find the set of features that best model the concept. This lead to the accuracy of machine learning techniques being limited because of the limitation of the hypothesis done.
- Low level point of view: find the set of features that produces the best classification rate.

From the machine learning point of view, features are not necessarily of equal importance or quality, and irrelevant or redundant features may lead to inaccurate conclusion. Experiments have shown that, although the performance can thus be improved to a certain extent, using too many features leads to performance degradation [13].
With an highly discriminant sets of features, is not true that their combination produces a better discriminant power, for example if the set of features is 60, the number of possible combinations are:

$$n_{combinations} = \sum_{n=1}^{60} \binom{60}{k} \tag{2.1}$$

which is clearly impossible to compute, for this reason is applied some feature selection algorithms.

An example of feature selection for the categorical approach is the Sequential Feature Selection. It starts from an initial condition, and features are added or removed from a candidate subset while evaluating the *criterion* in two possibilities:

1. Sequential Forward Selection (SFS): features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion.
2. Sequential Backward Selection (SBS): features are sequentially removed from a full candidate set until the removal of further features increases the chosen criterion.

Another feature selection method is the Minimum-Redundancy-Maximum-Relevance (mRMR) which select the features with the highest relevance to the target class. Relevance is characterized in terms of *mutual information* which is defined as (given X and Y a pair of random variables):

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{2.2}$$

where $p(x, y$ is the joint probability mass function of X and Y, $p(x)$ and $p(y)$ are the marginal probability mass function of X and Y respectively.

On the other side, for dimensional approach, feature selection is for example RReliefF [14]. Basic idea of this algorithm is that try to estimate the quality of each attribute (in this context the features) according to how well their values distinguish between instances that are close each other.
The pseudocode of the RReliefF feature selection algorithm from [14]:

```
INPUT: training data {x_i}_{i=1}^N, {y_i}_{i=1}^N, parameters K, σ, n
OUTPUT: vector W of estimations of the importance of features
  set N_dC, N_dM[m], N_dC&dM[m], W[m] to 0
  for t = 1 to n
    randomly select an instance i
    select k instances nearest to i
    for each neighbor j
      N_dC = N_dC + diff(y_i, y_j) · d(i, j)
      for m = 1 to M
        N_dM[m] = N_dM[m] + diff(x_im, x_jm) · d(i, j)
        N_dC&dM[m] = N_dC&dM[m] + diff(y_i, y_j) · diff(x_im, x_jm) · d(i, j)
      end
    end
  end
  for m = 1 to M
    W[m] = N_dC&dM[m]/N_dC − (N_dM[m] − N_dC&dM[m])/(n − N_dC)
  end
```

Figure 2.7: RReliefF pseudocode

Another feature selection for dimensional approach is Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The method starts with all features and reduces them one by one, and hence is similar to backward selection. The goal of ICA is to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible. While the other well known linear transformation methods (PCA) benefit from the gaussianity of the data, ICA improves the classifier performance in the opposite case.

2.2.8 Machine learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence (AI). Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

When most people hear about ML they picture a robot, but it not just fantasy, it's already here, it has been around for decades in some specialized applications like *Optical Character Recognition*. The first ML application that became mainstream was done in 1990s, the *spam filter* [15].
A classical definition came from *Arthur Samuel* in 1959:

"*Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed*"

Another definition, more engineering-oriented is by *Tom Mitchell* in 1997:

"*A computer program is said to learn form experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E*"

The main difference between traditional programming and ML is well schematized in the figure 2.8 There are many different Machine Learning

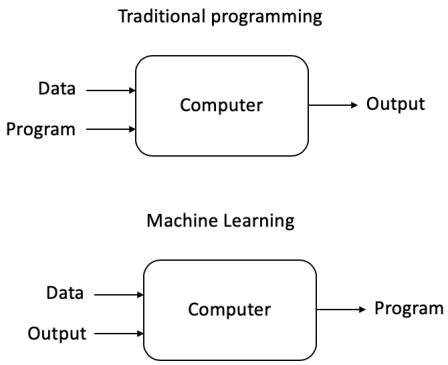

Figure 2.8: Traditional programming versus Machine Learning

systems. They can be classified in categories based on:

- Whether or not they are trained with human supervision (supervised, unsupervised, reinforcement learning).
- Whether or not they can learn incrementally on the fly (online and bach learning).
- Whether they work by comparing new data points to know data points, or instead detect patterns in the training data and build a predictive model (instance-based and model-based learning).

These criteria are not exclusive, they can be combined together.

In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each

image would have a label (the output) designating whether it contained the object.
Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.
In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range.
In this thesis the focus will be on **supervised learning**.

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as **training data**, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, called **feature vector**, and the training data is represented by a **matrix**. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.
In order to solve a problem of supervised learning, one has to perform steps:

1. Determine the training data type.

2. Gather a training set.

3. Determine the input feature representation of the learned function. Here input objects are transformed into feature vector which contains a number of features that describe the object.

4. Determine the structure of the learned function and corresponding algorithm.

5. Run the algorithm on the training set and optimize performances on a subset called *validation set* of the training set, or through *cross-validation*.

6. Evaluate the accuracy of the model.

There are several algorithms of supervised learning, there are no one that works best on all problems, due to this different algorithms are tested. Most widely used learning algorithms are:

- Support Vector Machines (SVM).
- Support Vector Regression (SVR).
- Linear Regression (LR).
- Decision Tree (DT).
- Neural Networks (NN).

Regression and Classification are both problems of supervised machine learning, the main difference between them is that the output variable in regression is numerical (or continuous) while that for classification is categorical (or discrete).

The task of MER is a regression problem both for dimensional, categorical and MEVD. In dimensional approach, the valence-arousal plane with a continuous space. Each point of the plane is considered an emotion state. This allow to overcome the categorical problem of granularity issue since the emotion plane implicitly offers an infinite number of emotion descriptions.
The regression approach applies a computational model that predicts the valence and arousal values of a music piece, which determine the placement of the music piece in the emotion plane [4].

A user can then retrieve music by specifying a point in the emotion plane according to his/her emotion state, and the system would return the music pieces whose locations are closest to the specified point. Because the 2D emotion plane provides a simple means for user interface, novel emotion-based music organization, browsing, and retrieval can be easily created for mobile devices.

Regression approach

A schematic diagram of the regression approach is in 2.9 where in the training phase, regression model are trained by learning the relationship between music features x and ground truth emotion values y. To denote regressors for valence and arousal are used r_V and r_A. In the test phase, given the features x_* of an input song, the regressors r_V and r_A can be applied to predict its emotion values $y_* = [v_*, a_*]^T = [r_V(x_*), r_A(x_*)]^T$. The regression theory aim at predicting a real value from observed variables, in MER application music features. The VA values are predicted directly from music features and due to this MER can be approached as a regression problem.

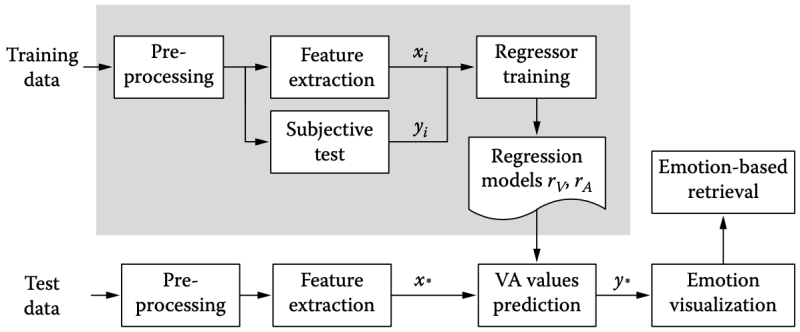

Figure 2.9: Schematic diagram of a regression approach

Given N inputs $(\mathbf{x}_i, y_i)$, with $i \in 1, ..., N$ where $\mathbf{x}_i$ is the feature vector of an object d_i (music piece), and y_i is the real value to be predicted (valence or arousal), a regressor $r(\cdot)$ is created by minimizing the mean squared error (MSE) ε:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}(y_i - r(\mathbf{x}_i))^2 \tag{2.3}$$

where $r(\mathbf{x}_i)$ is the prediction result for d_i.
In this thesis in mathematical expressions, **bold** font represent vectors and matrices.

To evaluate the performances of the regression approach with various ground truth data spaces, feature spaces and regression algorithms is used the R^2 statistics, which is a standard way for measuring the goodness of fit of regression models. It is calculated as:

$$R^2(\mathbf{y}, r(\mathbf{X})) = 1 - \frac{N\varepsilon}{\sum_{i=1}^{N}(y_i - \widehat{y})^2} = 1 - \frac{\sum_{i=1}^{N}(y_i - r(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \widehat{y})^2} \tag{2.4}$$

where $\widehat{y}$ is the mean of the ground truth. R^2 is comparable between experiments because of the normalization of the total squared error $N\varepsilon$ by the energy of the ground truth. The value of R^2 lies in $[-\inf; 1]$ where $R^2 = 1$ means the model perfectly fits the data, while a negative R^2 means the model is even worse than simply taking the sample mean.

The regression approach to MER, however, is not free of issues. First, the regression approach suffers from the subjectivity issue of emotion perception as it assigns the valence and arousal values to a music piece in a deterministic way. It is likely that different users perceive different emotion values in the music piece. Second, the regression approach requires numerical emotion ground truth to train the computational model, but performing such an emotion rating is a heavy cognitive load to the subjects.

3 Theoretical Background on EDA

This chapter introduces the readers to...

3.1 Some different sections

3.2 Remarks

4
State of the Art

This chapter introduces the models

4.1 Some different sections

4.2 Conclusive Remarks

In this chapter we introduce the main issues ...

5 Implementation and Results

In this chapter we present...
Then...
Finally ...

5.1 Some different sections

5.2 Conclusive Remarks

6 Dataset Improvements

In this chapter we present...
Then...
Finally ...

6.1 Some different sections

6.2 Conclusive Remarks

7 Conclusions and Future Works

This work of thesis proposes a methodology for...
The devised methodology is based on...
The main advantages are...
As far as the experiments are concerned...
The proposed approach has shown promising results both in simulation and in the experiments.

7.1 Future Works

Generalization We would like to generalize...

Challenging scenarios Another possible improvement is related to the extension of the proposed approach to...

Different approaches Finally we are moving towards a deeper analysis of... a a a a a a a a a a a a a a

Appendices

Bibliography

[1] 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*, (Hangzhou, China), pp. 375–376, 2003.

[2] E. A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal," *Reviews in the Neurosciences*, vol. 15, no. 4, pp. 241–252, 2004.

[3] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

[4] Y.-H. Yang and H. H. Chen, *Music emotion recognition*. USA: CRC Press, Inc., 1st ed., 2011.

[5] J. H. Lee and J. S. Downie, "Survey of music information needs, uses, and seeking behaviours: preliminary findings.," in *ISMIR*, vol. 2004, p. 5th, Citeseer, 2004.

[6] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of new music research*, vol. 33, no. 3, pp. 217–238, 2004.

[7] B. Van De Laar, "Emotion detection in music, a survey," in *Twente Student Conference on IT*, vol. 1, p. 700, 2006.

[8] K. Hevner, "Expression in music: a discussion of experimental studies and theories.," *Psychological review*, vol. 42, no. 2, p. 186, 1935.

[9] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365–389, 2018.

[11] K. F. MacDorman, Stuart Ough Chin-Chang Ho, "Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison," *Journal of New Music Research*, vol. 36, no. 4, pp. 281–299, 2007.

[12] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression.," 2001.

[13] J. L. Zhang, X. L. Huang, L. F. Yang, Y. Xu, and S. T. Sun, "Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods," *Multimedia Systems*, vol. 23, no. 2, pp. 251–264, 2017.

[14] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[15] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.