POLITECNICO
MILANO 1863

Dipartimento di Elettronica, Informazione e Bioingegneria

Master Degree in Computer Science and Engineering

Music Emotion Detection

A Framework Based on Electrodermal Activities

by:
Gioele Pozzi

matr.:
10454628
id.:
905756

Supervisor:
Sarti Augusto

Co-supervisor:
Borrelli Clara

Academic Year
2019-2020

POLITECNICO
MILANO 1863

Dipartimento di Elettronica, Informazione e Bioingegneria

Master Degree in Computer Science and Engineering

Riconoscitore di Emozioni Musicali

Un Framework Basato su Attività Elettrodermiche

Candidato:
Gioele Pozzi

matricola:
10454628
id.:
905756

Relatore:
Sarti Augusto

Co-relatore:
Borrelli Clara

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.
Music information behavior studies have identified emotion as an important criterion used by people in music searching and organization. It becomes significant the field of *music emotion recognition*.

Nowdays, is more and more important to retrieve and organize users music, due to the increasing platforms of streaming, which gives the access to a catalog of billions of songs.
The automatization of the recognition of perceived emotion in music allows users to organize and research music in a content-centric fashion.

Purpose of this thesis is to find a link between music and emotions during the listening of a song by combining audio and physiological signals analysis.

The inclusion of emotions is an hard task, due to the subjective nature of emotion perception. There are problems in the reliability of ground truth data and evaluation of prediction results, which are not troubles in problems as face recognition or speech recognition.

Sommario

Una delle funzioni più attrattive della musica è che questa può trasmettere e comunicare emozioni e modulare l'umore di una persona, come descritto in [1]. La musica può provocarci lacrime, consolarci quando siamo tristi, farci innamorare.
Gli studi fatti finora sulla musica, affermano che le emozioni sono un criterio importante per la ricerca e l'organizzazione dei brani musicali. Qui diventa fondamentale l'importanza del campo chiamato *music emotion recognition*.

Al giorno d'oggi, diventa sempre più importante il fatto di catalogare e organizzare la musica degli utenti, a causa dell'incremento di piattaforme di streaming musicale, le quali danno accesso ad un numero infinito di brani.
L'automatizzazione del riconoscimento delle emozioni percepite in musica, permette all'utente di organizzare e ricercare la musica in una visione più incentrata sul contenuto.

Lo scopo di questa tesi è quello di trovare il link tra emozioni percepite durante l'ascolto di un brano musicale attraverso l'analisi del segnale audio in primis, ma anche con l'utilizzo di segnali psicologici.

L'utilizzo delle emozioni , in generale, è un compito difficile, a causa della natura intrinseca delle emozioni percepite. Ci sono problemi di affidabilità dei dati empirici e la valutazione del modello di predizione, che d'altra parte non sono dei problemi nei casi ben noti di *face recognition* e *speech recognition*.

Acknowledgements

In primis vorrei ringraziare di cuore i miei genitori, che nonostante i problemi, le delusioni e le cadute mi hanno sempre sostenuto e dato una mano a rialzarmi, grazie Lorena e Paolo.
Un grazie infinito a chi mi ha sostenuto con spiegazioni, risposto alle mille domande e ha dedicato tanto tempo nella stesura e correzione di questa tesi, non ce l'avrei fatta senza, grazie Clara. Un grazie anche ai professori Sarti e Antonacci, che sia con le loro lezioni e i loro insegnamenti mi hanno fornito un notevole background per poter affrontare tutti i problemi.
Ringrazio anche tutto il *ISPG laboratory* e *ANTlab* anche solo per il sostegno morale durante la ricerca.
Un doveroso ringraziamento alla mia ragazza, che non ha mai smesso di credere in me e mi ha sempre spronato a dare il meglio, grazie Beatrice, non ce l'avrei fatta senza di te.
E come non potrebbero mancare gli inseparabili amici della biblioteca, Chri, Marci, Miry, Legre, Fede, senza di voi e la vostra compagnia non avrei passato neanche un esame.
Grazie anche ai miei compagni di università, Luca e Alessandro, ne abbiamo passate tante assieme, tra lezioni a Milano e la convivenza a Cremona, è stato un percorso con alti e bassi, non senza ostacoli, ma che ci ha fatto crescere come persone.
In fine, vorrei assolutamente non ringraziare questa situazione dovuta dalla pandemia, a tutto avrei pensato per il giorno della mia laurea, tranne che a questo.

"You never know how strong
you are until being strong
is the only choice you have."
Scott Jurek

Contents

List of Figures

List of Tables

Glossary

AC Alternate Current. 30

AI Artificial Intelligence. 19

ANOVA ANalysis Of VAriance. 77

ANS Autonomic Nervous System. 38

AUC Area Under the Curve. 75

BP Blood Pressure. 38

BVP Blood Volume Pulse. 42

CENS Chroma Energy Normalized Statistics. 67

CMIM Conditional Mutual Information Maximization. 36

CNN Convolutional Neural Network. 45, 52, 100

CNS Central Nervous System. 38

ComPaRe Computational Paralinguistic Evaluation. 60

DC Direct Current. 30

DCT Discrete Cosine Transform. 72

DISR Double Input Symmetrical Relevance. 36

DL Deep Learning. 45

DT Decision Tree. 21, 86, 87

DWT Discrete Wavelet Transform. 35, 44, 50, 71

ECG Electrocardiogram. 38, 41, 42, 44, 48, 49, 50

EDA ElectroDermal Activity. 25, 26, 28, 30, 33, 35, 36, 38, 42, 52, 55, 56, 58, 61, 62, 73, 74, 75, 84, 91, 93, 95, 96, 99, 100

EDR ElectroDermal Response. 25, 26

EEG Electroencephalogram. 38, 40, 41, 44, 45

EMD Empirical Mode Decomposition. 44

EMG Electromyogram. 38, 41, 43, 49

GSR Galvanic Skin Response. 26, 38, 48, 49

HL High Level. 17, 24

HRV Hearth Rate Variability. 38, 42, 49, 50, 51

ICA Independent Component Analysis. 18, 44

IIR Infinite Impulse Response. 28, 74

IOT Initial Orientation Time. 59

JMI Joint Mutual Information. 36

k-NN k-Nearest Neighbor. 14, 45, 50, 51, 52, 86

LASSO Least Absolute Shrinkage and Selection Operator. 85, 86

LDA Linear Discriminant Analysis. 45, 48, 49, 77

LL Low Level. 17, 24

LR Linear Regression. 21, 84

MER Music Emotion Recognition. 3, 4, 5, 7, 8, 13, 14, 15, 16, 21, 22, 24, 25, 35, 37, 55, 57, 60, 84, 99

MEVD Music Emotion Variation Detection. 12, 14, 21

MFCC Mel-Frequency Cepstral Coefficient. 16, 35, 72, 76

MIR Music Information Retrieval. 3, 4, 5, 8, 13, 55, 99

MIREX Music Information Retrieval Evaluation eXchange. 3

ML Machine Learning. 4, 13, 14, 19, 21, 33, 44, 45, 48, 51, 55, 57, 61, 62, 77, 84, 86, 87, 89, 99, 100

MLR Multivariate Linear Regression. 91, 93

MP Matching Pursuit. 48

mRMR minimum-Redundancy-Maximum-Relevance. 18, 45

MSE Mean Square Error. 22

NN Neural Network. 14, 21, 45, 49

PCA Principal Component Analysis. 18, 45, 48

PGR PsychoGalvanic Reflex. 26

PMEmo Popular Music with Emotional annotations. 55, 57, 59, 60, 61, 62, 75, 89, 91, 96, 97, 99

PNN Probabilistic Neural Network. 48

PNS Peripheral Nervous System. 38

PPG Photoplethysmography. 42

PSD Power Spectral Density. 45, 75

RF Random Forest. 45, 48, 87

RFE Recursive Feature Elimination. 80

RMSE Root Mean Square Error. 75, 84, 89, 90, 91, 94, 99

RMSSD Root Mean Square Squared Difference. 75

RNN Recurrent Neural Network. 45

RSP Respiration. 38, 43, 49

SBS Sequential Backward Selection. 17, 45

SC Skin Conductance. 26, 29, 38, 43, 49, 50

SCL Skin Conductance Level. 26

SCR Skin Conductance Response. 26, 28, 33, 35, 51, 52

SDSD Standard Deviation Discrete Differences. 75

SE Spectral Entropy. 45

SFS Sequential Forward Selection. 17, 45

SMNA SudoMotor Nerve Activity. 28

SNS Somatic Nervous System. 38

SPR Skin Potential Response. 26

SR Skin Resistance. 38

SSR Sympathetic Skin Response. 26

ST Skin Temperature. 38

STFT Short Time Fourier Transform. 66

SVC Support Vector Classification. 14

SVM Support Vector Machine. 14, 21, 45, 48, 49, 51

SVR Support Vector Regression. 21, 86, 91, 93

VA Valence Arousal. 10, 14, 15, 22, 59, 84, 96, 97

ZCR Zero Crossing Rate. 65

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.

As one can image, dealing with human emotion is not a simple task, due to their complexity and subjectivity. For this reason we used a data-driven method, basing our research on a large dataset, on data.
A data-driven model is based on the analysis of the data about a specific system. The concept of this model is to find relationships between the system state variables, input and output, with having an explicit knowledge of the behavior of the system.

1.2 Outline of the thesis

This thesis is organized as follows:

After a brief introduction about the objective of the thesis, in Chap-

ters 2 and 3 is presented a complete overview about the main arguments.
In Chapter 2, are presented Music Information Retrieval (MIR) and Music Emotion Recognition (MER).
In Chapter 3 Electrodermal Activity (EDA) and other physiological data using on-body sensors are given.

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. In the same we also illustrate 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 done by the creators 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.

The Music Information Retrieval deal with retrieving information from music. In the last few years compared a large variety of music streaming services. They are very useful, but they give the possibility to the user to find billion of songs and become necessarily to find a useful tool to search between songs.

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 Music Emotion Recognition

This chapter introduces the readers to the main basics about Music Information Retrieval (MIR) and Music Emotion Recognition (MER).
MIR is an interdisciplinary science where the goal is to retrieve relevant information from music. Researchers belonging to this community 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 a small but growing field of research with many real-world applications and is being used by businesses and academics to categorize, manipulate and even create music.
A few application to MIR can be:

- Music recommender systems, several already exist, but few are based upon MIR techniques, some systems do not use just similarity between subjects but also use audio retrieval to achieve better results in music recommendation as in Pandora[1].
- Intelligent and adaptive digital audio effects aim at design a system that determine the settings of audio effects based on the audio content.
- Music recording analysis as track separation, or also instrument recognition.
- Automatic music transcription, the process of converting an audio recording into symbolic, such score or a MIDI file.

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

- Automatic music tagging, as musical genre categorization or extraction of other high level features (the usual task for the yearly Music Information Retrieval Evaluation eXchange (MIREX)).

A broadly part of MIR is MER, where a useful application can be seen in this thesis.
In this chapter will be presented as first an introduction on MER. Due to the fact that this field is based on the emotions, they are explained and related to music. Is then presented the emotion space, where emotions are represented on a plane.
Is also shown the general framework of MER algorithms based on categorial or dimensional approaches, the music features extraction and selection, followed by some Machine Learning (ML) process to complete the general problem.
To conclude are mentioned some open issues related with MER.

2.1 Music Emotion Recognition

MER is an important topic in the field of MIR. Music is often referred as the language of emotion. People tend to listen different songs when in different emotional states. Therefore, categorizing music according to the type of emotion they express is becoming more and more important for internet music service provider. MER aims at modeling human emotion perception of music [9].
Automatic MER allows users to retrieve and organize their music collections in a fashion that is more content-centric than conventional methods based on metadata.
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 because 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.

Music plays an important role in human life, even more in the digital age. Never before such a large collection of music has 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.

However, 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 geographic area) assigned to music pieces by online users.
Even if emotion-based music retrieval was not yet well explored, a survey

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

conducted in 2004 from [10] 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?"* [10].

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?"*

Into another survey [11], they present findings from an exploratory questionnaire study featuring 141 music listeners (between 17 and 74 years of age) that offers some novel insights.
Emotions induced by music are subjective phenomena, there are significant differences between individuals. 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 retrieve 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 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.

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

2.2 Emotions and 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* that the performer tries to communicate with the listener.
- *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.

One of the aim of this thesis, is trying to link perceived and felt emotions, the former through the analysis of the music, the latter, through biometric signals and understand how are they related.

2.3 Emotion space

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 [12] from 1934, studied the relationship between music and emotions though experiments where subjects were 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.

Emotions, in the years, were conceptualized in two main approaches, the **categorical approach** and the **dimensional approach**. In the following sections will be presented these two different approaches, along with another one used for dynamic emotion recognition.

2.3.1 Categorical approach

The first assumption of this emotion conceptualization is that emotions are categorized and categories are distinct from each other. Within this

approach, it is necessary to assume that there are a limited number of innate and universal emotion categories, such as:

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

All the 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, a famous categorical approach to emotion conceptualization is Hevner's adjective checklist. He defined eight clusters positioned in circle as in Figure 2.1. Adjectives in the same cluster are nearly identical, neighbor clusters have similar meaning. The opposite position of a given cluster is its opposite in emotional sense.

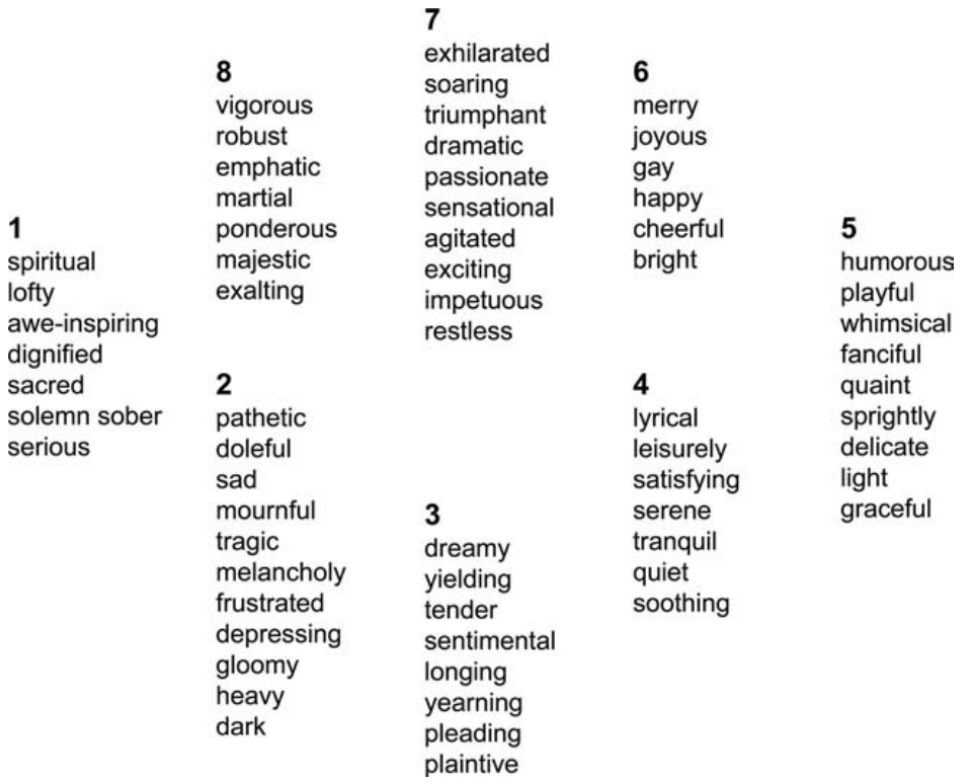


Figure 2.1: Eight clusters proposed by Hevner

Hevner's checklist proposed in 1935 was updated and regrouped into ten groups by Fansworth and into nine groups in 2003 by Schubert.

Drawback 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 that using a finer granularity it does not necessarily solve the issue because the language for describing emotions is inherently ambiguous and varies from person to person. Using a large number of emotion classes could confuse the subject and is impractical for psychological studies falsing results.

2.3.2 Dimensional approach

While 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.

Several names from researchers gave very similar interpretations of the resulting factors like tension/energy, intensity/softness, tension/relaxation. Most of the factors correspond to the two dimensions of emotion the *valence* (positive and negative affective states) and *arousal* (energy and stimulation level) to create the Valence Arousal (VA) space. Some studies found that valence as well as intensity, is triggered by the amygdala, while the arousal by the reptilian brain.
Russel, proposed a circumplex model of emotion in [13] which consist in a two-dimensional, circular structure, as in Figure 2.2 involving the dimensions of valence and arousal. In this structure, emotions that are inversely correlated, are placed across the circle from one another.

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.
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 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:

Figure 2.2: Russel's circumplex model of affect

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

as expressed in [14].
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.

Also dimensional approach have some drawbacks. For example, 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. To overcome this issue, some researchers tired to add a third dimension called *potency*, varying from dominant to submissive, to obtain a more complete picture of emotion. However, this would increase the cognitive

load on the subjects and at the same time requires a more complex interface and makes hard to annotate the process. The third dimension problem is still in discussion.

2.3.3 Music Emotion Variation Detection

An important aspect that is not addressed in the previous two Paragraphs (2.3.1 and 2.3.2) is 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 as Music Emotion Variation Detection (MEVD). Usually subjects are asked to rate valence and arousal in response of the stimulus over time.
For example, songs can be described by valence and arousal curves as in Figure 2.3.

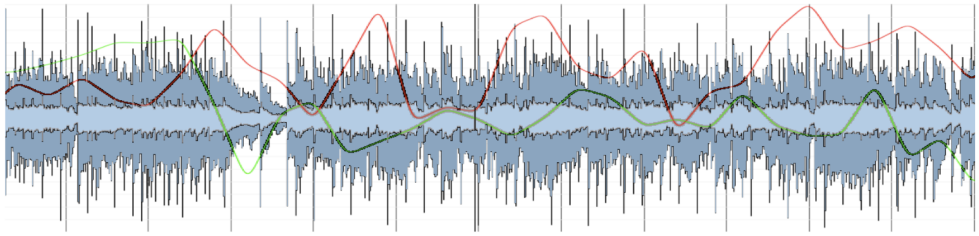


Figure 2.3: Valence and arousal curves for MEVD

2.4 MER algorithms

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.
In the following section will be presented the general algorithm path for MER problems.

2.4.1 General framework

Nowdays, main approaches are still context-based approaches based on human tagging, but it is not possible to annotate a great amount of songs and there is a possibility of human mistakes. To overcome those problems, ML and data mining techniques are used to model the relationship between music and emotion. ML is used to automatically infer mood and mood variation perceived in songs.
The training and automatic recognition model typically consists of the following steps:

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, 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 [15].

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

5. 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 have a standardized visualization.

6. Apply a learning algorithm between music features and emotion labels/values by training a ML model to learn the relationship between emotion and music. Music emotion classification is carried out with classification ML algorithms, such as Neural Network (NN), k-Nearest Neighbor (k-NN), decision tree, Support Vector Machine (SVM) and Support Vector Classification (SVC).

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

The music emotion recognition process can be schematized in the Figure 2.4 with a division in training and testing, in order to apply ML methods.

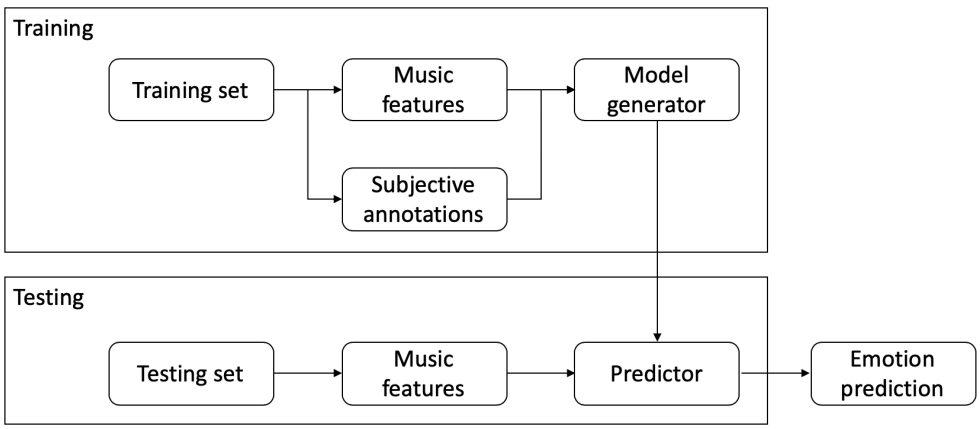


Figure 2.4: MER general framework process

Researches that work on MER can be classified into two approaches, categorical and dimensional, which are based on the emotion representation ideas.
In categorical approach, subjective annotations are global terms, in dimensional approach subjective annotations are point in the VA space and in MEVD they are sequences of points in the dimensional space.

2.4.2 Categorical approach

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

2.4.3 Dimensional approach

The dimensional approach to MER defines emotions as numerical values over VA plane. 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. Due to the fact that the emotion plane contain an infinite number of emotion descriptions, the granularity and ambiguity issues are relieved.
MER problem became a regression problem, and two independent models, as regressors, are trained to predict separately valence and arousal values.
The dimensional approach requires the subjects to annotate the numerical VA values. This requirement impose an high cognitive load on the subjects.

Pros and cons of categorical and dimensional approach are schematized in the Table 2.3.

	Pros	Cons
Categorical	Intuitive Natural language Atomic description	Lack a unifying model Ambiguous Subjective Difficult to offer fine-grained differentiation
Dimensional	Focus on 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

2.5 Music features

In MER analysis, an important step 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. [14] 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 predict the arousal dimension [16].
In [5] is shown a table summary of musical characteristics relevant to emotion, reported in Table 2.4.

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 [5]

Despite the identification of these relations, 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 Coefficient (MFCC) 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 (HL) 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 (LL) point of view: find the set of features that produces the best classification rate.

2.5.1 Feature selection

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 [16].
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 [17]. 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 [17]:

```
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 2.5: 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.

2.6 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.

ML is not just fantasy, it's already here, it has been around for decades in some specialized applications. The first ML application that became mainstream was done in 1990s, the *spam filter* [18].
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 2.6.

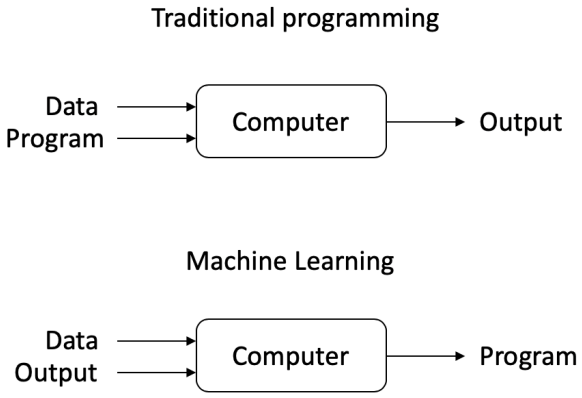


Figure 2.6: Traditional programming versus Machine Learning

There are many different ML 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 was 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* (a statistical method used to estimate the accuracy of ML models).

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:

- SVM.
- Support Vector Regression (SVR).
- Linear Regression (LR).
- Decision Tree (DT).
- 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 [9].

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.

2.6.1 Regression approach

A schematic diagram of the regression approach is in 2.7 where in the training phase, regression model are trained by learning the relationship between music features x and ground truth emotion values y.

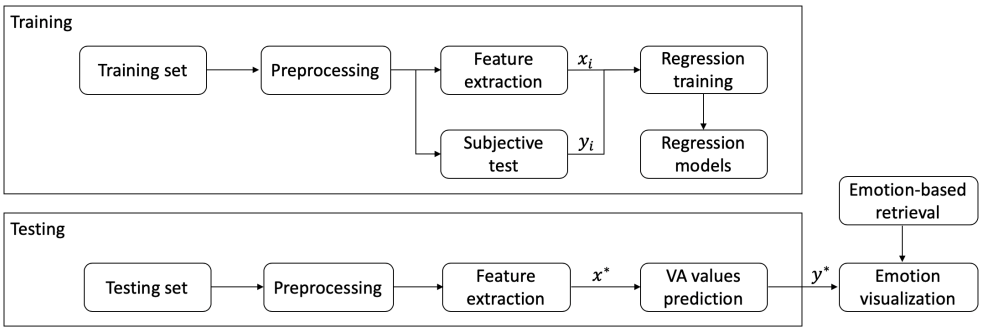


Figure 2.7: Schematic diagram of a regression approach

Regressors for valence and arousal are denoted with 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:

$$y_* = [v_*, a_*]^T = [r_V(x_*), r_A(x_*)]^T \tag{2.3}$$

The regression theory aims 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.

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 Square Error (MSE) ε:

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

where $r(\mathbf{x}_i)$ is the prediction result for d_i.
In this thesis in mathematical expressions, **bold** font is used to 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-squared model, R^2 statistics, which is a standard way for measuring the goodness of fit of regression models.
It is calculated as:

$$R^2(\mathbf{y}, r(\mathbf{X})) = 1 - \frac{N\varepsilon}{\sum_{i=1}^{N}(y_i - \widehat{y})^2} = 1 - \frac{\sum_{i=1}^{N}(y_i - r(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \widehat{y})^2} \tag{2.5}$$

where $\widehat{y}$ is the mean value of the ground truth. R^2 is comparable between experiments thanks to 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.

2.7 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 [19].

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 the management 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 LL audio signal and 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.

3 Theoretical Background on EDA

This chapter introduces the readers to Electrodermal Activities and how are they related with MER task.

First is presented a general and theoretical introduction of these ElectroDermal Activity (EDA), how were they discovered and how are processed to have significant results.

After a general introduction, some recording techniques are shown, using electrodes positioned in different parts of the body.

It follow an explanation of the different techniques for preprocessing the data, as artifacts removal.

At last is mentioned an explanation of the main features that can be extracted from EDA signals.

3.1 Electrodermal Activity

Already in the 80's, psychological factors related to electrodermal phenomena were observed. It became an important field of study, due to the fact its ease of obtaining a distinct ElectroDermal Response (EDR), the intensity of which seems apparently related to stimulus intensity and/or its psychological significance [20].

While there is still widespread disagreement and confusion about the nature and causes of musically evoked emotions, recent studies involving real-time observation of brain activity seem to show that areas of the brain linked with emotion (as well as pleasure and reward) are activated by music listening [21].

EDA is arguably the most useful index of changes in sympathetic arousal that are tractable to emotional and cognitive states as it is the only autonomic psychophysiological variable that is not contaminated

by parasympathetic activity. EDA has been closely linked to autonomic emotional and cognitive processing, and is a widely used as a sensitive index of emotional processing and sympathetic activity.
This coupling between cognitive states, arousal, emotion and attention enables EDA to be used as an objective index of emotional states. It can also be used to examine implicit emotional responses that may occur without conscious awareness or are beyond cognitive intent (i.e., threat, anticipation, salience, novelty).

3.1.1 Terminology and history

EDA was first introduced by Johnson and Lubin in 1966 [22] as a common term for all electrical phenomena in skin, including all active and passive electrical properties that can be traced back to the skin and its appendages.
EDA is the property of the human body that causes continuous variation in the electrical characteristics of the skin. Historically, EDA has also been known as Skin Conductance (SC), Galvanic Skin Response (GSR), EDR, PsychoGalvanic Reflex (PGR), Skin Conductance Response (SCR), Sympathetic Skin Response (SSR) and Skin Conductance Level (SCL). The long history of research into the active and passive electrical properties of the skin by a variety of disciplines has resulted in an excess of names, now standardized to EDA.

The use of the term *response* for electrodermal phenomena suggests that there is a distinct relationship to a stimulus producing an EDR. Sometimes there are parts that cannot be traced to any specific simulation, they are called *spontaneous* or *non-specific* EDR.

3.1.2 SCL and SCR division

There is ample empirical evidence that electrodermal phenomena are generated by sweat gland activity in conjunction with epidermal membrane processes. Skin conductance is characterized by:

- Tonic also called SCL, smooth underlying slowly changing level, it accounts for the general levels of the conductivity of the skin.
- Phasic, SCR rapidly changing peaks, results from momentary sympathetic activation when arousing stimuli are present.

When sweat gland activity is abolished in humans, either as a result of congenital absence, by sympathectomy, by peripheral sudomotor nerve discharge, or by pharmacological blocking, SCR and Skin Potential Response (SPR) are normally eliminated and SCL is reduced [23].

In the Figure 3.1 can be seen the plot over time of an EDA signal and its decomposition in tonic and phasic parts extracted using pyphysio[1] library [24] on an EDA signal.

[1] https://github.com/MPBA/pyphysio

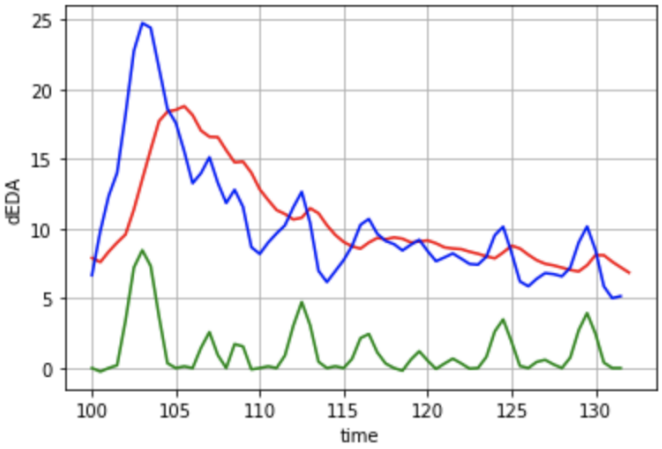


Figure 3.1: Representation EDA signal (in blue), driver signal (in red) and phasic signal (in green)

The time series of the change of SC is characterized by a slowly varying tonic activity and fast varying phasic activity. The SCR shows a steep incline to the peak and a slow decline to the baseline. The successions of SCR usually results in a superposition of subsequent SCR as one SCR arises on top of the declining trail of the preceding one.
The Figure 3.2 from [2] shows a SC data section. The upper row shows the original SC data. The middle row shows the driver signal which results from deconvolution of the SC data. Inter-impulse data are used to estimate the tonic part of the driver at 10-s intervals (tonic grid points). The tonic driver is used to compute the tonic SC (see upper row). Subtraction of the tonic part from the driver results in the phasic driver (lower row). The phasic driver shows a virtually zero baseline and distinct phasic responses.

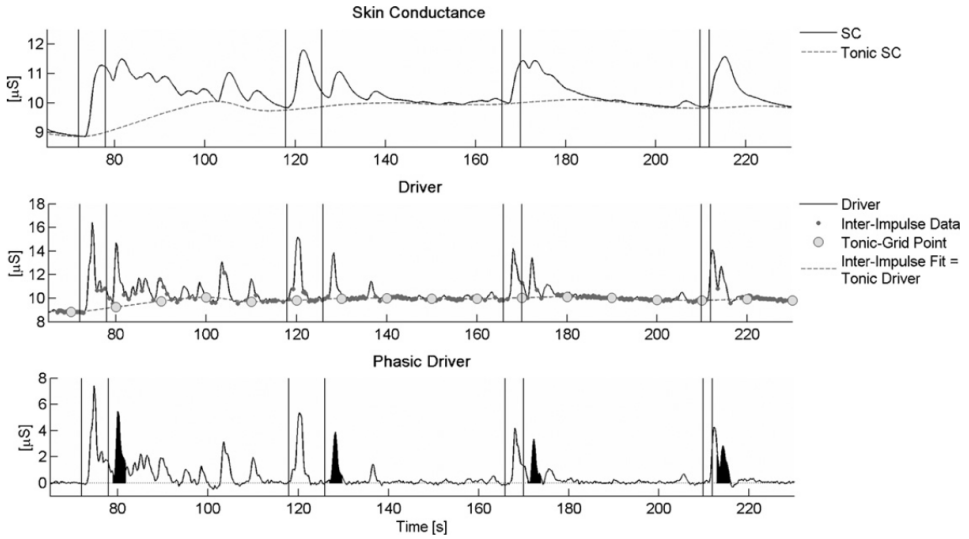


Figure 3.2: Skin conductance and phasic driver extraction from [2]

3.1.3 Decomposition algorithms and tools

To process EDA data dividing in phasic and tonic component, in the literature, several algorithms and tools were presented.

To separate the signal components, generally is designed an Infinite Impulse Response (IIR) Butterworth low-pass filter with a cut-off frequency of $0.001Hz$ (stop frequency of $1Hz$ at $-60dB$). The tonic component is then extracted from the output of the IIR filter, while the phasic component is obtained from the difference between the original signal (supplied to the IIR filter) and the tonic component.

To separate phasic and tonic components, is also used an algorithm called *cvxEDA*. The model of the *cvxEDA*[2] assumes that the observed SCR (y) is the sum of the phasic activity (r), a slow tonic component (t), and an additive independent and identically distributed zero-average Gaussian noise term (ε):

$$y = r + t + \varepsilon \tag{3.1}$$

Physiologically-plausible characteristics (temporal scale and smoothness) of the tonic input signal can be achieved by means of a cubic spline with equally-spaced knots every $10s$, an offset and a linear trend term:

$$t = Bl + Cd \tag{3.2}$$

where:

- B is a tall matrix whose columns are cubic B-spline basis functions
- l is the vector of spline coefficients
- C is a $Nx2$ matrix with $C_{i,1} = 1$, and $C_{i,2} = \frac{i}{N}$
- d is a $2x1$vector with the offset and slope coefficients for the linear trend

Phasic component is the result of a convolution between the SudoMotor Nerve Activity (SMNA) p and an impulse response $h(t)$ shaped as a biexponential Bateman function:

$$h(t) = (e^{-t/\tau_1} - e^{-t/\tau_2})u(t) \tag{3.3}$$

where τ_1 and τ_2 are the slow and the fast time constants of the phasic curve shape and $u(t)$ is the unitary step function.
Referring to [25], the final model can be written as:

$$y = Mq + Bl + Cd + \varepsilon \tag{3.4}$$

[2]https://github.com/Iciti/cvxEDA

Given the EDA model 3.4, *cvxEDA* formulated the problem as a minimization problem as:

$$minimize\ \frac{1}{2}||Mq + Bl + Cd - y||^2 + \alpha\delta||Aq||_1 + \frac{\gamma}{2}||l||_2^2 \tag{3.5}$$

$$subject\ to\ Aq >= 0$$

This problem can be solved using one of the many sparse-QP solvers in order to find the optimal $[q, l, d]$, than find tonic component t from 3.2.

One tool is EDAtool[3]. It is a function developed to preprocess EDA signal including removal of electrical noise and artifact detection. It separates also the signal in phasic and tonic components.

Another tool that is able to separate signal components is Ledalab[4]. This software aims to provide EDA analysis though two methods:

1. Continuous decomposition analysis, which performs a decomposition of SC data into continuous signals of phasic and tonic activity.

2. Discrete decomposition analysis, which performs a decomposition of SC data into distinct phasic and tonic activity by means of non-negative deconvolution.

[3]http://www.musicsensorsemotion.com/2012/06/21/edatool/
[4]http://www.ledalab.de

3.2 Measurement principles

EDA can be measured both without externally applied voltage (endosomatic method) or with application of Direct Current (DC) or Alternate Current (AC) (exosomatic method). The widespread used method is the exosomatic with DC recordings. With direct voltage, skin resistance measurements will result when current is constant, while skin conductance measurement will result when voltage is kept constant.

There are some factors that should be controlled as possible sources or variance in EDA recordings, like environmental conditions as the climatic conditions and physiological factors like age, gender and ethnic differences.

EDA can be measured in many different ways electrically including skin potential, resistance, conductance, admittance, and impedance. It achieves this by passing a minuscule amount of current between two electrodes in contact with the skin. The units of measurement for conductance are microSiemens (μS).

3.2.1 Recording techniques

Electrodermal recording is usually performed with two electrodes. Exosomatic techniques use two active sites, while endosomatic recording requires an active and an inactive site.
Figure 3.3 illustrate the preferred palmar recording areas for exosomatic and endosomatic EDA recordings. Sites A and B for bipolar recordings. C and D for volar electrode sites.

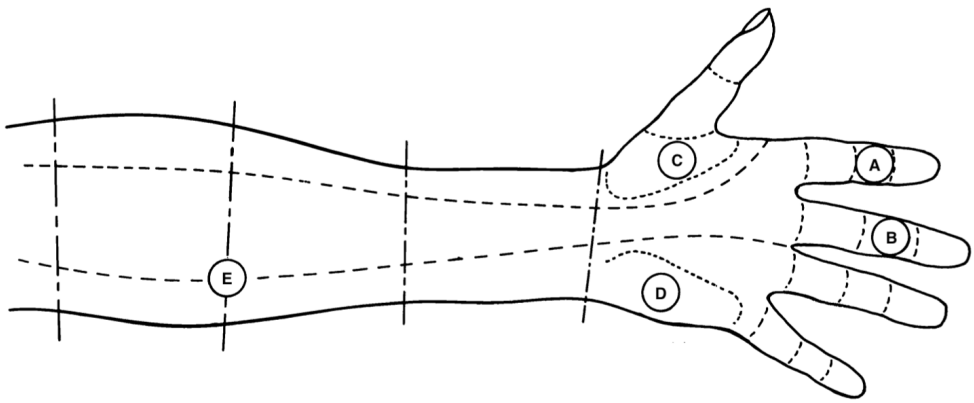


Figure 3.3: Preferred palmar recording areas for exosomatic and endosomatic EDA recordings

3.2.2 Wearable technologies

During the last years, some wearable devices were made, in order to extract EDA data through sensors.

For example, one wearable device is from Empatica[5]. Product examples can be seen in Figure 3.4.

(a) e4 (b) embrace 2

Figure 3.4: Empatica wearable devices

Empatica has designed a system support real-world applications for seizure detection and characterization. Empatica is running a clinical trial, open to Embrace users, to collect and validate biometric signals from epilepsy patients using the Empatica Embrace watch and Alert app and compare them to e-diary seizure report information.

Another device is from Bitbrain[6] as in Figure 3.5.

(a) Biosignal device from Bitbrain (b) Positioning the sensors

Figure 3.5: Bitbrain wearable devices

The sensors are located on the fingers' first and second phalanges (optimal measurement points) as shown in 3.5.

Another one device is from iMotions[7]. The name of the device is *Shimmer3 GSR+* which monitors skin conductivity between two electrodes attached to two fingers of one hand as can be seen in Figure 3.6. Caused by a stimulus the sweat glands become more active, increasing

[5]https://www.empatica.com/en-eu/
[6]https://www.bitbrain.com
[7]https://imotions.com

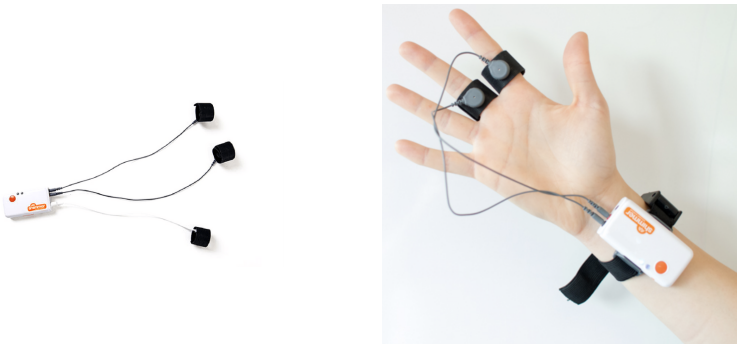


Figure 3.6: Bitbrain wearable devices

moisture on the skin and allowing the current to flow more readily by changing the balance of positive and negative ions in the secreted fluid (increasing skin conductance).

3.3 EDA preprocessing

EDA data is often captured by wearable devices, which makes the signal collected vulnerable to several types of noise. Artifacts can be generated from electronic noise or variation in the contact between the skin and the recording electrode caused by pressure, excessive movement or adjustment of the device [3].
They may be mistaken for a skin conductance response, and this must be avoided.

Typically, as Boucsein [20] report, the shape of an SCR lasts between $1s$ to $5s$, has a steep onset and an exponential decay and reaches an amplitude of at least $0.01\mu S$. An example of a typical SCR in Figure 3.7.

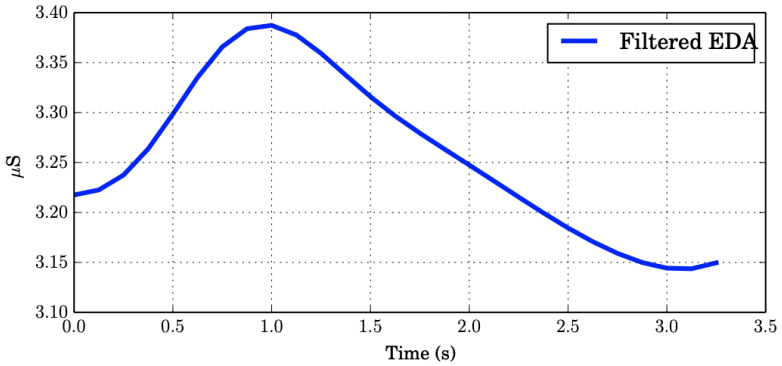


Figure 3.7: Example of a SCR shape

Currently, many researchers deal with signal artifacts and noise by applying exponential smoothing or low-pass filtering.
Additionally, filter cutoff frequencies are based only loosely on prior knowledge of typical characteristics of SCR shape, and vary widely study to study (from $1Hz$ to $5Hz$). The cutoff frequency ultimately chosen for a study is specific to that particular study, making generalization difficult.

There are much relevant techniques that are also able to recognize and compensate for large-magnitude artifacts that can result from pressure or movement of the device during recordings.
In [3] is presented a figure reported at Figure 3.8, which shows a portion of signal that contains three artifacts, in which the fast decrease could not be produced by human physiology. Comparing the raw signal and the filtered version, the low-pass filter has not removed the artifacts.
Some researchers, as Boucsein analysis [20], develop heuristic techniques for removing atypical portion of the EDA signal. Someone decide to discard portion of their data where the signal increased more than 20% per second or decreased more than 10% per second.

In another case, a study which collected EDA from two sensors (on both the ankle and wrist) [26] was able to detect artifacts by looking for epochs when only one of the two sensors had an abnormally low signal,

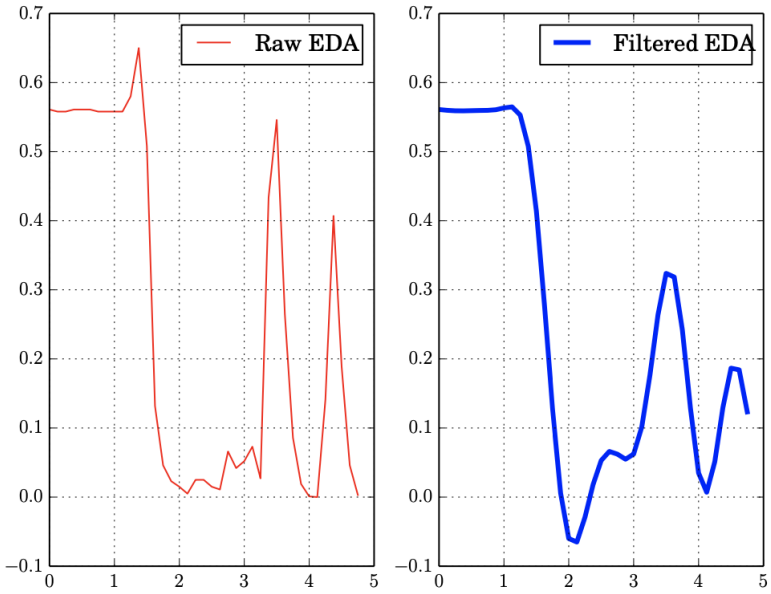


Figure 3.8: Portion of an EDA signal, the raw signal on the left in red, a $1Hz$ low-pass filter applied on the signal to the left in blue in [3]

or showed an unusually rapid increase or decrease.

In [3] developed a ML algorithm for automatically detecting EDA artifacts, providing empirical evaluation of classification performances.

3.4 EDA features

As for MER analysis, also in EDA data analysis, is important to find which features need to be extracted and then which feature selection method must be carried.

Emotion recognition from EDA has been commonly used for the assessment of user's experience in a variety of contexts such as recreational and games [27] and driving [28]. Previous research has explored the predictive power of a diverse set of EDA features of different types, including time domain, frequency domain, and time-frequency domain features.

Regarding time domain features, most usually features considered are the statistical parameters of the signal as:

- Mean value: μ is the central value of a discrete set of numbers $x_1, x_2, ..., x_n$, specifically, the sum of the values divided by the number of values:
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.6}$$

- Standard deviation: is a measure of the amount of variation or dispersion of a set of values:
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3.7}$$

- Kurtosis: is a measure of the "tailedness" of the probability distribution of a real-valued random variable:
$$kurt = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4} \tag{3.8}$$

- Skewness: is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean:
$$skew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3} \tag{3.9}$$

In [6], as an example, were extracted the features for EDA data shown in table 3.1, where Discrete Wavelet Transform (DWT) is an implementation of the wavelet transform using a discrete set of wavelet scales.
Other cases, researchers have focused on event-related features of EDA. They are useful when are presented to the subjects some events, stimulus, like images or sounds.
Examples of event-related aspects of EDA considered in other studies are SCR amplitude, SCR peak count, mean SCR rise time, or the sum of SCR areas.

Domain	Feature vector	Number of features
Time	SCR related	7
	Statistical features	8
	Hjorth features	2
	Higher Order Crossing	5
Frequency	Statistical features	8
	Band power	9
Time-Frequency	DWT coefficients	56
	SWT features	40
	MFCC	481
	Statistical features MFCC	5

Table 3.1: Features extracted in [6]

Fewer researches, as [6] remarks, has focused on the predictive power of EDA related to the frequency domain. The frequency domain analysis has shown superior capability for the gradient component's detection of individual SCR.
Due to the different rate of physiological process, EDA signals vary significantly with the frequency [29].
Frequency oscillations of EDA signals can be divided into different frequency sub-bands to analyze it. Indeed previous researchers has considered statistical aspects (variance, range, signal magnitude area, skewness, kurtosis, harmonics summation) and spectrum power of five frequency bands, as well as their minimum, maximum, and variance.

As for audio, also for EDA data, after constructing a feature matrix, need to apply an algorithm of feature selection to improve data reliability.

Some examples of feature selection could be:

- Joint Mutual Information (JMI): focuses on the increasing complementary information between features.
- Conditional Mutual Information Maximization (CMIM): it can properly identify truly redundant features and noisy features, and gives preference to informative, uncorrelated features.
- Double Input Symmetrical Relevance (DISR): a normalized variant of JMI.

In general it is not known which features are most appropriate for emotion recognition from EDA and previous works have made limited contributions on a systematic comparison of EDA features.
In [6] there is a table showing various features extracted for EDA signals already presented in Table 3.1 with also references in the literature.

4
State of the Art

This chapter introduces the readers to a complete review of the problem and all the different resolution possibilities.
First section deal with a general explanation of various physiological signals that are used in different application. Later, a general methodology of physiological signal processing is presented with an evidence on various issues.
Fourth section of the chapter deals with a complete state of the art review on the task of emotion recognition through physiological signals.

4.1 Physiological signals

In this section will be defined a general overview on physiological signals that can be used in order to achieve a solution to the MER problem.

Emotions, which affect both human physiological and psychological status, play a very important role in human life. Positive emotions help improve human health and work efficiency, while negative emotions may cause health problems. Long term accumulations of negative emotions are predisposing factors for depression, which might lead to suicide in the worst cases.
The emotion often refers to a mental state that arises spontaneously rather than through conscious effort and it is accompanied by physical and physiological changes, relevant to the human organs and tissues such as brain, heart, skin, blood flow, muscle, facial expressions, voice, etc. [30].

Emotion recognition has been applied in many areas such as safe driving [31], health care especially mental health monitoring [32], social security [33], and so on.

In general, emotion recognition methods could be classified into two major categories:

- Using human physical signals such as facial expression [34], speech [35], gesture, posture, etc. This method has the advantage of easy collecting and is a chapter which has been studied for years. On the other side, the reliability cannot be guaranteed, as it is relatively easy for people to control the physical signals like facial expression or speech to hide real emotions, especially during social communications.

- Using internal signals as:

  - Electroencephalogram (EEG)
  - Electrocardiogram (ECG)
  - Electromyogram (EMG)
  - Blood Pressure (BP)
  - Hearth Rate Variability (HRV)
  - EDA as:
    * Skin Resistance (SR)
    * Skin Temperature (ST)
    * SC
    * GSR
  - Respiration (RSP)

  These signals are produced by the Nervous System which is divided into:

  - Central Nervous System (CNS)
  - Peripheral Nervous System (PNS): consist of the Autonomic Nervous System (ANS) and Somatic Nervous System (SNS).

  EEG, ECG, EMG, GSR and RSP change in a certain way when people face some specific situations. Physiological signals are in response to the CNS and ANS. Due to the fact that CNS and ANS are involuntarily activated, they cannot be controlled.

In the Table 4.1 is shown a summary of various papers using different biological signals.

Biological signal	Paper
ECG	[36], [37], [38], [39], [40]
ECG, EMG, RSP	[41]
ECG, GSR	[42]
HRV, SR	[43]
EEG	[44]
HRV	[45]

Table 4.1: Papers with correspondent biological signal used

In Table 4.2 is presented the relationship between emotions and physiological features, thanks to [30]. Arrows indicate increased (↑), decreased (↓), no change in activation from the baseline (−) or both increases and decreases in different studies (↑↓).

Signal	Anger	Anxiety	Embarrassment	Fear	Amusement	Happiness	Joy
Cardiovascular							
HR	↑	↑	↑	↑	↑↓	↑	↑
HRV	↓	↓	↓	↓	↑	↓	↑
LF		↑		(−)		(−)	
LF/HF		↑			(−)		
PWA				↑			
PEP	↓		↓	↓	↑	↑	↑↓
SV	↑↓	(−)		↓		(−)	↓
CO	↑↓	↑	(−)	↑	↓	(−)	(−)
SBP	↑	↑	↑	↑	↑	↑	↑
DBP	↑	↑	↑	↑	↑	↑	(−)
MAP			↑	↑	↑	↑	
TPR	↑			↓	↑	↑	(−)
FPA	↓	↓		↓	↓	↑↓	
FPTT	↓	↓		↓		↑	
EPTT		↓		↓		↑	
FT	↓	↓		↓	(−)	↑	
Electrodermal							
SCR	↑	↑		↑	↑		
nSRR	↑	↑		↑	↑	↑	↑
SCL	↑	↑	↑	↑	↑	↑	(−)
Respiratory							
RR	↑	↑		↑	↑	↑	↑
Ti	↓	↓		↓	↓	↓	
Te	↓	↓		↓		↓	
Pi	↑			↑		↓	
Ti/Ttot				↑	↓		
Vt	↑↓	↓		↑↓	↑↓	↑↓	
Vi/Ti						↑	
Electroencephalography							
PSD(α wave)	↑	↑		↓	↑	↑	↑
PSD(β wave)	↓				↑		
PSD(γ wave)				↓	↑	↑	↑
DE (avg)	↑	(−)		↓		↑	↑
DASM (avg)	(−)			↑	↓	↓	↓
RASMs (avg)	↑			↑		↓	

Table 4.2: Relationship between emotions and physiological features

The position of different biosensors is shown in Figure 4.1.

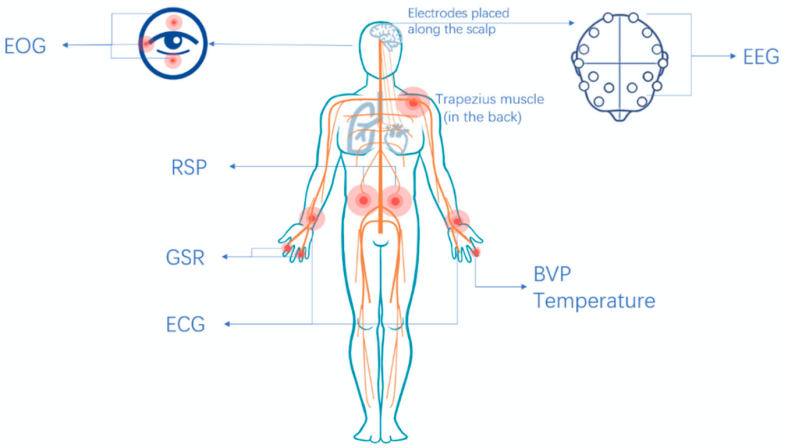


Figure 4.1: Position of the bio-sensors

4.1.1 Electroencephalogram

EEG is an electrophysiological monitoring method to record electrical activity of the brain. EEG measures voltage fluctuations resulting from ionic current within the neurons of the brain. Clinically, EEG refers to the recording of the brain's spontaneous electrical activity over a period of time, as recorded from multiple electrodes placed on the scalp.

EEG is most often used to diagnose epilepsy, which causes abnormalities in EEG readings. It is also used to diagnose sleep disorders, depth of anesthesia, coma, encephalopathies, and brain death.
Many studies have indicated that the physiological correlates of emotions are likely to be found in the central nervous system rather than simply in peripheral physiological responses. Researchers have supported this viewpoint using EEG or other neuroimaging (e.g., functional Magnetic Resonance Imaging) approaches to investigate the specificity of brain activity associated with different emotional states.
However, most of the available studies on emotion-specific EEG response have focused on EEG characteristics at the single-electrode level, rather than at the level of EEG-based functional connectivity.

4.1.2 Electrocardiogram

ECG is a recording of the electrical activity of the hearth using electrodes placed on the skin. These electrodes detect small electrical changes that are a consequence of cardiac muscle depolarization followed by repolarization during each cardiac cycle, the heartbeat.

There are three main components to an EEG: the P wave, which represents the depolarization of the atria; the QRS complex, which represents the depolarization of the ventricles; and the T wave, which represents the repolarization of the ventricles, as in Figure 4.2.

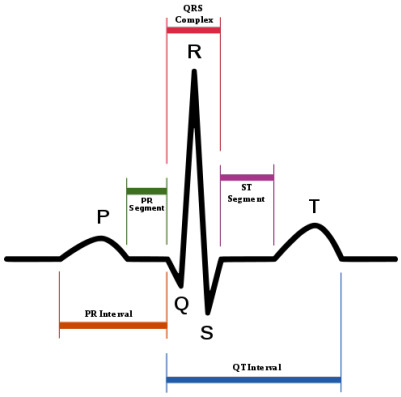


Figure 4.2: ECG of a heart in normal sinus rhythm, PQRST wave

4.1.3 Electromyogram

EMG is an electrodiagnostic medicine technique for evaluating and recording the electrical activity produced by skeletal muscles.
An electromyograph detects the electric potential generated by muscle cells when these cells are electrically or neurologically activated. The signals can be analyzed to detect medical abnormalities, activation level, or recruitment order, or to analyze the biomechanics of human or animal movement.

Therefore, the best readings are obtained when the sensor is placed on the muscle belly and its positive and negative electrodes are parallel to the muscle fibers. Since the number of muscle fibers that are recruited during any given contraction depends on the force required to perform the movement, the intensity (amplitude) of the resulting electrical signal is proportional to the strength of contraction.

In psychophysiology, EMG was often used to find the correlation between cognitive emotion and physiological reactions. In the work by Sloan [46], for example, the EMG was positioned on the face (jaw) to distinguish *smile* and *frown* by measuring the activity of zygomatic major and corrugator supercilli. In experiment of [41], bipolar electrodes were placed at the upper trapezius muscle (near the neck) in order to measure the mental stress of the subjects.

4.1.4 Hearth Rate Variability

HRV measure the beat-to-beat temporal changes of the hearth rate, sometimes it is calculated from ECG, but the usability of measuring the ECG is limited. HRV can be evaluated also through the Blood Volume Pulse (BVP) or Photoplethysmography (PPG).

A reduced HRV is linked to psychiatric illness as depression, anxiety. The heart rate is the most natural choice for arousal detection using comparison of sympathetic and parasympathetic frequency bands of the time series. However, it is highly dependent on the position of the body during monitoring.

4.1.5 Electrodermal Activity

As already been studied in Chapter 3, EDA measures the resistance of the skin and the skin conductivity applying electrodes to the skin. The skin conductivity decreases during relaxed states, and increase when exposed to effort.

4.1.6 Respiration

RSP is the process of moving air into an out of the lungs to facilitate gas exchange with the internal environment, mostly bringing in oxygen and flushing out carbon dioxide.
The respiration can be measured with a latex rubber band, the amount of stretch in the elastic is measured as a voltage change and recorded. The most common measures of RSP are the depth of breathing and the rate of RSP.

RSP rate generally decreases with relaxation, tense situations may result in momentary RSP cessation. Irregularity in the RSP pattern could be the cause of negative emotions.
Due to the fact that RSP is closely linked to the cardiac function, RSP can be affect other measures, like EMG and SC measurements.
Positions to the left, and typical waveform of the signals to the right are presented in the Figure 4.3.

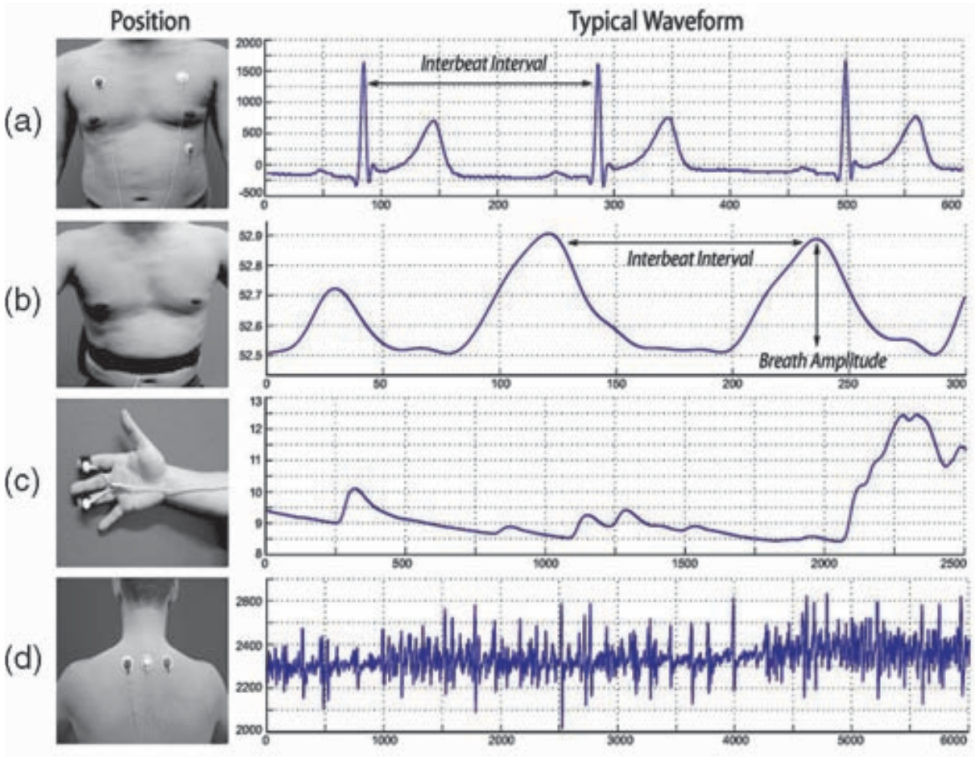


Figure 4.3: Positions (left) and waveform of the signals (right), (a) ECG, (b) RSP, (c) SC, (d) EMG

4.2 General methodology

For physiological signal-based emotion recognition, there is a common methodology which can be divided into two categories:

- Traditional ML methods: model specific methods, which require carefully designed hand-crafted features and feature optimization methods.
- Deep learning methods, which are model-free methods and can learn the inherent principle of the data and extract features automatically.

The whole emotion recognition framework is shown in Figure 4.4.

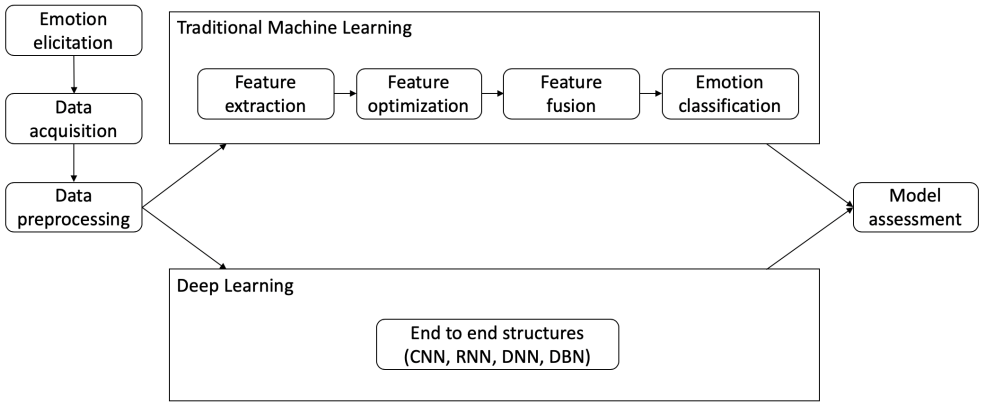


Figure 4.4: Emotion recognition process using physiological signals under target emotion stimulation

4.2.1 Preprocessing

After the part of data acquisition, which is different for each physiological signal, a data preprocessing step is usually performed. It is necessary to eliminate the noise effect, artifacts and other signal parts that may lead to wrong results. Due to the complex and subjective nature of raw physiological signals and the sensitivity to noises, electromagnetic interference, movement artifacts, ... this step is mandatory.

Some of the common steps can be summarized in:

- Filtering: is commonly used a low-pass filter to remove noises, or also adaptive band pass filters to remove artifacts.
- DWT: used to reduce the noise of physiological signals
- ICA: used to extract and remove respiration sinus arrhythmia from ECG.
- Empirical Mode Decomposition (EMD): used to remove the eye-blink from EEG.

4.2.2 Traditional Machine Learning

Main steps for traditional ML methods, as already presented in 2.6. There are processes including feature extraction, feature selection and classification as reviewed in [47].

Feature extraction

Feature extraction plays a fundamental role in the emotion recognition model. Several major features are extracted for each physiological signal, since it is important to extract the most prominent features for emotion recognition task. For example EEG is a complex and non-stationary signal, so some statistical features like Power Spectral Density (PSD) and Spectral Entropy (SE) are commonly used.
Often are extracted statistical features as mean, standard deviation, Kurtosis, Skewness, entropy.
However, each bio-signal has to be investigated separately as extracted features might vary in their usefulness for the classification of emotions.

Feature selection

After the feature extraction process, there might be a quantity of features, some of which may be irrelevant, some that are probably correlated each other, there might be some redundant features.
This lead to a longer time to analyze the features and train the model. This results in an overfitting problem and as consequence the decreasing of the model performance.
Some of the main feature selection algorithms are RfeliefF, mRMR, SBS and SFS, PCA, ...

In general there are several feature selection algorithms, some reduce the dimensionality by taking out some redundant or irrelevant features, other transform the original one into a new set of features. Performances of the feature selection algorithm depends on the classifier and the dataset, due to this, the perfect feature selection algorithm does not exist.

Classification

In emotion recognition, the major task is to assign the input signal to one of the given class sets. There are several classification models like Linear Discriminant Analysis (LDA), k-NN, SVM, Random Forest (RF), ...

4.2.3 Deep Learning

Deep Learning (DL) methods have the benefit to be model-free methods, so they do not depend on the specific model considered.
Examples of DL algorithms are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), ... NNs in general are particular types of ML methods which have as fundamental unit the node, loosely based on the biological neuron.

The relevant aspect of DL is that it can learn the inherent principle of the data and extract features automatically, so there is no need to extract features and select the most relevant ones, which could lead to a better generalization of the problem.

4.2.4 Model assessment and selection

The generalization error of the classifier can be evaluated by experiments, where a testing set should be used to test the ability of the classifier to classify the new samples, and the testing error on the testing set could be viewed approximately as the generalization error.
The dataset is divided into two mutually exclusive sets, the training and the testing set. It is important to maintain the consistency of the data distribution as much as possible. In general, the experiment has to be repeated several times with random division and then calculate the average value as the evaluation result.

To reinforce model assessment it is possible to use k-fold cross-validation, where the initial sampling is divided into K sub-samples. One sub-sample is used as the testing set and the other $K-1$ samples are used for training. Cross-validation is repeated K-times. Each sub-sample is validate once and the average result of K times is used as the final result. The most common is the 10-fold cross-validation.

To evaluate the performance of the experiment there is the accuracy, which is the proportion of samples that are correctly classified to the total samples.

4.3 Issues of physiological signals

A lot of efforts have been made in revealing the relationships between explicit physiological signals and implicit psychological feelings. However, there are still several challenges in emotion recognition based on physiological signals:

- Need a very well designed setup in order to obtain high quality physiological data. The standard setup is the standard lab setting with subjects with earphones sit motionless in front of a screen where the emotion stimuli materials are played. This system is fixed, data are noiseless and stable, but is hard to obtain genuine emotions.
- Stimulus materials are artificially selected and labels of the materials are manually set. The problem is that for the same stimulus, human emotion vary from each other. This could lead to a large deviation in the rating as explained in [48].
- There is not still clear evidence on which combination of physiological signals and extracted features is the most significant to emotion changes.
- For most studies, the number of subjects is small. Due to limited samples, the performance of the classifier with subjects who have not been analyzed during training would be poor. The clear efficient method is to include more subjects from different ages and backgrounds.
- Emotion perception and experience lead to strong differences.
- The reliability of facial expressions cannot be guaranteed sometimes.

4.4 Related works based on physiological signals

In this section some recent works regarding the relationship between EDA signals and human emotions will be presented.

4.4.1 ECG and GSR signal emotion recognition

In the work [42] ECG and GSR of 11 healthy students were collected while subjects were listening to emotional music clips. They extracted Matching Pursuit (MP) coefficients from ECG and GSR signals.

Then, a set of statistical indices are extracted from MP coefficients and three dimensionality reduction methods has been applied, like PCA, kernel PCA and LDA. These features were fed into the Probabilistic Neural Network (PNN) in subject-dependent and subject-independent modes.

The PNN is fed with a feature vector in the input layer, then the distance between input and a the weight vector is determined. The summation of these contributions is computed for each input class to yield the probability. Finally, the maximum of the resulting probabilities is selected by a competitive layer and the label 1 is assigned to the class that produces the maximum, 0 otherwise.

Using PNN was achieved the highest recognition rate of 100%.

4.4.2 ECG sensors for human emotion recognition

In the work presented in [36], it is suggested that ensemble learning approach for developing a machine learning model that can recognize four major human emotions (anger, sadness, joy and pleasure) incorporating ECG signals.

As feature extraction method, the analysis combines four ECG signals techniques. Several ML methods have been applied to the model, the most accurate (with an accuracy of 70%) is achieved by using an Extra Tree Classifier (a variant of the RF that introduce more variation in the ensemble).

4.4.3 Automatic ECG emotion recognition

An automatic ECG-based emotion recognition algorithm is presented in [37]. They recorded ECG signal from subjects and extracted some features from the signal from the time and frequency domain. Then, performed an algorithm of feature selection, a sequential forward floating selection-kernel-based class separability-based.
Valence and arousal and four types of of emotions are recognized using Least Square-SVM recognizer. They gained a classification rate for

positive/negative valence, high/low arousal, and four types of emotion classification tasks are 82.78%, 72.91%, and 61.52%, respectively.

4.4.4 Classification of music emotions with forehead biosignals and ECG

In the work presented in [38], a fusion of three-channel (left and right temporal channel and frontalis) and ECG are used to recognize music-induced emotions. They employed two parallel SVM as arousal and valence classifiers.

The inputs of the classifiers were obtained by applying a fuzzy-rough model feature evaluation criterion and sequential forward floating selection algorithm.
The average classification accuracy is 88.78% (valence classification accuracy of 94.91% and arousal classification accuracy of 93.63%).

4.4.5 Emotion classification with forehead biosignals

The work [39] the feasibility of using 3-channel forehead biosignals is investigated. Classification in valence arousal space is performed by employing two parallel cascade-forward NN.
The inputs of the classifiers were obtained by applying a fuzzy rough model feature evaluation criterion and sequential forward floating selection algorithm. An averaged classification accuracy of 87.05% was achieved, corresponding to average valence classification accuracy of 93.66 % and average arousal classification accuracy of 93.29 %.

4.4.6 Physiological changes in music listening

The paper [41] investigates the potential of physiological signals as reliable channels for emotion recognition. Were used four-channel biosensors to measure EMG, ECG, SC and RSP changes, as can be seen in Figure 4.3. They extracted some features in various analysis domain i.e. the time/frequency, entropy, geometric analysis, subband spectra.

Classification of four musical emotion (one for each quadrant of the valence-arousal diagram) is performed by using an extended LDA (pLDA). They also provided a novel scheme of emotion-specific multilevel dichotomous classification, gaining an accuracy of 95% for subject-dependent and 70% for subject-independent classification.

4.4.7 NN based emotion estimation

In order to build a human-computer interface that is sensitive to a user's expressed emotion, in [43] a NN based emotion estimation algorithm is proposed, using HRV and GSR. In this study, a video clip method was used to elicit basic emotions from subjects while ECG and GSR signals

were measured. These signals reflect the influence of emotion on the autonomic nervous system. The extracted features that are emotion-specific characteristics from those signals are applied to an artificial neural network in order to recognize emotions from new signal collections. Results show that the proposed method is able to accurately distinguish a user's emotion.
They gain a total accuracy of 80.2%.

4.4.8 Recognize emotions by affective sound through HRV

The research in [45] reports on how emotional states elicited by affective sounds can be effectively recognized by means of estimates of ANS dynamics.

The ANS dynamics is estimated through standard and nonlinear analysis of HRV exclusively, which is derived from the ECG. A group of 27 people were administered with ECG recordings, then HRV features showing significant changes between valence and arousal dimensions were used as input of an automatic classification system.

The best accuracy was achieved for a quadratic discriminant classifier, to 84.72% on the valence dimension and 84.26% on the arousal one.

4.4.9 Emotion recognition from ECG

In [40] carried out the work of affective ECG signal acquisition from 391 subjects through stimulation of film clips. They recognized emotions divided into Joy and Sadness.
Then, features extraction and feature selection algorithms based on the DWT and a Fisher-k-NN are implemented to classify the test data.

4.4.10 Relationship between music emotion and physiological signals

In [7] the study explores the possibility of using physiological signals to detect users emotion response to music, considering individual characteristics (as personality, music preferences, etc.).

A user experiment was conducted with 23 participants, during music listening, a series of physiological signals like HRV and SC were recorded using a wearable wristband.
Here, arousal and mood values rated by participants were grouped into three main categories (i.e. positive, negative, neutral), for mood ratings, they combined the mood categories into positive, negative and neutral moods.

After some data preprocessing, were extracted the features in Table 4.3.

Category	Features
Descriptive statistics of raw signal	Mean, Standard deviation, median, range
Time series features	Means of the abs of the $1^{st}/2^{nd}$ differences of the raw/normalized signals
Physiological signal specific features	SCR, HRV

Table 4.3: Features extracted from physiological signals in [7]

A ML approach was applied to measure the extent to which physiological signals could be used to recognize users' emotion responses to music listening, in positive and negative categories of arousal and mood. Specifically, they trained and compared the performance of several classification models, namely decision tree, k-NN, naïve Bayes and SVM.

4.5 Related works based on EDA signals

In this section some recent works regarding the relationship between EDA signals and human emotions will be presented.

4.5.1 DL model for human emotion recognition with EDA

The work in [49] had the main objective of ensure that elderly and/or disabled people perform/live well in their immediate environments. This can be monitored by among others the recognition of emotions based on non-highly intrusive sensors such as EDA sensors.
However, designing a learning system or building a machine-learning model to recognize human emotions while training the system on a specific group of persons and testing the system on a totally a new group of persons is still a serious challenge in the field, as it is possible that the second testing group of persons may have different emotion patterns.

They contributed to the field of human emotion recognition by proposing a CNN architecture which promises robustness for both subject-dependent and subject independent human emotion recognition.
Subject-independent emotion recognition is a challenging field due to:

- Physiological expressions of emotion depend on age, gender, culture and other social factors.
- Depends on the environment in which a subject lives.
- The lab-setting independent nature of emotion recognition is related to the fact that the classifier can/will be trained locally once using sensors of a given lab-setting and after that tested considering different datasets that are collected based on different lab settings.

Authors converted EDA signals into matrices whereby the goal is to make the application of CNN model possible.
They tested the CNN on two datasets, MAHNOB and DEAP, which are four-classes labeled and they increased the accuracy up to 78% for MAHNOB and 82% for DEAP in subject-independent classification, while up to 81% for MAHNOB and 85% for DEAP in subject-dependent classification.

4.5.2 VA recognition of affective sounds based on EDA

In [25] tried to automatically classify the emotional state of healthy subjects. They proposed the use of convex optimization based on EDA framework and clustering algorithms to automatically discern arousal and valence levels induced by affective stimuli.

EDA recordings were gathered from 25 healthy volunteers, using only one EDA sensor to be placed on fingers.
In model-based approaches, models describe and estimate the underlying psychological process that generates the observed data (EDA measurements).
The model based analysis of EDA has fundamental advantages, such as a propensity to reduce the effects of measurement noise and the essential ability to improve the temporal resolution of inference in rapid event-related paradigms.

EDA data, in this experiment was analyzed with the *cvxEDA*[1] algorithm, presented in [50], which proposed a representation of the SCR parts of EDA as the output of a linear time-invariant system to a sparse non-negative driver signal.

They extracted several features form EDA both from phasic and tonic components output of the *cvxEDA*. For each feature, two levels of valence (positive and negative) and three levels of arousal (low, medium and high) were compared.
The supervised classification was implemented using a k-NN classifier.

Results, thanks to *cvxEDA* showed a recognition accuracy of 80 % on the arousal dimension and 84 % in valence classification.

[1] https://github.com/Iciti/cvxEDA

4.6 Conclusions

In this chapter we have shown a review of the state of the art about human emotion recognition based on physiological data. A schematic block of the general algorithm can be seen in Figure 4.4.
Summing up, the majority of the works deal with a small number of subjects (25/30 maximum) and they evaluated their accuracy based on valence-arousal space grouped in four main labels as schematized in Figure 4.5.

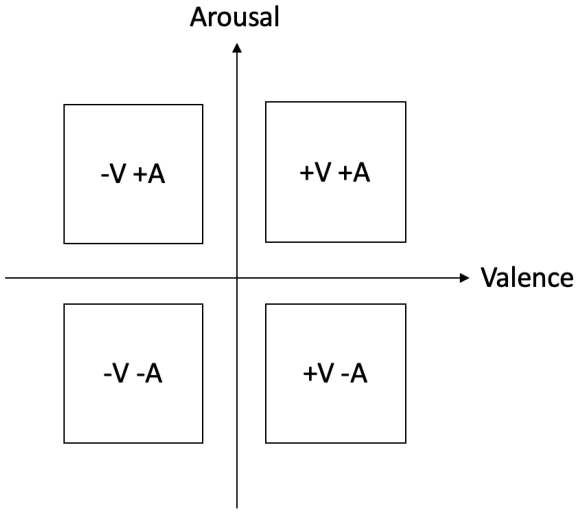


Figure 4.5: Valence and Arousal space divided in four main groups

5 Proposed Framework

In this chapter is presented the overall description of the dataset used in the experiment and the related work already done.

5.1 PMEmo dataset

This thesis is based on the paper *The PMEmo dataset for Music Emotion Recognition* [8]. K. Zhang, H. Zhang and S.Li created a novel dataset called Popular Music with Emotional annotations (PMEmo) containing emotions of 794 songs as well as EDA signals.

Most researchers working on MER adopt methods in supervised ML to implement music emotion prediction [51], which usually need a large number of songs with emotion labels provided by listeners to train the models.

A musical experiment was well-designed for collecting the affective annotated music corpus oh high quality, which recruited 457 subjects. The dataset (about 1.3Gb) is publicly available to the research community at this link[1].
It is intended for benchmarking in MIR and MER, it involves precomputed audio features sets and manually selected chorus excerpts (in .mp3) of songs, to facilitate the development of chorus-related research.

5.1.1 Dataset structure

The dataset contains 794 music clips annotated by 457 subjects, where participants come from different countries and majors, in order to elimi-

[1]https://drive.google.com/drive/folders/1qDk6hZDGVlVXgckjLq9LvXLZ9EgK9gw0

nate the effects of cultural and educational background [52].
Chorus parts are manually selected from students majoring in music.

Meanwhile, the EDA of subjects when listening to these music pieces are also recorded, making it possible to analyze emotion states in multiple modes. All annotations are stored in CSV files delimited by comma. The dataset is composed of:

- annotations: valence and arousal values for each song. There are:
  - *static_annotations*: valence and arousal standard deviation values for each song, one value for each song.
  - *static_annotations_std*: valence and arousal mean values for each song, one value for each song.
  - *dynamic_annotations*: valence and arousal standard deviation values for each song, acquired at a sampling rate of $2Hz$.
  - *dynamic_annotations_std*: valence and arousal mean values for each song, acquired at a sampling rate of $2Hz$.
- chorus: all chorus excerpts of 794 songs manually selected.
- comments: songs comments taken from NetEase[2] and SoundCloud[3].
- EDA: EDA data for each song, each one extracted by at least 10 subjects, with a sampling rate of $50Hz$.
- features: all features extracted by the authors of [8]:
  - *EDA_features*: features extracted from the EDA signals:
    - *EDA_features_static*: EDA static features for each song for each subject.
    - *EDA_features_dynamic*: EDA dynamic features for each song for each subject with a sampling rate of $50Hz$.
    - *static_features*: audio static features for each song.
    - *dynamic_features*: audio dynamic features for each song with a sampling rate of $50Hz$.
- *lrc_dataset*: lyrics text of all music excerpts.
- *lyrics*: lyrics text of all music excerpts divided by each timestamp.
- *metadata*: metadata of the songs, containing *music_ID*, title, artist, album, duration, *chorus_start_time* and *chorus_end_time*.

[2]https://music.163.com
[3]https://soundcloud.com

Since the early years of MER, there have been numerous efforts to build datasets with emotional annotations, to facilitate the development and evaluation of music emotion recognition. Table 5.1 from [8] summarize some works on that.

Name	Stimulus	Data	Audio
Emotify[4]	400 excerpts	induced emotion	yes
Moodswing[5]	240 excerpts (30s)	valence and arousal	no
Amg1608[6]	1608 excerpts (30s)	valence and arousal	no
emoMusic[7]	744 excerpts (45s)	valence and arousal	yes
DEAM[8]	1802 excerpts	valence and arousal	yes
SoundTracks[9]	360 + 110 excerpts	valence, energy, tension, mood	yes
GMD[10]	1400 songs	genre, valence and arousal	yes
DEAPDataset[11]	120 music excerpts	valence, arousal, dominance and physiological data	no
PMEmo	794 music chorus	valence, arousal and physiological data	yes

Table 5.1: Some existing music datasets with emotion annotations from [8]

PMEmo is a valid alternative to the datasets listed in Table 5.1 due to the fact that it is wide enough compared with the others and it gives the possibility to apply ML methods.
PMEmo is also great because it gives original audio files, chorus parts manually selected.

5.1.2 Song acquisition and subject selection

They collected 1000 songs from the "*Billboard Hot 100*", the "*iTunes top 100 songs*" and the "*UK top 40 single charts*". They late discovered a set of duplicates and filtered double music obtaining a full song set of 794 pop songs.
Each datasets in MER utilize music segments, here each clip is manually selected as one of the chorus parts of each song, which is implemented

[4] http://www.projects.science.uu.nl/memotion/emotifydata/
[5] http://music.ece.drexel.edu/research/emotion/moodswingsturk
[6] https://amg1608.blogspot.ca/
[7] http://cvml.unige.ch/databases/emoMusic/
[8] http://cvml.unige.ch/databases/DEAM/
[9] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/pastprojects/coe/materials/emotion/soundtracks/Index
[10] https://hilab.di.ionio.gr/en/music-information-research/
[11] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html

by university students in music major. The clips are of various length, exactly the duration of the chorus parts.

A total of 457 subjects, 236 females and 221 males were recruited to participate. Among them, 366 are Chinese university students who were in non-music majors while 44 were majoring in music recruited to ensure high quality labeling. To weaken the impact of cultural background, 47 English speakers were invited to annotate the datasets.
Each song received a total of at least 10 emotion annotations including one by music-majoring and one by English speaker.

5.1.3 Experiment design

To monitor and obtain EDA continuously they used MP150 Biopac system[12] with a sampling rate of $50Hz$ and export signals from *AcqKnowledge* software.
To annotate songs was developed a desktop application shown in Figure 5.1.

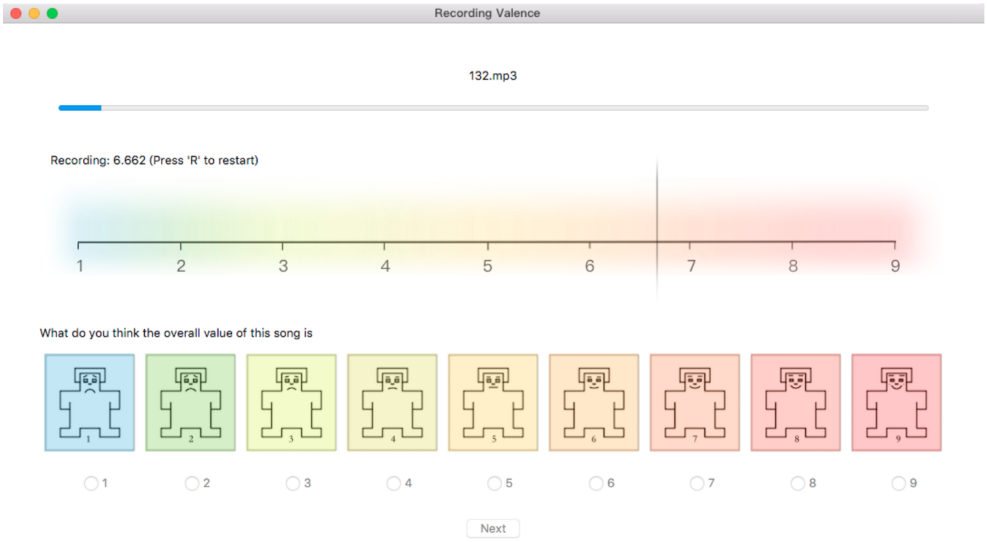


Figure 5.1: Annotation interface for PMEmo

The annotation was done with the sliding area collecting dynamics annotations , from 1 to 9, at a sampling rate of $2Hz$. Annotators should make a statistic annotation for the whole music excerpts on nine-point scale after the dynamic labeling. Furthermore, annotators were asked to listen to the same music twice to annotate on valence and arousal separately.

[12]https://www.biopac.com/product/eda-finger-transducer-bsl/

In Figure 5.2 is shown the flow diagram of the experiment, where each subject spent 50 minutes on average.

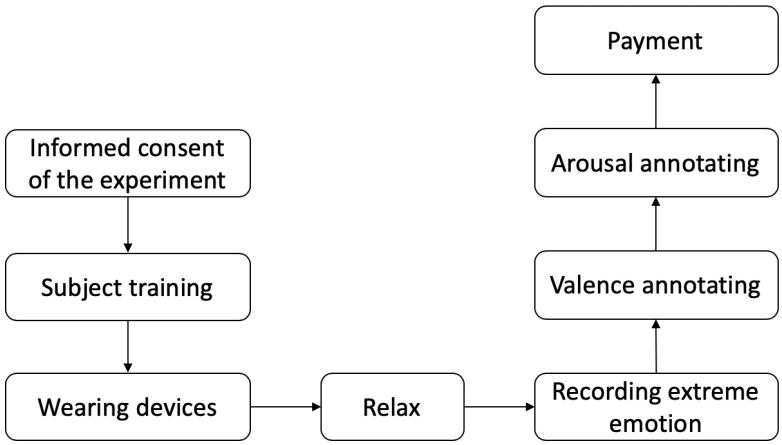


Figure 5.2: Experimental procedure for PMEmo

Each subject listened to 20 excerpts and one of those was duplicated to guarantee the high quality data as what was done in [53]. The annotations from this subjects were accepted only if the bias between duplicate clips were within 0.25 in the VA space (they did not inform subjects about the duplicated excerpt).
In total 457 subjects have participated but 401 were considered valid annotations (87.7%). Each music clip was annotated by at least 10 subjects including English speakers and semi-experts from music academy.

5.1.4 Data reliability

Annotators need some preliminary time before they can give meaningful and reliable annotations, this is called Initial Orientation Time (IOT). Schubert in [54] found that median IOT for valence was $8s$ while for arousal $12s$. Other researchers showed that annotations began to converge after $10s$. The PMEmo authors decided to discard first $15s$ for the dynamic annotations from the data.

To evaluate annotation consistency they used the Chronbach's α, which represents the degree to which a set of items measures a single unidimensional latent construct. In [8] computed the Chronbach's α on the sequence of annotations for each song.
They processed annotations by:

$$a_{j,i} = a_{j,i} + (\bar{A}_j - \bar{A}) \tag{5.1}$$

where:

- $a_{j,i}$ is the label annotated by subject j at time i
- $\bar{A}$ is the mean of all the labels for this song by all subjects

- $\bar{A}_j$ is the mean of dynamic labels by subject j

The mean (averaged across songs) and the standard deviation of the Chronbach's α for the annotation in the PMEmo dataset are shown in Table 5.2.

Dimension	Mean	Std Dev
Valence	0.998	0.005
Arousal	0.998	0.008

Table 5.2: Mean and standard deviation of the Chronbach's α for PMEmo dataset annotations

5.1.5 Feature set

As already mentioned before in Chapter 2.5, for generic MER there has been no attempt made at defining a "standard" feature set. In PMEmo work [8] they based on the INTERSPEECH 2013 Computational Paralinguistic Evaluation (ComPaRe) [55] and extracted a feature set of 6373-dimension scale.
They provided all 6373-dimension features in song level for the sake of static emotion task. They extracted only the core of 260-dimension features in segment level (calculated in $1s$ window with $0.5s$ overlap) for dynamic recognition task to properly reduce the computing load.
Extraction of the features is done with the open-source toolkit openSMILE[13] [56].
No feature selection procedure was implemented in PMEmo work.

[13]https://www.audeering.com/opensmile/

5.2 General framework

We decided to start from the results of PMEmo and try to improve them. We will first look at a general framework description and then will follow explanation of the single parts.
The main idea of the thesis is to find if there is a real connection between audio and EDA signals, if emotions *felt* during the listening of the music are correlated to the ideal emotions extracted from the audio data, the *perceived* emotions.
Our goal is to improve PMEmo work, focusing on the lacks of PMEmo work in [8]. Their main lacks are:

- The use of a huge number of features, because this worsens the performances of the model predictor. Considering too many features this cause an overfitting in the model, because many of them could be strongly correlated and just worst the model. Having a feature vector of 6373-dimension is definitely too large.
- Related to the previous point, they used audio features that are automatically extracted from a software, open SMILE, which was created more for the analysis of the speech, unlike for audio.
- No feature selection method was applied to the feature space, they brought all the features as an input to the ML model.

In the following section we introduce the reader to the framework we decided to implement, based on the idea of resolving PMEmo lacks listed before. For example, we extracted audio features that are more audio-related, starting from audio low level descriptors as tempo and beats and moving to higher level features as harmony.

Another improvement we implemented was to add a feature selection step before passing to the ML model. We deployed different algorithms of feature selection, in order to use the one that best fits with our model and gives best results.

As already presented in Chapter 4.2 in Figure 4.4, it is possible to infer a model analyzing the emotion elicitation. Since we will try to deal with both audio and EDA data, we decided to develop a model based on a traditional ML framework.
General framework implemented is represented in Figure 5.3. Differently from Figure 4.4, here, the starting point is dataset files given from the PMEmo dataset.

Figure 5.3: General framework

The input of this process is the PMEmo dataset, and we starts from

taking audio files from all the songs, and also EDA files from every subject in the experiment.
Will follow now, a list of symbols and terminology used in the explanation of the framework:

- $\mathbf{A}$ is the vector of all the audio files contained in the dataset.
- a_n is the n^{th} song excerpt of the vector $\mathbf{A}$.
- $\mathbf{E}$ is the vector of all the EDA files contained in the dataset.
- e_n is the n^{th} EDA signal of the vector $\mathbf{E}$.
- $\mathbf{F_A}$ is the matrix of features related to audio, while $\mathbf{F_E}$ is the matrix of features related to EDA.
- f_A^n is the vector of the n^{th} feature contained into the matrix $\mathbf{F_A}$.
- f_E^n is the vector of the n^{th} feature contained into the matrix $\mathbf{F_E}$.
- $\mathbf{F_A^*}$ is the matrix of audio features that are extracted after a feature selection method.
- $\mathbf{F_E^*}$ is the matrix of EDA features that are extracted after a feature selection method.
- The sum of $\mathbf{F_A^*}$ and $\mathbf{F_E^*}$ is denoted as $\mathbf{x}$, as the vector of audio and EDA features.
- The output variable of the system is Y, which is composed of the Valence and the Arousal values, $Y = (Y_a, Y_v)$ and the objective of the model is to estimate $mean(Y_a)$, $mean(Y_v)$, $std(Y_a)$ and $std(Y_v)$.

From the dataset containing audio and EDA files, we extracted a certain number of features, creating a matrix for all the features extracted. This is the way to represents the input data to the ML model.
From the feature matrix, several feature selection algorithms were applied in order to reduce the number of features, to improve the model, because having too many features it will cause an overfitting in the model and worst overall performances.
At last, some ML process is applied to create the model for the emotion classification task and then it is tested. This is the main step, that creates the bridge from the input data to the output, the step that gives the real answer.
In this dataset were present two different types of data, audio and EDA and our main goal was to combine:

- audio data connected to perceived emotions
- EDA data related to felt emotions

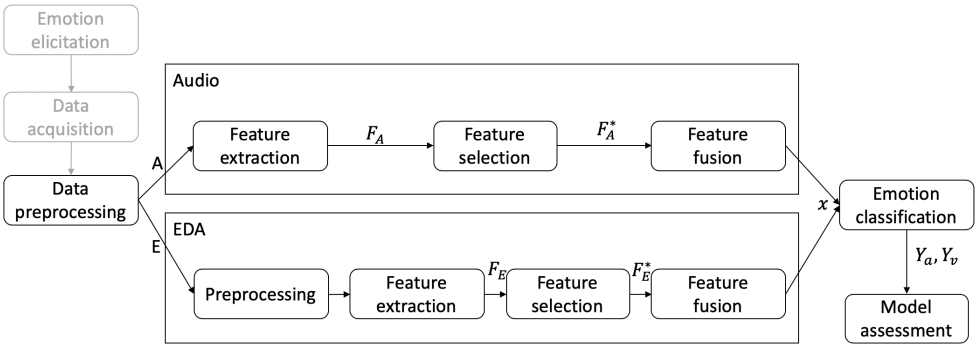


Figure 5.4: General framework with audio and EDA division

The general framework, with these two process divided is shown in Figure 5.4 which is based on the one presented before at Figure 5.3.
As can be seen in Figure 5.4, the first step is to take from the PMEmo dataset **A** and **E**, this will be the starting line of the whole process, represented in Figure 5.4 as the first block of *Data preprocessing*.
Now, audio and EDA will be treated as two separate parts that will be fused together in the last part of the process.
EDA features need a preprocessing step, which will be explained later on.
After this preprocessing step, both on **A** and **E** are extracted features in the *feature extraction* step ending up with $\mathbf{F_A}$ and $\mathbf{F_E}$. This step is needed because the raw input data is often too large, noisy and redundant for analysis.
It follows the step of *feature selection* both for audio and EDA files, which takes as input $\mathbf{F_A}$ and $\mathbf{F_E}$ and with different algorithms gives as output subsets of $\mathbf{F_A}$ and $\mathbf{F_E}$, called $\mathbf{F_A^*}$ and $\mathbf{F_E^*}$.
Than, $\mathbf{F_A^*}$ and $\mathbf{F_E^*}$ are fused together and they are sent to the emotion classification step, which using ML models is able to create a model assessment.

In the next sections we are going to present all the blocks of this general framework, starting with feature extraction. In particular, several features can be extracted from audio, thanks to the hard work done in the past years. On the other side, EDA features are less common in the literature and they are mostly statistical features.

5.3 Audio feature extraction

In this section, for explaining audio feature extraction, we will use as an example song, a_4 under the name *4.mp3*. Its waveform and audio specs can be seen in Figure 5.5.

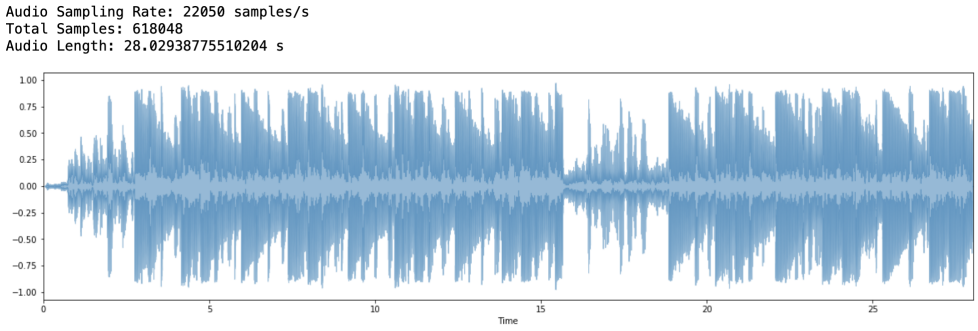


Figure 5.5: Waveform and audio specs of the song number 4

Features were extracted both in a static way, taking into account the whole excerpt and, in a dynamic way, by dividing the musical excerpt in windows of $1s$ with 50% overlap.
After the process of feature extraction, every feature was normalized in the range $[0, 1]$.
Audio features extracted can be grouped into:

- Temporal features: Tempo, Beats and Zero Crossing Rate.
- Chroma features: Chroma STFT, Chroma cqt and Chroma cens.
- Spectral features: Spectral contrast, centroid, bandwidth, rolloff, poly.
- Cepstrum features: MFCCs, tonal centroid.

5.3.1 Tempo

In musical terminology, tempo is the speed or pace of a given piece. In classical music, tempo is typically indicated with an instruction at the start of a piece and is usually measured in beats per minute (or bpm).
Tempo for a_4 is about $152bpm$.

5.3.2 Beats

To extract beats is used a beat extractor, which output is an estimation of the tempo and an array of frame numbers corresponding to detected beat events.
Beats are detected in three stages:

1. Measure onset strength

2. Estimate tempo from onset correlation

3. Pick peaks in onset strength approximately consistent with estimated tempo

In Figure 5.6 it is shown beats detected and the array of frame numbers.

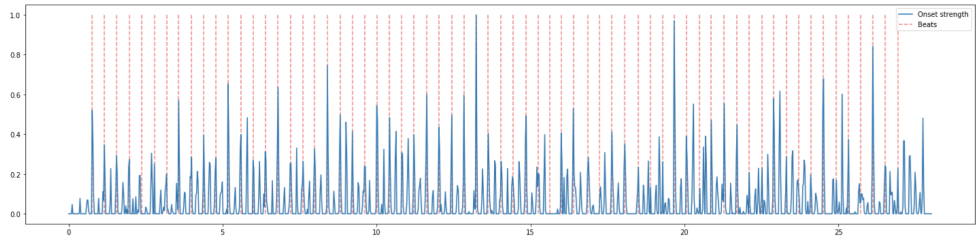


Figure 5.6: Beat and array of frames of a_4

5.3.3 Zero crossing rate

The Zero Crossing Rate (ZCR) is the rate of sign-changes along a signal. It is the rate at which the signal changes from positive to zero to negative or to negative or from negative to zero to positive. It is useful to recognize percussive sounds. The ZCR is evaluated as:

$$ZCR = \frac{1}{T}\sum_{t=1}^{T-1}\frac{|sign(s_t)| - |sign(s_{t-1})|}{2} \tag{5.2}$$

where T is the length of the time window, s_t is the magnitude of the t^{th} time domain sample.
The ZCR of song 4 is shown in Figure 5.7.

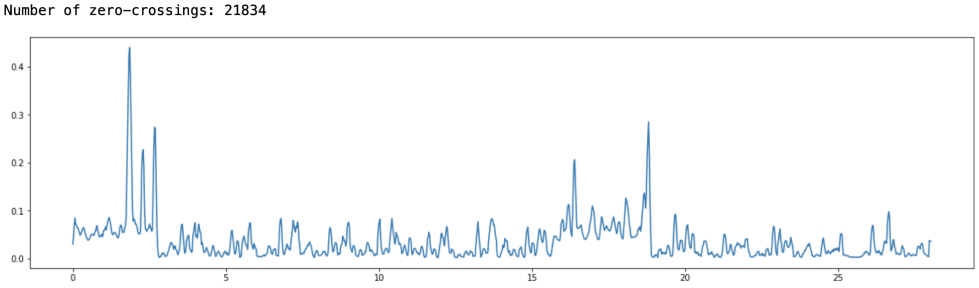


Figure 5.7: ZCR extracted from song a_4

ZCR is computed on windows for A and then mean, std and variance are computed.

5.3.4 Chroma

The chroma feature, also called chromagram, relates to the twelve different pitch classes. Chroma features capture harmonic and melodic characteristic of music, while being robust to changes in timbre and instrumentation.
Humans perceive two musical pitches as similar if they differ by an octave. A pitch can be separated into two components, referred as *height* (the octave where the pitch is) and *chroma*. The twelve chroma values are represented by the set:

$$C, C\#, D, D\#, E, F, F\#, G, G\#, A, A\#, B$$

that consists of the twelve pitch spelling attributes as used in Western music notation.
In the Figure 5.8 from [4] it is shown a chromagram (b) obtained from the score (a) and a chromagram (d) obtained from an audio recording of the C-major scale played on a piano (c).

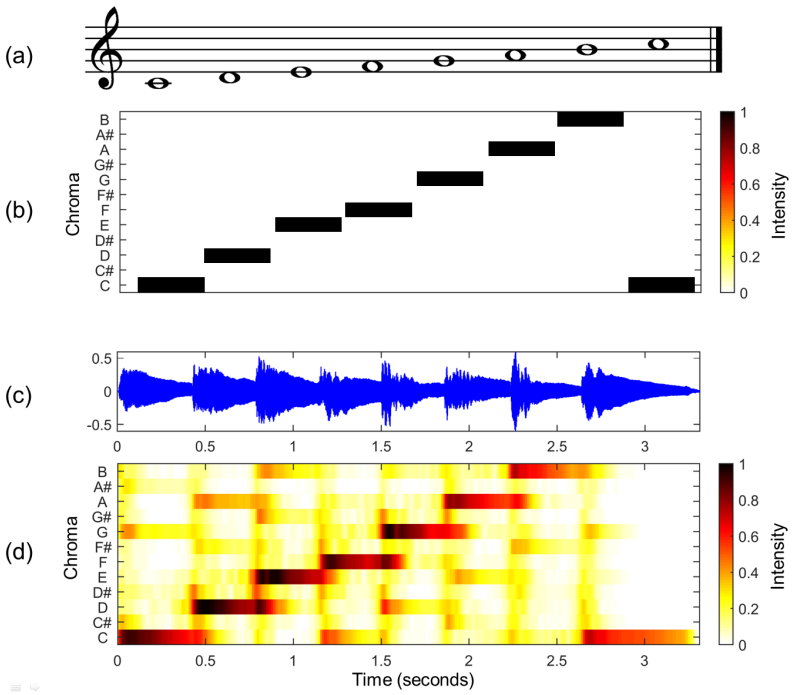


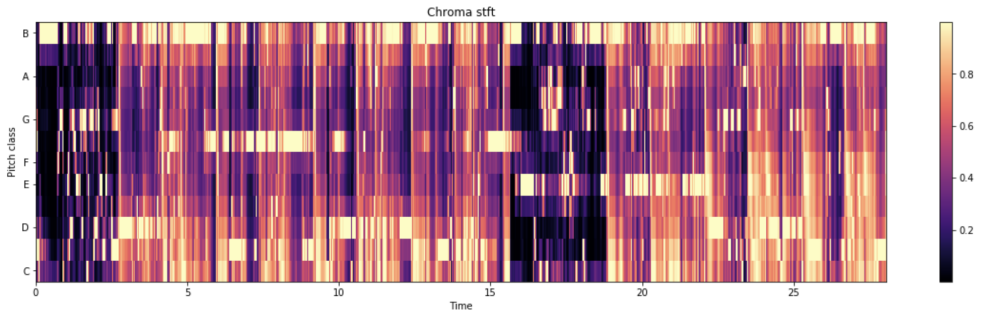
Figure 5.8: (a) Musical score of a C-major scale, (b) Chromagram obtained from the score, (c) audio recording of the C-major scale played on a piano, (d) chromagram obtained from the audio recording from [4]

There are different ways to convert an audio recording into a chromagram, as performing Short Time Fourier Transform (STFT) in combination with binning strategies or using multirate filter banks.

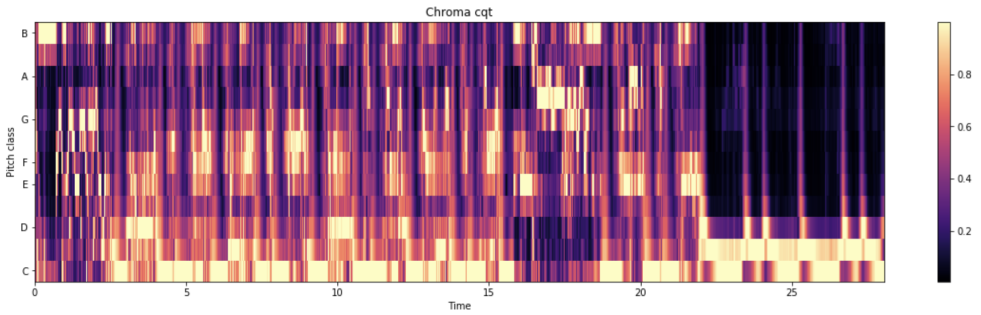
Chroma features can be significantly changed by introducing preprocessing and post-processing steps that modify spectral, temporal and dynamical aspects. This leads to a large number of chroma variants.
For the chromagram, we extracted three different types of chroma:

- Chroma STFT, which is obtained through the STFT.
- Chroma cqt, extracted using the constant-Q transform.
- Chroma cens which consider short-time statistics over energy distribution. In Chroma Energy Normalized Statistics (CENS) features, a temporal smoothing is introduced.

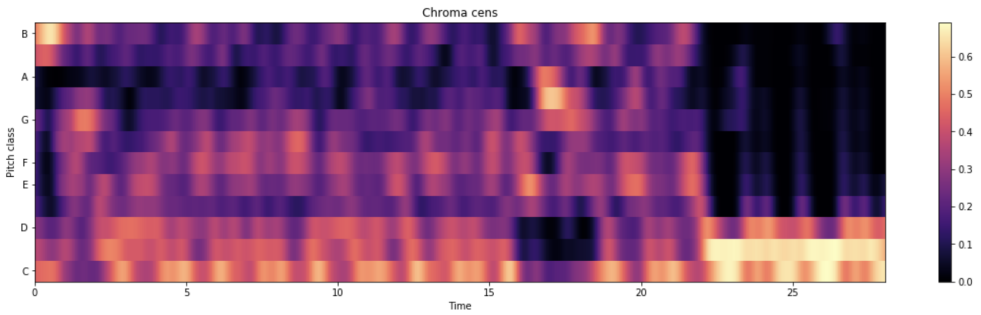
Three different chromagrams are shown in Figure 5.9 for a_4.



(a) Chroma stft



(b) Chroma cqt



(c) Chroma cens

Figure 5.9: Different chromagram extracted from song number 4

We compute chromagram on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.5 Spectral contrast

Spectral contrast divide the signal in sub-bands, and from each sub-band it works with peaks and valleys.

More in general spectral peaks correspond to harmonic components and spectral valleys correspond to non-harmonic components or noise in a music piece as evaluated in [57].
Therefore, the difference between spectral peaks and spectral valleys will reflect the spectral contrast distribution.
The spectral contrast of a_4 is shown in Figure 5.10.

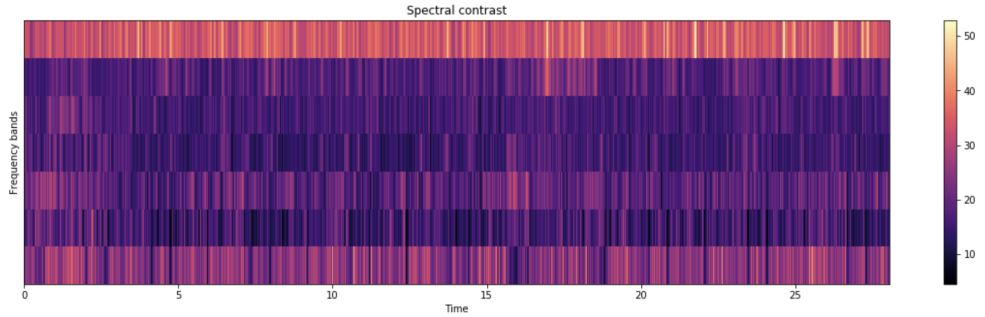


Figure 5.10: Spectral contrast extracted from a_4

We compute spectral contrast on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.6 Spectral centroid

Spectral centroid is a measure to characterize a spectrum. It indicates where is located the center of mass of the spectrum. It has connection with the brightness of a sound.
Spectral centroid is calculated as a weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights:

$$cent = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{5.3}$$

where $x(n)$ is the weighted frequency value (or magnitude) of bin number n and $f(n)$ represents the center frequency of that bin.
The spectral centroid of a_4 is shown in Figure 5.11.

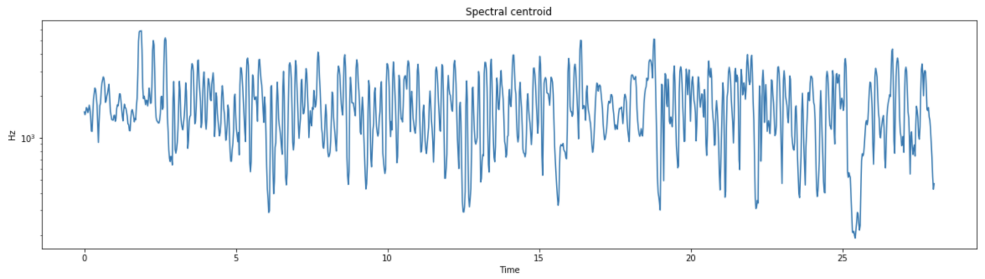


Figure 5.11: Spectral centroid extracted from a_4

We compute spectral centroid on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.7 Spectral bandwidth

The spectral bandwidth is the order-p spectral bandwidth as:

$$\left(\sum_k S(k)(f(k) - f_c)^p\right)^{1/p} \tag{5.4}$$

where $S(k)$ is the spectral magnitude at frequency bin k, $f(k)$ is the frequency at bin k and f_c is the spectral centroid.
The spectral bandwidth of a_4 is shown in Figure 5.12.

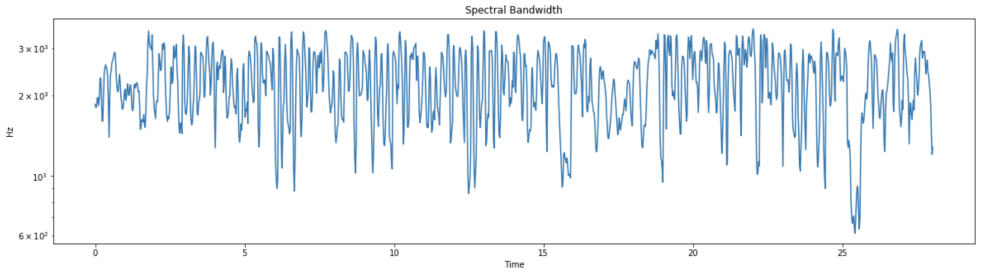


Figure 5.12: Spectral bandwidth extracted from song 4

We compute spectral bandwidth on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.8 Spectral rolloff

Spectral rolloff is defined as the N^{th} percentile of the power spectral distribution, where N is usually 85% or 95%. The rolloff point is the frequency below which the N of the magnitude distribution is concentrated. This can be used to, e.g., approximate the maximum (or minimum) frequency by setting roll_percent to a value close to 1 (or 0).
The spectral rolloff of a_4 is shown in Figure 5.13.

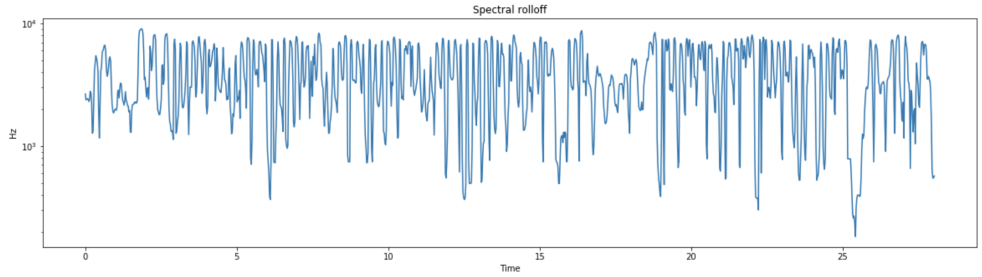


Figure 5.13: Spectral rolloff extracted from a_4

We compute spectral rolloff on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.9 Spectral poly

Get coefficients of fitting an n^{th} order polynomial to the columns of a spectrogram.
In the Figure 5.14 can be seen different poly extracted from a_4, with different degrees, 0-order fit a degree-0 polynomial (constant) to each frame, 1-order fit a linear polynomial to each frame and 2-order fit a quadratic to each frame.

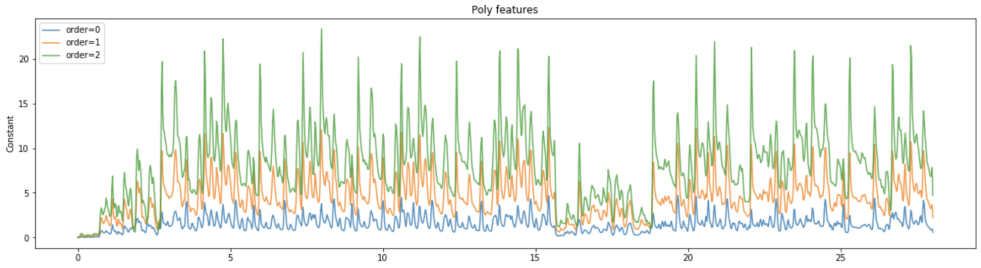


Figure 5.14: Spectral poly extracted from a_4

We compute spectral poly on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.10 Tonal centroid

Computes the tonal centroid features (tonnetz), following the method of [58].
In musical tuning and harmony, the Tonnetz, is a conceptual lattice diagram representing tonal space first described by Euler in 1739. Various visual representations of the Tonnetz can be used to show traditional harmonic relationships in European classical music.
Close harmonic relations are modeled as short distances on an infinite Euclidian plane. Chords become geometric structure on the plane and chords become geometric structures on the plane, keys are defined by regions in the harmonic network
An example is shown in Figure 5.15.
The Tonnetz of the a_4 is shown in Figure 5.16.

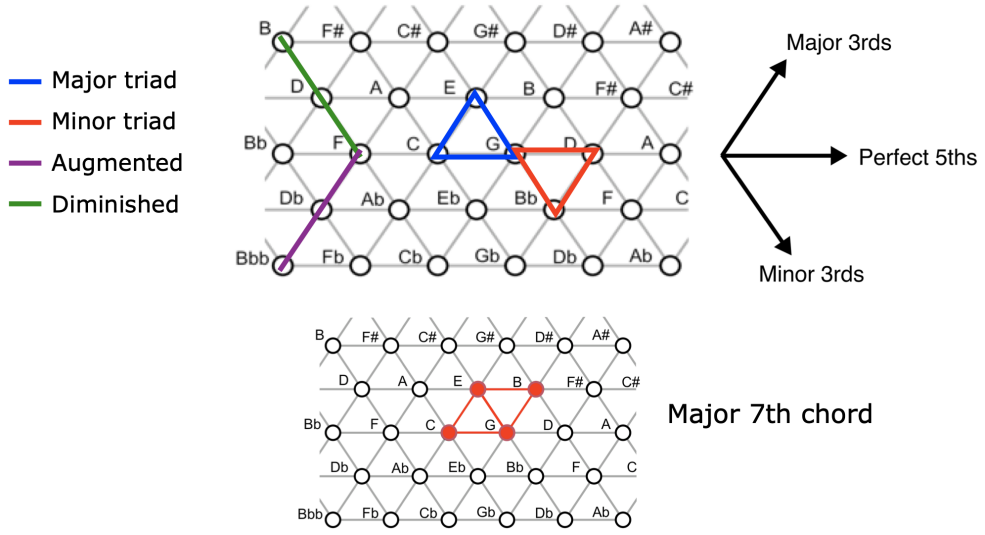


Figure 5.15: Representation in the Euclidian plane of the tonnetz

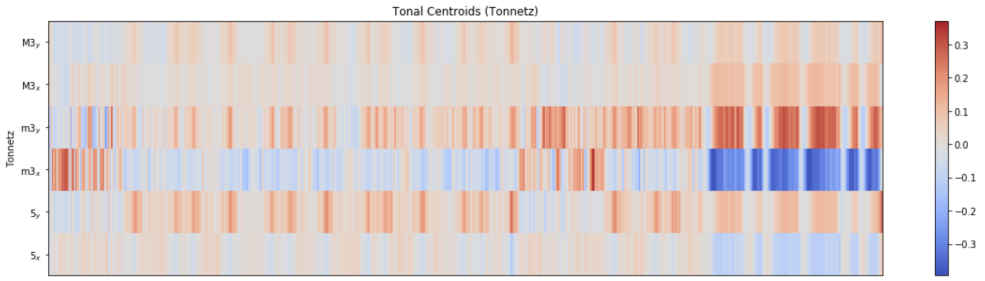


Figure 5.16: Tonnetz extracted in a_4

5.3.11 Melspectrogram

The melspectrogram is a mel-scaled spectrogram. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. It can be generated by a bank of band-pass filter, by Fourier transform or DWT.

In order to have a more comprehensible spectrogram, when dealing with audio signals, it is scaled. The axis representing the frequencies is transformed to log scale, and the *color* axis representing the amplitude, is represented in Decibels.

The Mel-scale is a different scale, based on non-linear transformation of the frequency scale. It is constructed such that sounds of equal distance from each other on the Mel-scale also *sound* to humans as they are equal in distance from one another.

In practice it partitions the Hz scale into bins, and transforms each bin into a corresponding bin in the Mel Scale, using a overlapping triangular filters.

To convert a frequency in Hz into its equivalent in *mel*, the following

formula is used:

$$pitch[mel] = 1127.0148 \log \left[1 + \frac{f}{100}\right] \tag{5.5}$$

Finally, a melspectrogram is a spectrogram properly filtered such that the frequency axis is in mel-scale.
The melspectrogram of a_4 is shown in Figure 5.17.

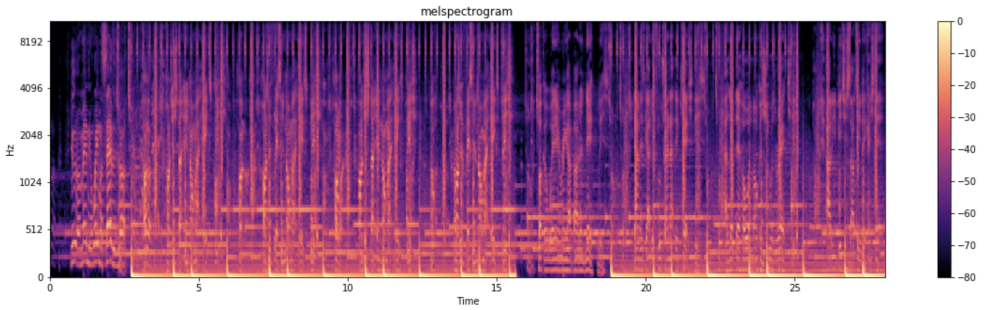


Figure 5.17: Melspectrogram extracted from a_4

We compute melspectrogram on windows of the signal and then we extracted mean, standard deviation, skewness and kurtosis.

5.3.12 Mel Frequency Cepstral Coefficients

The MFCC are coefficients based on the extraction of the signal energy within critical frequency bands by means of a series of triangular filters (in Figure 5.18) whose center frequencies are equally spaced according to the mel scale.

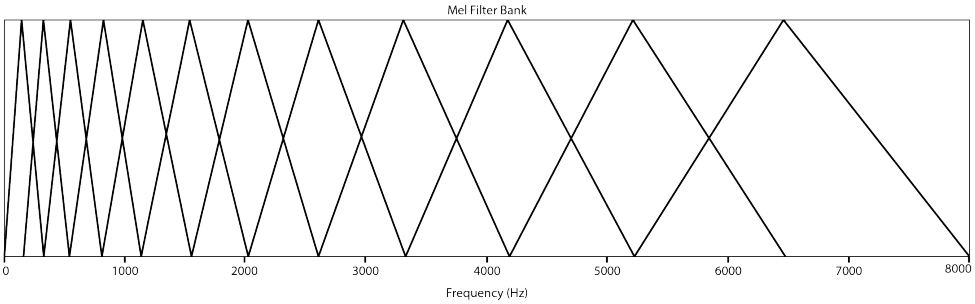


Figure 5.18: Triangular filters for the MFCC extraction

The log-energy of the spectrum is measured within the pass-band of each filter, resulting in a reduced representation of the spectrum. The cepstral coefficients are finally obtained through a Discrete Cosine Transform (DCT) of the reduced log-energy spectrum.
In Figure 5.19 is shown the MFCC graph for 12 mfccs for a_4.
For each of the 12 MFCCs are extracted the mean, standard deviation, median, kurtosis and skewness.

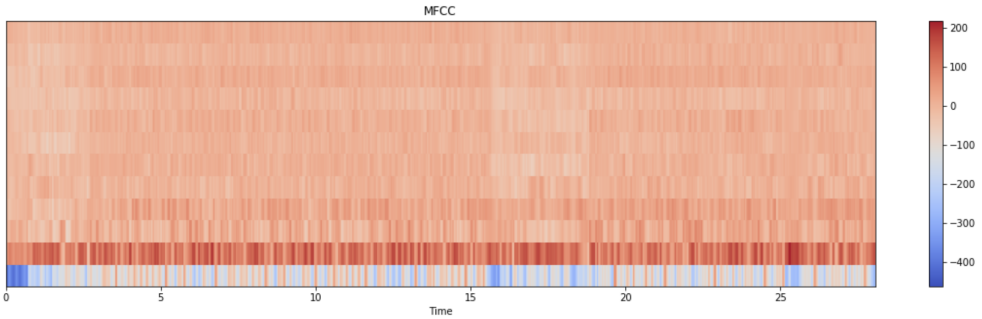


Figure 5.19: MFCC for a_4

5.4 EDA feature extraction

For the continuity, also in the explanation of EDA features extracted, will be used the e_4 Its EDA signal and its specs, can be seen in Figure 5.20.

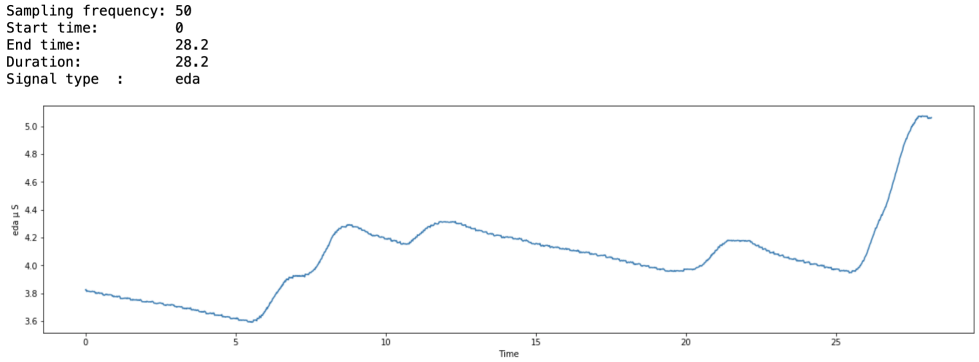


Figure 5.20: EDA for e_4

EDA signal was preprocessed following the pipeline highlighted in the pyphysio library [24] and resumed in Figure 5.21.

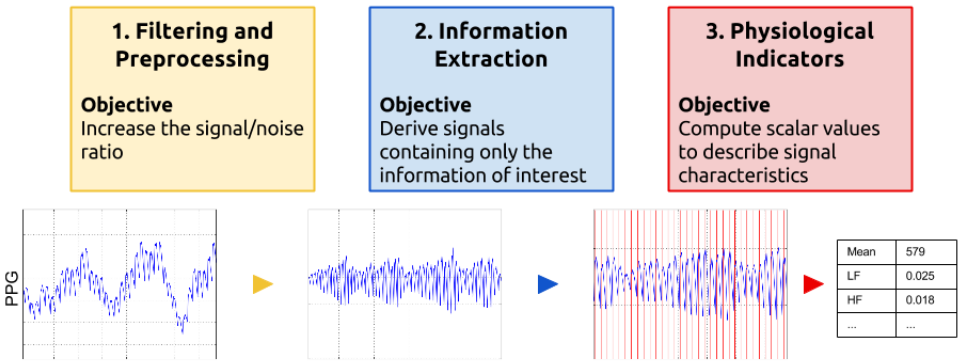


Figure 5.21: Pyphysio pipeline

They divide the pipelines into three separate steps:

1. Filtering and Preprocessing: this step includes all the procedures aiming at increasing the signal/noise ratio, typically band-pass filtering, smoothing, removal of artifacts. The output of this step is a new version of the input signal with improved signal quality (less noise).

2. Information Extraction: this step aims at extracting the information of interest from the physiological signal. The output is a new signal containing only the information of interest.

3. Physiological Indicators: this steps produces a list of scalar values able to describe the characteristics of the input signal. This step is usually performed on small segments of the input signals which are extracted using a sliding window on the whole length of the signal.

Taking into account the initial signal shown in Figure 5.20 the filtering and preprocessing stage was done by an IIR filter to remove high frequency noise, a low-pass filter of $0.6Hz$ to diminish the noise from motion and artifacts as can be seen in Figure 5.22.

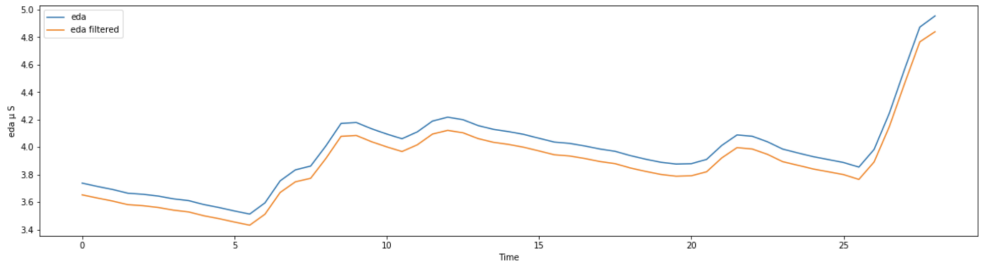


Figure 5.22: EDA signal in blue and filtered EDA in orange for e_4

Following the procedure introduced by [24], the library is able to extract tonic and phasic component of an EDA data by evaluating a driver function. The graph can be seen in Figure 5.23.

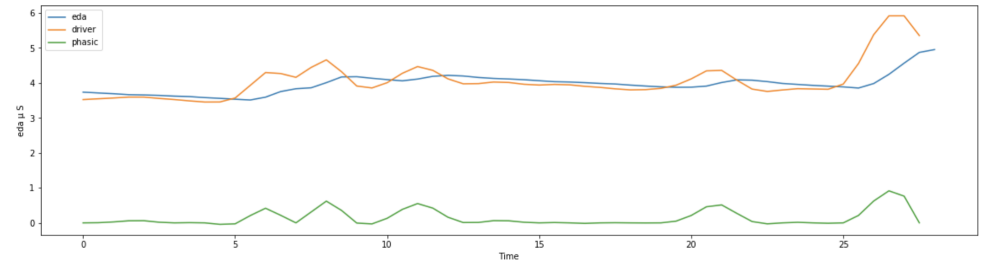


Figure 5.23: EDA signal in blue, driver in orange and tonic part in green for e_4

Now will follow a list of features extracted during the process.
Features were extracted both in a static way, taking into account the

whole excerpt and in a dynamic way, by dividing the musical excerpt in windows of $1s$ with 50% overlap.
After the process of feature extraction, every feature was normalized in the range $[0, 1]$.

5.4.1 Statistic

As already mentioned before in 3.4, most extracted EDA features are statistical. We extracted the mean of the signal, the standard deviation, kurtosis and the skewness. They were extracted both in the time domain and in the frequency domain.

Median value was also extracted, where the median is the value separating the higher half from the lower half of the data. It is like the *middle* value, but differently from the mean, the median gives a better idea of a typical value.
While other statistics were already described in 3.4, the median can be described in a caseless formula as:

$$median(a) = \frac{a\left\lceil \frac{l+1}{2} \right\rceil + a\left\lfloor \frac{l+1}{2} \right\rfloor}{2} \tag{5.6}$$

where a is an ordered list of l numbers, $\lceil \cdot \rceil$ is the ceil function (gives the least integer greater than or equal to the input) and $\lfloor \cdot \rfloor$ is the floor function (gives the greatest integer less than or equal to the input).

We extracted also the maximum value, the minimum and the difference between these two, the range.

5.4.2 Other features

Some other features were extracted thanks to pyphisio library [24], they are:

- Area Under the Curve (AUC) between two points can be found by doing a definite integral between the two points:

$$AUC = \int_{t_1}^{t_2} e_n de \tag{5.7}$$

- Root Mean Square Squared Difference (RMSSD) compute the Root Mean Square Error (RMSE) of the squared 1^{st} order discrete differences

- Standard Deviation Discrete Differences (SDSD) calculate the standard deviation of the 1^{st} order discrete differences

5.4.3 Power and Peak inband

The power spectrum $S_{xx}(f)$ of a time series a_n describes the distribution of power into frequency components composing the signal. Any discrete signal, according to Fourier analysis, can be decomposed into a number of discrete frequencies, or a spectrum of frequencies over a continuous range.
The statistical average of a certain signal as analyzed in terms of its frequency content, is called its spectrum.

The PSD, also called power spectrum, applies to signal over all time, that theoretically could be an infinite time interval. The PSD than, refers to the spectral energy distribution that would be found per unit time, since the total energy of such a signal over all time would generally be infinite.
We extracted the power and the peak frequency for each frequency band. We decided to set the same frequency ranges as in PMEmo:

- $0Hz - 0.1Hz$
- $0.1Hz - 0.2Hz$
- $0.2Hz - 0.3Hz$
- $0.3Hz - 0.4Hz$
- $0.4Hz - 0.5Hz$

5.4.4 Mel Frequency Cepstral Coefficients

As for the audio features, we extracted MFCC features. We kept 12 MFCC coefficients.
For each of the 12 MFCCs are extracted the mean, standard deviation, median, kurtosis and skewness.

5.5 Feature selection

Feature selection become an important step while performing a ML task. Given a dataset, every column of the dataset is a feature, and not necessarily every feature is going to have an impact on the output variable. Adding these irrelevant features in the model, it will make the model worst.
The feature selection methods that we are going to present are valid for a regression problem, where both the input and the output variables are continuous in nature.

Feature selection can be done in multiple ways but there are broadly 3 categories of it:

1. Filter Method by filtering and taking only the subset of the relevant features, the filtering is done using correlation matrices.
2. Wrapper Method which needs a ML algorithm and uses its performance as evaluation criteria.
3. Embedded Method,an iterative method. It takes care of each iteration of the model training process and extract those features which contribute the most to the training for a particular iteration.

Feature selection step is a very important passage, because the output of this step goes directly into the ML process. If is putted in the ML process noise data, will come out noise results and a model less accurate. This became more important where the number of features are large. It reduces the training time and the evaluation one.
Using less features, applying feature selection algorithms:

- It reduces overfitting.
- It enables the ML algorithm to train faster.
- It reduces the complexity of the model and make it easy to interpret.
- It improves the accuracy of the model if the right subset is chose.

The last point of the list is very important, because feature selection improve the model only if the right subset is chosen. In the implementation we evaluated the model score, and we saw that applying a feature selection algorithm, the score is improved.

Feature selection methods can be divided into:

1. Filter methods:
   Filter methods are independent from the ML algorithms. Here, features are selected on the basis of their scores in statistical tests for their correlation with the outcome variable. Filter methods are Pearson's correlation, LDA, ANalysis Of VAriance (ANOVA) and Chi-square.

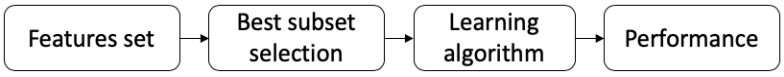


Figure 5.24: Filter methods scheme

2. Wrapper methods:

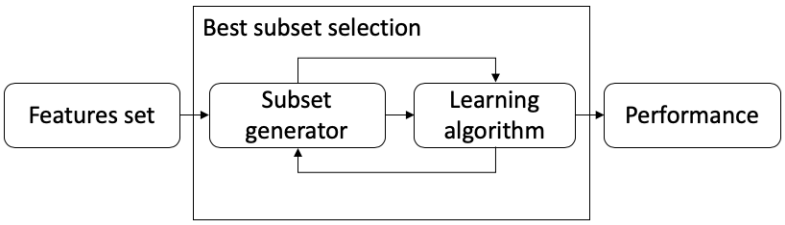


Figure 5.25: Wrapper methods scheme

The idea is to use a subset of features and train the model using them. Basing on the inferences that are drawn from the previous model, is decided to add or remove features from the subset. This problem can be seen as a search problem and of course, these models are computationally expensive due to their nature. Wrapper methods are forward selection, backward elimination and recursive feature elimination.

3. Embedded methods:

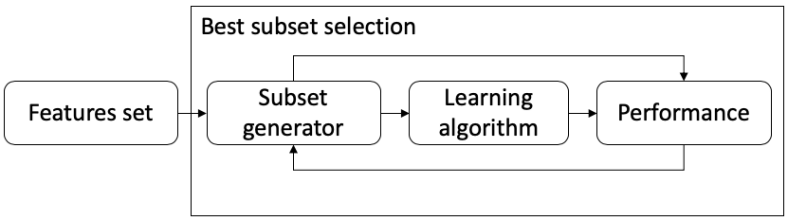


Figure 5.26: Embedded methods scheme

They combine the quality of the filter and wrapper methods. It is implemented by algorithms that have their own built-in feature selection methods.

In the following paragraphs are taken into account all the audio features extracted for the whole song, in the static case.

5.5.1 Pearson correlation

Pearson correlation is a filter method for feature selection. In this method a filter process is done which select only the subset of the relevant features. The model is built after selecting the features.
The filtering is done using Pearson correlation.

In statistics, the Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y.
The correlation coefficient has values between -1 to 1:

- A value closer to 0 implies weaker correlation (exact 0 implying no correlation)
- A value closer to 1 implies stronger positive correlation
- A value closer to -1 implies stronger negative correlation

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a *product moment*, that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{5.8}$$

where:

- $cov(X,Y)$ is the covariance (the measure of the joint variability of the two random variables X and Y)
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

In Figure 5.27 is shown the Pearson correlation heatmap of all the features, before applying any feature selection method, which represents the correlation of independent variables with the output variable.
Then are selected just the features that have a Pearson coefficient greater than a certain value, for example features that have a Pearson coefficient greater than 0.5 based on the output variable of *Valence (mean)* are shown in Figure 5.28.
Features that have a Pearson coefficient greater than 0.5 based on the output variable of *Valence (mean)* are:

Feature	P-coeff. > 0.5
Valence (mean)	1.0
medianMFCC[0]	0.606916
meanMFCC[0]	0.593028
stdMFCC[1]	0.541788
stdMFCC[2]	0.538732
spec_bw_std	0.514848
poly_mean	0.503954

Table 5.3: Selected features with Pearson correlation method

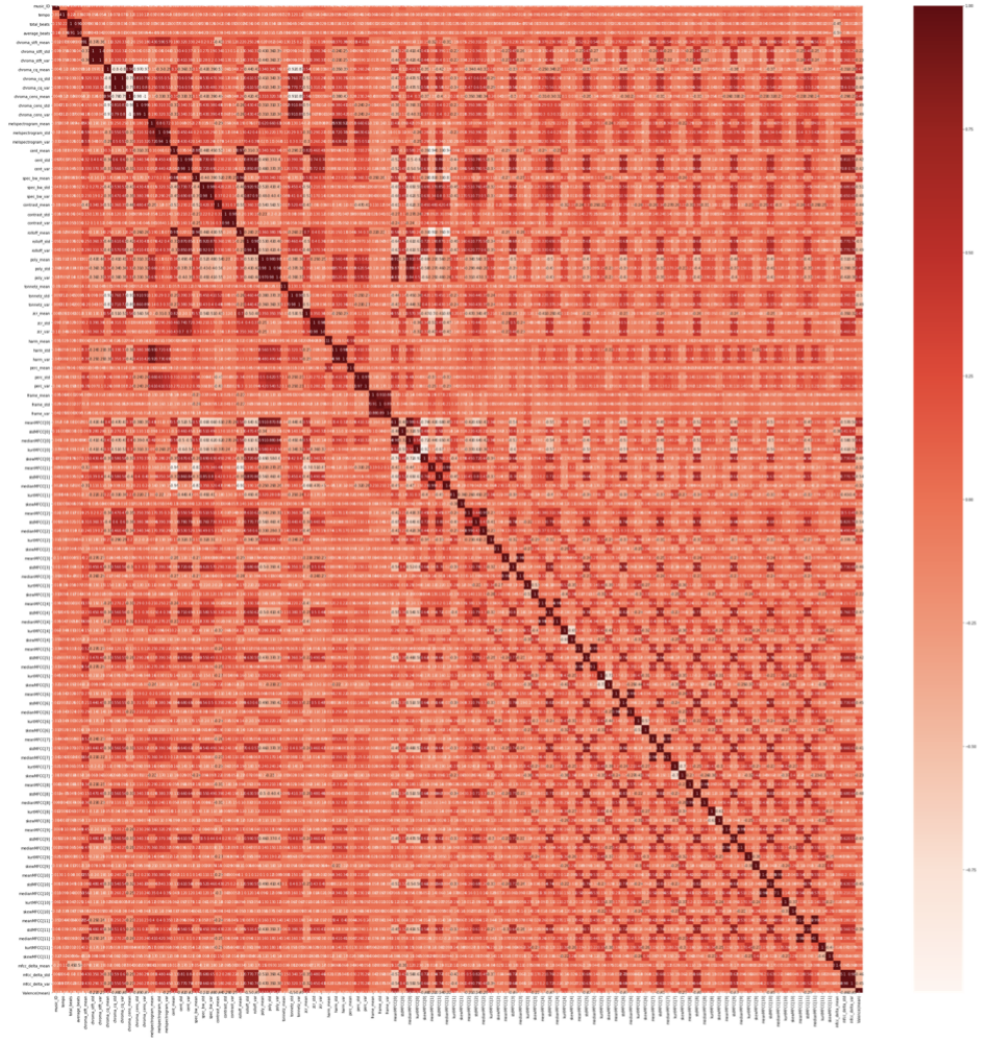


Figure 5.27: Pearson heatmap

5.5.2 Backward elimination

Backward elimination is a wrapper feature selection method. The model is fed at first with all the features, than the performance of the model is checked and iteratively is removed the worst performing features one by one till the overall performance of the model comes in acceptable range. The performance metric used here to evaluate feature performance is pvalue. If the pvalue is above 0.05 then we remove the feature, else it is kept.
Features that have a pvalue smaller than 0.05 are shown in Table 5.4.

5.5.3 Recursive feature elimination

The Recursive Feature Elimination (RFE) method is another wrapper method and works by recursively removing attributes and building a model on those attributes that remain.
It uses accuracy metric to rank the feature according to their importance.

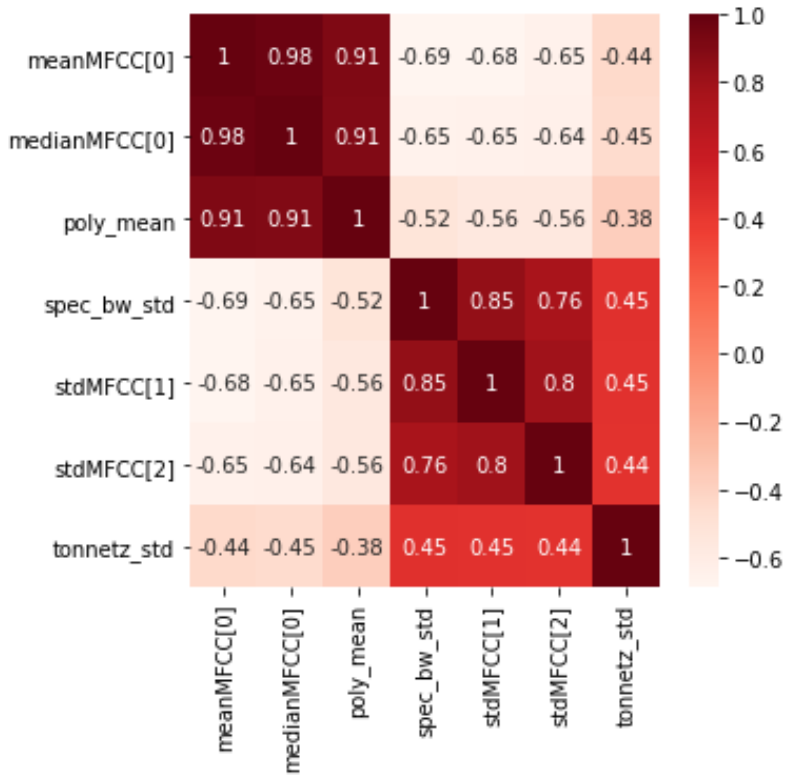


Figure 5.28: Pearson most relevant features

Feature	pvalue < 0.05	Feature	pvalue < 0.05
chroma_cq_mean	1.178195e-03	chroma_cq_std	1.880319e-02
chroma_cq_var	1.078829e-02	chroma_cens_mean	1.128383e-07
melspectrogram_mean	2.689379e-02	cent_mean	2.045630e-04
contrast_mean	3.953641e-03	contrast_std	3.293580e-02
tonnetz_std	1.684936e-02	tonnetz_var	2.160709e-02
harm_var	5.003698e-03	perc_std	2.241721e-03
frame_std	1.172305e-02	frame_var	1.691725e-02
stdMFCC[0]	6.128907e-05	skewMFCC[0]	3.544928e-02
medianMFCC[1]	7.279576e-07	skewMFCC[1]	3.116653e-05
meanMFCC[2]	1.603156e-02	medianMFCC[2]	4.386178e-03
stdMFCC[3]	1.041871e-02	medianMFCC[3]	2.085640e-02
skewMFCC[3]’	1.353325e-03	kurtMFCC[4]	1.293951e-03
skewMFCC[4]	3.583275e-03	stdMFCC[5]	2.307732e-02
medianMFCC[5]	5.428031e-03	skewMFCC[5]	2.791598e-02
meanMFCC[7]	1.521483e-02	stdMFCC[8]	3.149996e-06
medianMFCC[8]	1.441910e-04	meanMFCC[9]	4.847923e-04
meanMFCC[11]	1.331274e-03	mfcc_delta_std	8.903809e-03
mfcc_delta_var	1.039878e-02	const	1.503598e-8

Table 5.4: Selected features with Backward elimination method

The RFE method takes the model to be used and the number of required features as input. As output it gives the ranking of all the variables, where 1 is the most important.
Most relevant features extracted are shown in Table 5.5.

Feature	Ranking
chroma_stft_std	1
chroma_stft_var	1
chroma_cq_mean	1
chroma_cens_std	2
melspectrogram_mean	2
cent_mean	3
contrast_std	3
contrast_var	3
zcr_std	3
zcr_var	4
harm_std	5
harm_var	5
perc_std	5
frame_var	5
meanMFCC[0]	7
meanMFCC[1]	7
medianMFCC[2]	8
meanMFCC[2]	9
meanMFCC[3]	9
medianMFCC[6]	9
meanMFCC[2]	9
meanMFCC[6]	9
meanMFCC[7]	10

Table 5.5: Selected features with RFE method

5.5.4 Embedded method

Embedded methods are iterative in a sense that takes care of each iteration of the model training process and carefully extract those features which contribute the most to the training for a particular iteration. Regularization methods are the most commonly used embedded methods which penalize a feature given a coefficient threshold. Here we will do feature selection using Lasso regularization. If the feature is irrelevant, lasso penalizes its coefficient and make it 0. Hence the features with coefficient equal to 0 are removed and the rest are taken.
Most relevant features extracted are in Table 5.6.

Feature
rolloff_var
cent_var
total_beats
chroma_cens_std
melspectrogram_mean
melspectrogram_var
spec_bw_var

Table 5.6: Selected features with Embedded method

5.5.5 RReliefF

As already explained in 2.5.1, the RReliefF algorithm is based on the quality estimation of each features, between instances that are close each other.
The main benefit of Relief-based algorithms is that they identify feature interactions without having to exhaustively check every pairwise interaction, thus taking significantly less time than exhaustive pairwise search. In this case, features were automatically ordered given a score, based on the RReliefF algorithm and a random forest classifier with 100 estimators. In this implementation, as input of the function we need to specify the number of features we want to keep.
Features selected with this method are shown in Table 5.7.

Feature
stdMFCC[0]
rolloff_var
chroma_cens_mean
chroma_cens_std
meanMFCC[1]
frame_var
chroma_cq_var
poly_mean
poly_std
melspectrogram_var

Table 5.7: Selected features with RReliefF method

5.6 Machine Learning methods

The task of MER is a problem that can be classified in the field of supervised ML, because we have an input variable which is given by the sum of $\mathbf{F}^*_\mathbf{A}$ and $\mathbf{F}^*_\mathbf{E}$, called x.
x is a vector of features composed of audio and EDA features.
The output variable Y, the emotion, and we want to find an algorithm that lean the mapping function from the input to the output, $Y = f(x)$.
Following the general framework, the input of this process are the audio features and EDA features, reduced by a feature selection method, defined as $\mathbf{F}^*_\mathbf{A}$ and $\mathbf{F}^*_\mathbf{E}$ and the output of the process is the emotion value on the Valence and Arousal plane. This output is described by two scorer, the RMSE and the R^2. They gives the possibility to show how much the model is predicting well the emotion based on the initial data of VA values given by the subjects.

General goal is to approximate the mapping function so well that when there is a new input, the algorithm can predict the output data.
Supervised ML is divided in two categories, regression and classification. The main difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification.

A regression problem is when the output variable is a real or continuous value. Many different models can be used. They are explained in detail in the following paragraphs.

We will use two different regressors, one for the Valence and one for the Arousal, due to the fact that emotions are mapped on the 2-Dimension space and we want to find separate values for Valence and Arousal.

5.6.1 Linear Regression

LR is a type of regression analysis where there is a linear relationship between the independent x variable and the dependent one y.
The Figure 5.29 show a red line referred to the best fit straight line. Dots are the data points and the task of LR is to try to plot a line that models the points the best.
The line can be modeled based on the linear equation:

$$y = a_0 + a_1 \cdot x \tag{5.9}$$

Aim of LR is to find the best value for a_0 and a_1. This is a search problem, which can be converted into a minimization problem where we would like to minimize the error between the predicted value and the actual one.
The minimization problem is the cost function J:

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2 \tag{5.10}$$

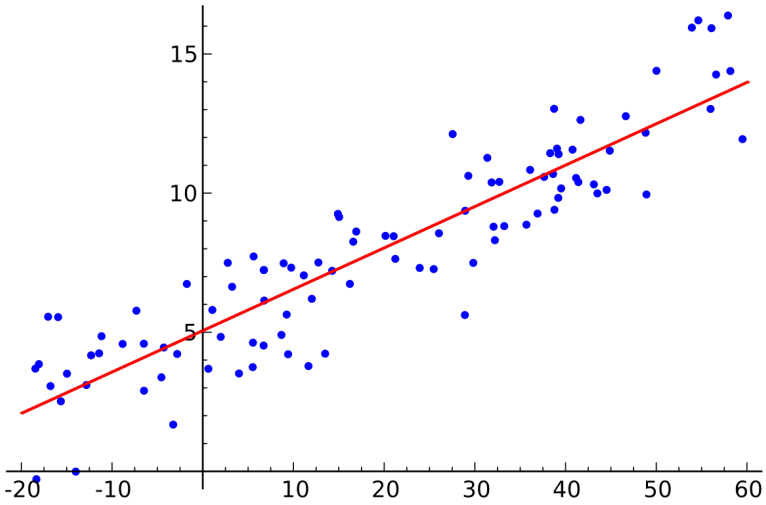


Figure 5.29: Linear regression

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)^2 \tag{5.11}$$

To update a_0 and a_1 values in order to reduce the cost function J is used the gradient descent, which idea is to start with random values of a_0 and a_1 and then iteratively update the values reaching minimum cost.

5.6.2 Lasso

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Taking into account a linear model based on n features represented as:

$$\hat{y} = w[0]x[0] + w[1]x[1] + ... + w[n]x[n] + b \tag{5.12}$$

Assuming the dataset has M instances and p values, the cost function of the regression problem can be written as:

$$J = \sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j x_{ij}\right)^2 \tag{5.13}$$

The LASSO cost function can be written as:

$$J = \sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j x_{ij}\right)^2 + \lambda \sum_{j=0}^{p}|w_j| \tag{5.14}$$

It is evident that for $\lambda = 0$, the equation 5.6.2 reduces to equation 5.6.2. LASSO can have zero coefficient, which lead to neglecting some features

for the evaluation of the output. This helps as a feature selection step. Feature selection using LASSO regression can be depicted well by changing the regularization parameter λ. This is called L_2 regularization.

5.6.3 Ridge

Ridge is another regression analysis method, similar to LASSO. Assuming the linear model described before in 5.6.2, the cost function for Ridge is:

$$J = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} (w_j)^2 \qquad (5.15)$$

Where is added a penalty equivalent to square of the magnitude of the coefficients w_j. This is called L_1 regularization.
The penalty term λ regularizes the coefficients such that if the coefficients take large value the optimization function is penalized. So, it shrinks the coefficients and helps to reduce model complexity.

5.6.4 Elastic Net

Elastic Net is a regularized regression that linearly combines the L_1 and L_2 penalties of the Ridge and LASSO methods. Absolute value penalization and squared penalization are combined with a coefficient, L_r

$$J = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + (1-r)\lambda \sum_{j=0}^{p} (w_j)^2 + r\lambda \sum_{j=0}^{p} (w_j)^2 \qquad (5.16)$$

5.6.5 k-nearest neighbors

The k-NN algorithm, is a non-parametric algorithm used for regression, where input consists of the k closest training examples in the feature space and the output is the property value for the object. This value is the average of the values of k nearest neighbors.
k-NN is sensitive to the local structure of the data.

5.6.6 Support Vector

Goal of a regression problem is to minimize the error rate. In SVR tries to fit the error within a certain threshold.
The error term is handled in constraints by setting the absolute error less than or equal to a specified margin, called maximum error ε. The problem is a minimization problem $minimize \frac{1}{2}||w||^2$ with the constraint $|y_i - w_i x_i| \leq \varepsilon$ showed in Figure 5.30.

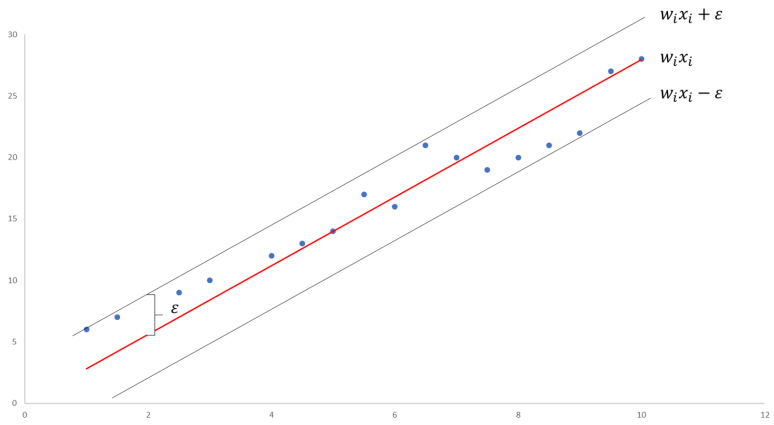


Figure 5.30: Support Vector regression

5.6.7 Decision Tree

DT is a supervised ML problem used to predict a target by learning decision rules from features. Its model is based on a data division by making a decision based on asking a series of questions.

DT is constructed by recursive partitioning, starting from the root node, and then each node can be split into left and right child nodes. These nodes can then be further split and they themselves become parent nodes of their resulting children nodes.

5.6.8 Random Forest

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.
An ensemble method combines the predictions from multiple ML algorithms together to make more accurate predictions.
Figure 5.31 show an ideal representation of RF.
RF operates by constructing a multitude of decision trees at training time and outputting the class that is the mean prediction of the individual trees.
RF allows to aggregate many DT and this gives the possibility to split the features but limited to some percentage of the total (the hyperparameter). This allows to have a balanced weight on all the features.

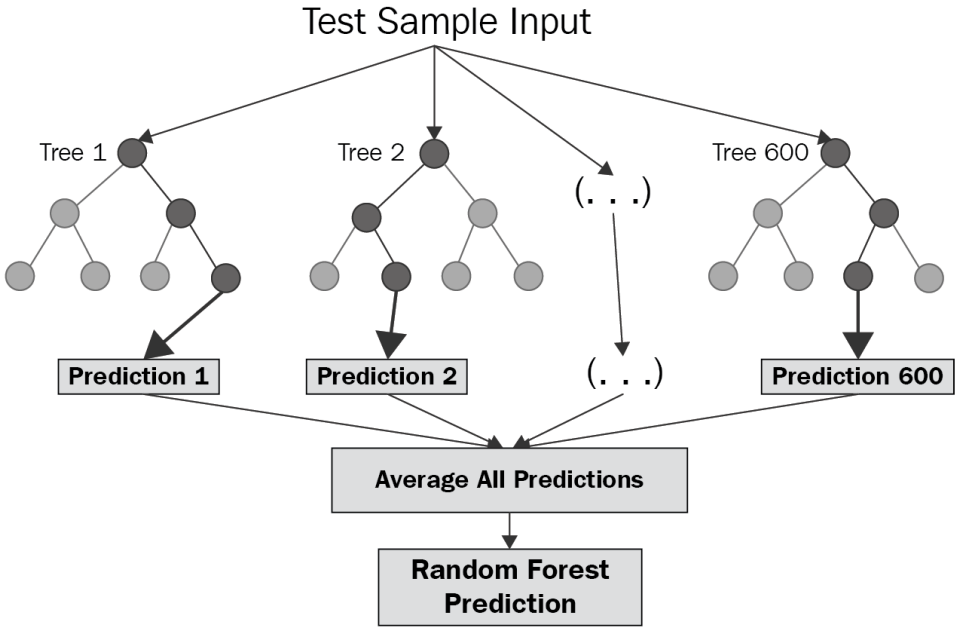


Figure 5.31: Random Forest regression

6 Results

In this chapter we present as first PMEmo results and performances, then we will show how our system perform based on different feature selection methods and different ML regression methods.

To evaluate how much the ML model is precise and correct we used two different parameters, the RMSE and the coefficient of determination R^2.

For both cases, to evaluate the score we applied a cross-validation with 10 fold to train the data and one remaining to test the data.
A division example is shown in figure 6.1.

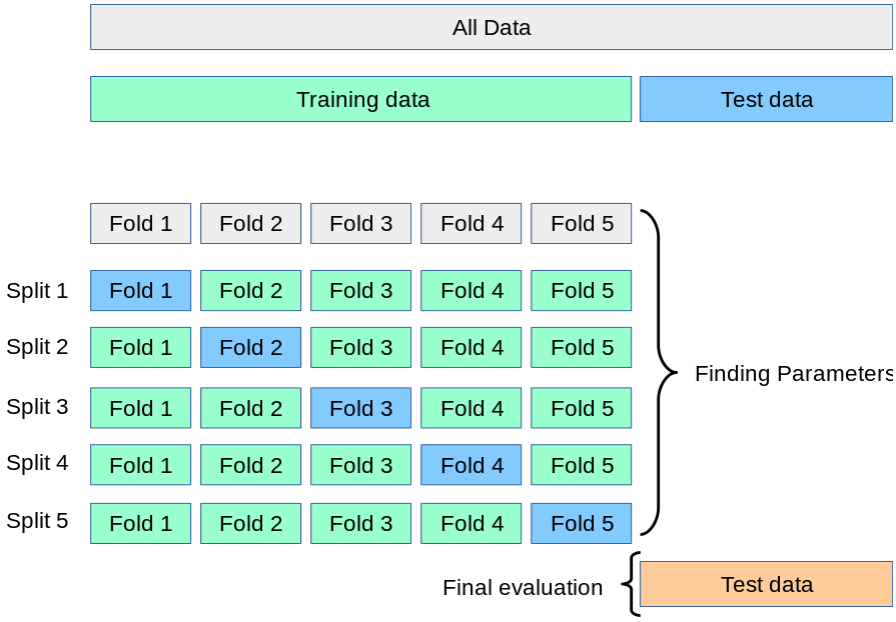


Figure 6.1: k-fold cross validation

A well-fitting regression model results in predicted values close to the observed data values.
The RMSE is the square root of the variance of the residuals:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}} \tag{6.1}$$

It indicates the absolute fit of the model to the data, how much close the observed data points are to the model's predicted values. RMSE is an absolute measure of fit. It is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by variables in a regression model.
R^2 is evaluated as:

$$R^2 = \frac{ESS}{TSS} \tag{6.2}$$

Where ESS is the sum of squares terms and TSS is the total sum of squares, defined as:

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{6.3}$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6.4}$$

In this case $\bar{y}$ is the mean of the data observed.
It scales from 0 to 1 where 0 indicates that the proposed model does not improve prediction over the mean model and 1 indicates the perfect prediction. It can be also negative, in the case where the model can be arbitrarily worse.

6.1 PMEmo performances

In the article, they adopt Multivariate Linear Regression (MLR) and SVR as the base classifiers to model emotions in valence and arousal.
For the static part, they trained and tested the classifiers using all the 6373-dimension features $x_1, x_2, ..., x_{6373}$ and separate static labels of valence $y_{valence}$ and arousal $y_{arousal}$ respectively.

$$X_1, X_2, ..., X_m \quad \rightarrow \quad e_1, e_2, ..., e_m \tag{6.5}$$

where:

- m is the number of songs
- $X_i = x_1, x_2, ..., x_{6373}$ is the feature set of the i^{th} song
- e_i is the value of valence or arousal for this song

With respect of continuous mood of a song, is natural to consider a decoupling into two scales and then recognize them separately.
For the dynamic emotion, defined as:

$$L_i = \bar{L}_i + D_i^{t_i} \tag{6.6}$$

where:

- t_i is the number of timestamps in the i^{th} song
- $\bar{L}_i$ is the mean of dynamic emotion
- $D_i^{t_i}$ is the fluctuation at each timestamp

the global model is:

$$X_1, X_2, ..., X_m \quad \rightarrow \quad \bar{L}_1, \bar{L}_2, ..., \bar{L}_m \tag{6.7}$$

while, the local model is:

$$Y_1^{t_1}, Y_2^{t_2}, ..., Y_m^{t_m} \quad \rightarrow \quad D_1^{t_1}, D_2^{t_2}, ..., D_m^{t_m} \tag{6.8}$$

where:

- m is the number of songs
- X_i is the global feature set of it
- $Y_i^{t_i}$ is a matrix of 260 columns and t_i rows

Dimension	Classifier	RMSE	r
Valence	MLR	0.136	0.546
Valence	SVR	0.124	0.638
Arousal	MLR	0.111	0.719
Arousal	SVR	0.102	0.764

Table 6.1: Evaluation results on static emotions

Before the regression models, they resized all the annotations (both for static and dynamic annotations) into $[0, 1]$.

Static task is to predict the overall emotion of a whole song, represented by a single valence value and arousal value. To train and test, they divided the dataset in 11 folds, 10 constituted the training set and the remaining set used to test the train model. A 10-fold-cross-validation was used for parameter optimization.
RMSE and Pearson Correlation Coefficient (r) were calculated separately for valence and arousal. In Table 6.1 is shown the results on static emotions.
About the dynamic case, a hierarchical regression model aiming to recognize the global trend as well as local variation was built. For Global-scale they extracted, for each song, one global feature and mapped it into one global emotion. For Local-scale operation, for each song, they divided it into $1s$ segment with 50% overlap, then extracting the local features from these fragments and project them onto mood space.

In Table 6.2 is presented the evaluation results on dynamic emotions.

Dimension	Classifier	Scale	RMSE	r
Valence	MLR	global	0.103	0.673
Valence	MLR	local	0.016	0.047
Valence	SVR	global	0.106	0.675
Valence	SVR	local	0.016	0.095
Arousal	MLR	global	0.113	0.816
Arousal	MLR	local	0.020	0.103
Arousal	SVR	global	0.101	0.844
Arousal	SVR	local	0.019	0.115

Table 6.2: Evaluation results on dynamic emotions

In PMEmo work, as already mentioned, they also recorded EDA subjects data when they were listening to music.
On EDA, they employed a low-pass filter of $0.6Hz$ to diminish the noise due to motion artifacts.

Then skin electric conductance was scaled in z-score:

$$z - score = \frac{X - \mu}{\sigma} \tag{6.9}$$

where μ is the mean of vector X and σ is the standard variation. Last passage on EDA signal was to resample them, from $50Hz$ to $2Hz$ due to different acquisition of EDA and continuous emotions.
They trained and tested MLR and SVR with pre-processed EDA data in the dynamic case and results are shown in Table 6.3.

Dimension	Classifier	Scale	RMSE	r
Valence	MLR	global	0.139	0.063
Valence	MLR	local	0.016	0.060
Valence	SVR	global	0.141	0.017
Valence	SVR	local	0.016	0.059
Arousal	MLR	global	0.186	0.011
Arousal	MLR	local	0.019	0.097
Arousal	SVR	global	0.194	0.040
Arousal	SVR	local	0.019	0.099

Table 6.3: Evaluation results on dynamic EDA

6.2 Our model performances

In this section we present all the results acquired with different parameters, with and without feature selection, different data types and different regressors.

6.2.1 No feature selection

Here, in Table 6.4 are shown the results for audio data, where no feature selection algorithm has been applied. In the Tables, are highlighted the best results for every dimension, one for the RMSE and one for r2.

Arousal (mean) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.102	0.184	0.184	0.100	0.119	0.113	0.209	0.106	0.129	0.136
R2	0.669	-0.039	-0.039	0.680	0.558	0.606	-1.115	0.644	0.477	0.459
Valence (mean) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.122	0.162	0.162	0.120	0.126	0.119	0.211	0.127	0.143	0.127
R2	0.373	-0.056	-0.056	0.400	0.357	0.418	-2.233	0.333	0.148	0.356
Arousal (std) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.047	0.047	0.047	0.046	0.047	0.050	0.050	0.049	0.051	0.045
R2	0.007	-0.013	-0.013	0.051	-0.009	-0.143	-0.123	-0.097	-0.184	0.136
Valence (std) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.051	0.046	0.046	0.051	0.049	0.047	0.048	0.048	0.052	0.045
R2	-0.349	-0.026	-0.026	-0.334	-0.187	-0.071	-0.136	-0.104	-0.397	-0.011

Table 6.4: No feature selection for audio data, with RMSE and r2 score

In Table 6.5 are shown the results for EDA data, where no feature selection algorithm has been applied.

Arousal (mean) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.188	0.184	0.184	0.182	0.136	0.074	0.096	0.055	0.018	0.182
R2	0.560	-0.039	-0.039	0.515	0.435	0.831	0.685	0.807	0.800	0.019
Valence (mean) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.163	0.162	0.162	0.170	0.121	0.073	0.100	0.053	0.018	0.158
R2	0.500	-0.056	-0.056	0.416	0.410	0.780	0.480	0.886	0.855	-0.006
Arousal (std) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.047	0.047	0.047	0.047	0.035	0.048	0.048	0.045	0.044	0.046
R2	-0.013	-0.013	-0.013	0.011	0.439	-0.072	-0.042	0.067	0.070	0.035
Valence (std) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.046	0.046	0.046	0.046	0.035	0.045	0.045	0.43	0.030	0.044
R2	-0.025	-0.026	-0.026	0.205	0.406	0.015	0.012	0.101	0.101	0.026

Table 6.5: No feature selection for EDA data, with RMSE and r2 score

In Table 6.6 are shown the results for fusion data, given by the union of audio and EDA features, where no feature selection algorithm has been applied.

Arousal (mean) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.109	0.184	0.184	0.103	0.126	0.110	0.139	0.112	0.131	0.147
R2	0.622	-0.039	-0.039	0.621	0.510	0.769	0.346	0.599	0.465	0.332
Valence (mean) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.134	0.162	0.162	0.127	0.329	0.645	0.448	0.138	0.142	0.137
R2	0.243	-0.056	-0.056	0.322	0.333	0.389	-0.014	0.204	0.162	0.244
Arousal (std) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.052	0.047	0.047	0.049	0.046	0.050	0.050	0.049	0.051	0.045
R2	-0.222	-0.013	-0.013	-0.101	0.027	-0.160	-0.159	-0.090	-0.223	0.081
Valence (std) dimension										
	LR	Lasso	ElasticNet	Ridge	kNN	SVRrbf	SVRpoly	SVRlinear	DT	RF
RMSE	0.056	0.046	0.046	0.053	0.048	0.047	0.048	0.052	0.052	0.046
R2	-0.547	-0.026	-0.026	-0.410	-0.140	-0.087	-0.115	-0.116	-0.356	-0.009

Table 6.6: No feature selection for fusion data, with RMSE and r2 score

In the previous Table, 6.4, 6.5 and 6.6 are shown results, where any feature selection method was applied. We expect that using all the set of features for audio and EDA that are extracted, it will improve the PMEmo baseline results. This assumption is based on the idea that PMEmo extracted features from a software that deal more with speech features, which may not be so much useful.

We extracted all these results also for every different feature selection methods and again, we expect to improve the model, thanks to the idea of feature selection, which should reduce the overfitting in the model and reduce the computation complexity.

6.2.2 Feature selection

After getting results from the three set of data, audio, EDA and fusion with all the features, we implemented all the feature selection algorithms already explained in Chapter 5.5 and get results for every regression method.

We compared all the different possibilities following the scheme in Figure 6.2.
To clarify, all regression methods are applied for every feature selection method. The same is valid for the last part, where every regression method is analyzed for every VA space.

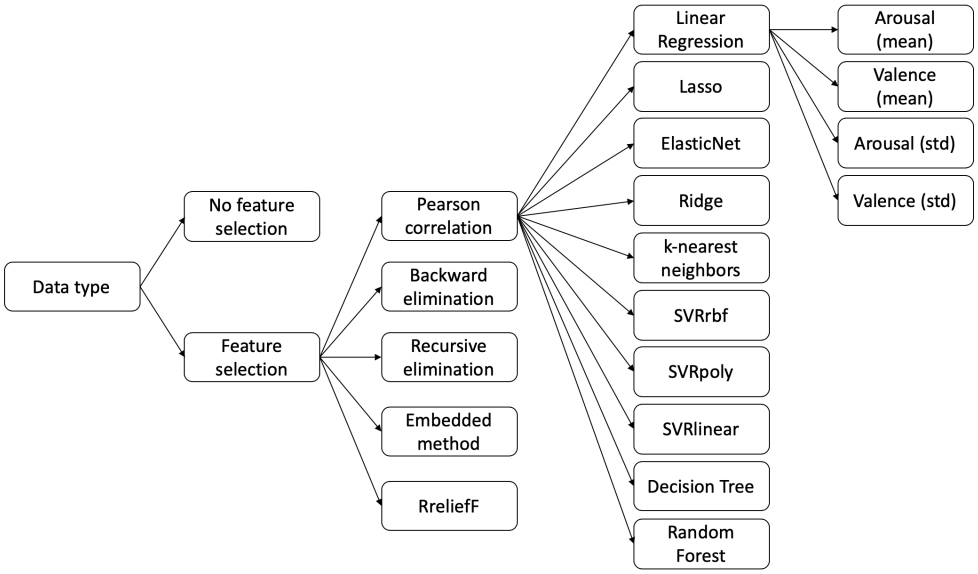


Figure 6.2: Evaluation possibilities

6.2.3 Best performances

We observed that the best couple feature selection method and regression approach, for both the VA in mean and standard deviation is using the **backward elimination** method and the **Ridge** regressor.

An important analysis is that we gain better results not for just audio data type or EDA type, but for the fusion data type, so it become relevant to use both audio and EDA combined.

For the different evaluation spaces, either Valence or Arousal (mean and standard deviation) the Backward elimination algorithm extracted from 30 to 50 features.

Understood that the Ridge regression on fusion data with Backward elimination algorithm gives the best results, we decided to calibrate the Ridge regression method.

A parameter of the Ridge regressor is α, called *complexity parameter* which controls the amount of shrinkage, the larger value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

The Ridge coefficients minimize a penalized residual sum of squares:

$$||x - tw||_2^2 + \alpha||w||_2^2 \tag{6.10}$$

where y is the training data and t the target value (in these cases VA data) and w the weights.

In the sklearn documentation is reported a graph that related the complexity parameter to the weights, shown in Figure 6.3.

We have found better results for small α values, around $\alpha = 0.001$. In the following Table, 6.7 are compared results from PMEmo, results with our model with no feature selection and results in the best case, mentioned before.

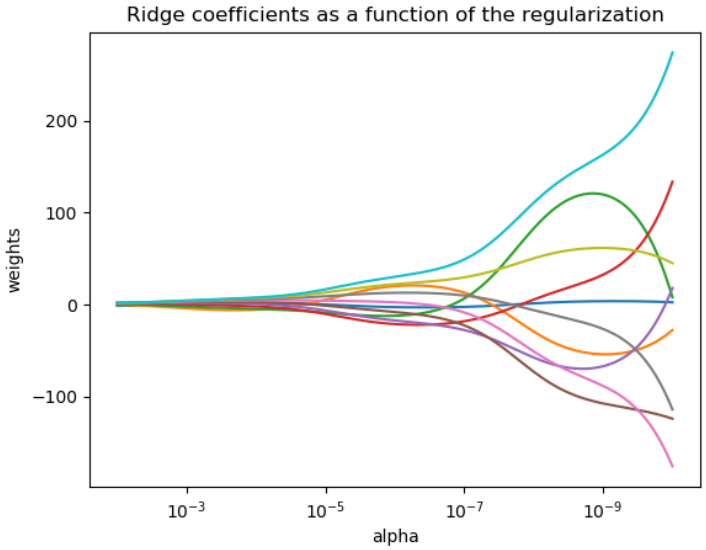


Figure 6.3: Relationship between complexity parameter α and weights of ridge regressor

Since PMEmo evaluated their data only on VA space in mean values, to have a fair comparison here are shown results for mean values of VA.

Dimension	Scorer	PMEmo	No F.S.	F.S. best
Arousal	RMSE	0.107	0.103	0.0417
Valence	RMSE	0.121	0.115	0.0435
Arousal	R2	0.764	0.769	0.780
Valence	R2	0.638	0.645	0.834

Table 6.7: Comparisons between PMEmo results, our algorithm with no feature extraction and with feature extraction and best setup of the regressor

7 Conclusions and Future Works

In this chapter we will review the work presented in this thesis and we will introduce some possible evolution of our system.

7.1 Conclusions

This thesis presented a complete work on a complex task of MER, find a relationship between music and emotions perceived by human during the listening.
We tried to solve this task by combining both audio and EDAdata, extracting several features which theoretically are relevant for the two data types.

As one could have already understand, the first problem was understanding which features are relevant for audio and EDA. Even if audio, in MIR is a well studied task, there is no evidence on which features are relevant respect to other, so moves are made by empirical attempts.
Much more complicated is the feature extraction procedure for EDA data, since there are several problem on figuring out how to threat the data, how to preprocess them and which features are relevant.

Once features were find out, the following problem was to find a good algorithm of feature selection, because how teach the ML theory, not all features have the same value and not all are relevant in the same way.
Most of the feature extraction methods are based on statistical processes, and they are useful to discard redundant features, which may lead to an overfitting model. So, also the feature extraction part is done in a certain empirical sense, by trying different possibilities.

As last, also the ML method is a complex process, it has many different implementation and is not clear which is best for the task of MER.

To summarize, the whole work is not a standard one, there are not standard rules to be followed and many decision are taken by trying different possibilities and choosing the best solution.

These problems sum to the fact that it is really hard to transform human being emotions into a numerical vision. What is certain is that music conveys emotion and modulate a listener's mood. Also while listening the same musical piece, one can feel different emotions, due to the fact that emotions are very complex.
After all these problems we started from the baseline of the work done by [8] and tried to improve their results. As one can see in the chapter 6 we have increase the model, by resulting in a smaller RMSE and bigger R^2 scores.

By looking the Table 6.7, our results are very positive, because, as firstly we improved the results, but also because they are aligned with our theoretical expectations.
We expect that by extracting features that are more relevant and suitable for audio we would improve the model, and that is what results tell us about. Scorer of our model with no feature selection are better than PMEmo results. This mean that just using audio features, the performance increases.

As another important step is to analyze which is the best combination of data type, feature selection method and regressor.
As data type, the best result is given with the fusion one, it means that combining audio and EDA features is the right path to follow in order to have a better model. Perceived and felt emotions are linked together, and having the possibility to combine them is a grateful opportunity.
This is the crucial point, EDA data are now fundamental to have a better prediction of the Valence and Arousal values, which means a better model and a better detector of music emotions.

7.2 Future Works

As the algorithm of recognizing emotion through audio and EDA data is a novel task, the algorithm proposed in this thesis opens further improvements.

As first, better features for audio can be extracted and can be studied in a deeper way EDA data, which is not very clear how to process them. It can be applied different feature selection algorithms and different ML methods though the regression problem.

As further improvement, we can think to apply also ML implementations that are not traditional, as deep ML implementations.
A nice test might be to threat the data as images, and apply some ML methods directly to the images, as the well known CNN.
There are several studies on the use of CNN on the spectrogram of a music piece and they are very interesting. In this case, the problem remain on how to convert EDA data in images, as the spectrogram for the audio.

An interesting try is done in the work of *Chaspari* in [59] which creates a similar spectrogram, called *EDA-Gram*, but for EDA.
A possible approach would be to combine the spectrogram for audio and *EDA-Gram* for EDA and use them as input of a deep neural ML process in order to avoid all the problems due to feature extraction and selection.

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] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[3] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1934–1937, IEEE, 2015.

[4] A. Shah, M. Kattel, A. Nepal, and D. Shrestha, "Chroma feature extraction," 01 2019.

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

[6] J. Shukla, M. Barreda-Angeles, J. Oliver, G. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, 2019.

[7] X. Hu, F. Li, and T.-D. J. Ng, "On the relationships between music-induced emotion and physiological signals," in *ISMIR*, pp. 362–369, 2018.

[8] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 acm on international conference on multimedia retrieval*, pp. 135–142, 2018.

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

[10] 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.

[11] 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.

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

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

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

[15] 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.

[16] 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.

[17] 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.

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

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

[20] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[21] W. Trost, T. Ethofer, M. Zentner, and P. Vuilleumier, "Mapping aesthetic musical emotions in the brain," *Cerebral Cortex*, vol. 22, no. 12, pp. 2769–2783, 2012.

[22] L. C. Johnson and A. Lubin, "Spontaneous electrodermal activity during waking and sleeping," *Psychophysiology*, vol. 3, no. 1, pp. 8–17, 1966.

[23] D. C. Fowles, "Electrodermal activity and antisocial behavior: Empirical findings and theoretical issues," in *Progress in electrodermal research*, pp. 223–237, Springer, 1993.

[24] A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito, and C. Furlanello, "pyphysio: A physiological signal processing library for data science approaches in physiology," *SoftwareX*, vol. 10, p. 100287, 2019.

[25] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2016.

[26] E. B. Hedman, *In-situ measurement of electrodermal activity during occupational therapy*. PhD thesis, Massachusetts Institute of Technology, 2010.

[27] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, pp. 49–54, 2010.

[28] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[29] P. Ghaderyan and A. Abbasi, "An efficient automatic workload estimation method based on electrodermal activity using pattern classifier combinations," *International Journal of Psychophysiology*, vol. 110, pp. 91–101, 2016.

[30] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[31] S. De Nadai, M. D'Incà, F. Parodi, M. Benza, A. Trotta, E. Zero, L. Zero, and R. Sacile, "Enhancing safety of transport by road by online monitoring of driver emotions," in *2016 11th System of Systems Engineering Conference (SoSE)*, pp. 1–4, Ieee, 2016.

[32] R. Guo, S. Li, L. He, W. Gao, H. Qi, and G. Owens, "Pervasive and unobtrusive emotion sensing for human mental health," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 436–439, IEEE, 2013.

[33] B. Verschuere, G. Crombez, E. Koster, and K. Uzieblo, "Psychopathy and physiological detection of concealed information: a review," *Psychologica Belgica*, vol. 46, no. 1-2, 2006.

[34] Y.-D. Zhang, Z.-J. Yang, H.-M. Lu, X.-X. Zhou, P. Phillips, Q.-M. Liu, and S.-H. Wang, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.

[35] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[36] T. Dissanayake, Y. Rajapaksha, R. Ragel, and I. Nawinne, "An ensemble learning approach for electrocardiogram sensor based human emotion recognition," *Sensors*, vol. 19, no. 20, p. 4495, 2019.

[37] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ecg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, 2017.

[38] M. Naji, M. Firoozabadi, and P. Azadfallah, "Classification of music-induced emotions based on information fusion of forehead biosignals and electrocardiogram," *Cognitive Computation*, vol. 6, no. 2, pp. 241–252, 2014.

[39] M. Naji, M. Firoozabadi, and P. Azadfallah, "Emotion classification during music listening from forehead biosignals," *Signal, Image and Video Processing*, vol. 9, no. 6, pp. 1365–1375, 2015.

[40] J. Cai, G. Liu, and M. Hao, "The research on emotion recognition from ecg signal," in *2009 International Conference on Information Technology and Computer Science*, vol. 1, pp. 497–500, IEEE, 2009.

[41] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[42] A. Goshvarpour, A. Abbasi, and A. Goshvarpour, "An accurate emotion recognition system using ecg and gsr signals and matching pursuit method," *biomedical journal*, vol. 40, no. 6, pp. 355–368, 2017.

[43] S. K. Yoo, C. K. Lee, Y. J. Park, N. H. Kim, B. C. Lee, and K. S. Jeong, "Neural network based emotion estimation using heart rate variability and skin resistance," in *International conference on natural computation*, pp. 818–824, Springer, 2005.

[44] O. Sourina, Y. Liu, and M. K. Nguyen, "Real-time eeg-based emotion recognition for music therapy," *Journal on Multimodal User Interfaces*, vol. 5, no. 1-2, pp. 27–35, 2012.

[45] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart

rate variability," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 385–394, 2015.

[46] D. M. Sloan, "Emotion regulation in action: Emotional reactivity in experiential avoidance," *Behaviour Research and Therapy*, vol. 42, no. 11, pp. 1257–1270, 2004.

[47] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pp. 410–415, IEEE, 2011.

[48] C. Mühl, A. Brouwer, N. van Wouwe, E. van den Broek, F. Nijboer, and D. K. Heylen, *Modality-specific affective responses and their implications for affective BCI*. Graz, Austria: Verlag der Technischen Universität, 2011.

[49] F. Al Machot, A. Elmachot, M. Ali, E. Al Machot, and K. Kyamakya, "A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors," *Sensors*, vol. 19, no. 7, p. 1659, 2019.

[50] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[51] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1–30, 2012.

[52] X. Hu and Y.-H. Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 228–240, 2017.

[53] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 693–697, IEEE, 2015.

[54] E. Schubert, "Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music," *Psychology of music*, vol. 41, no. 3, pp. 350–371, 2013.

[55] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[56] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835–838, 2013.

[57] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 113–116, IEEE, 2002.

[58] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pp. 21–26, 2006.

[59] T. Chaspari, A. Tsiartas, L. I. S. Duker, S. A. Cermak, and S. S. Narayanan, "Eda-gram: Designing electrodermal activity fingerprints for visualization and feature extraction," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 403–406, IEEE, 2016.