UNIVERSIDADE D
COIMBRA

Renato Eduardo Silva Panda

Emotion-based Analysis and Classification of Audio Music

Doctoral thesis submitted to the Doctoral Program in Information Science and Technology, supervised by Prof. Dr. Rui Pedro Pinto de Carvalho e Paiva, and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

January 2019

Faculdade de Ciências e Tecnologia

EMOTION-BASED ANALYSIS AND CLASSIFICATION OF AUDIO MUSIC

Renato Eduardo Silva Panda

Doctoral thesis submitted to the Doctoral Program in Information Science and Technology, supervised by Prof. Dr. Rui Pedro Pinto de Carvalho e Paiva, and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

January 2019

UNIVERSIDADE D
COIMBRA

AUTHOR, SUPERVISOR AND INSTITUTION INFORMATION

Author:	Renato Eduardo Silva Panda panda@dei.uc.pt
Supervisor:	Prof. Dr. Rui Pedro Pinto de Carvalho e Paiva ruipedro@dei.uc.pt
Host Institution:	Centre for Informatics and Systems of the University of Coimbra
Institution granting the academic degree:	University of Coimbra, Faculty of Sciences and Technology

Financial support by Fundação para a Ciência e a Tecnologia,

SFRH/BD/91523/2012.

Emotion-based Analysis and Classification of Audio Music

Thesis submitted to the
University of Coimbra
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Informatics Engineering

This work was carried out under the supervision of

Professor Doutor Rui Pedro Pinto de Carvalho e Paiva

Professor Auxiliar do
Departamento de Engenharia Informática da
Faculdade de Ciências e Tecnologia da
Universidade de Coimbra

ABSTRACT

This research work addresses the problem of music emotion recognition using audio signals. Music emotion recognition research has been gaining ground over the last two decades. In it, the typical approach starts with a dataset, composed of music files and associated emotion ratings given by listeners. This data, typically audio signals, is first processed by computational algorithms in order to extract and summarize their characteristics, known as features (e.g., beats per minute, spectral metrics). Next, the feature set is fed to machine learning algorithms looking for patterns that connect them to the given emotional annotations. As a result, a computational model is created, which is able to infer the emotion of a new and unlabelled music file based on the previously found patterns.

Although several studies have been published, two main issues remain open and are the current barrier to progress in field. First, a high-quality public and sizeable audio dataset is needed, which can be widely adopted as a standard and used by different works. Currently, the public available ones suffer from known issues such as low quality annotations or limited size. Also, we believe novel emotionally-relevant audio features are needed to overcome the *plateau* of the last years. Supporting this idea is the fact that the vast majority of previous works were focused on the computational classification component, typically using a similar set of audio features originally proposed to tackle other audio analysis problems (e.g., speech recognition). Our work focuses on these two problems.

Proposing novel emotionally-relevant audio features requires knowledge from several fields. Thus, our work started with a review of music and emotion literature to understand how emotions can be described and classified, how music and music dimensions work and, as a final point, to merge both fields by reviewing the identified relations between musical dimensions and emotional responses. Next, we reviewed the existent audio features, relating them with one of the eight musical dimensions: melody, harmony, rhythm, dynamics, tone color, expressive techniques, musical texture and musical form. As a result, we observed that audio features are unbalanced across musical dimensions, with expressive techniques, musical texture and form said to be emotionally-relevant but lacking audio extractors.

To address the abovementioned issues, we propose several audio features. These were built on previous work to estimate the main melody notes from the low-level audio signals. Next, various musically-related metrics were extracted, e.g., *glissando* presence, articulation information, changes in dynamics and others. To assess their relevance to

emotion recognition, a dataset containing 900 audio clips, annotated in four classes (Russell's quadrants) was built.

Our experimental results show that the proposed features are emotionally-relevant and their inclusion in emotion recognition models leads to better results. Moreover, we also measured the influence of both existing and novel features, leading to a better understanding of how different musical dimensions influence specific emotion quadrants. Such results give us insights about the open issues and help us define possible research paths to the near future.

Keywords: audio music emotion recognition, music information retrieval, emotionally-relevant audio features, musical texture, expressive techniques, bi-modal approaches, music and emotion;

Resumo

Este trabalho aborda o tema do reconhecimento emocional em música utilizando sinais áudio polifónicos. A área do reconhecimento de emoções em música tornou-se um foco de estudo nas últimas duas décadas. Nesta área, a abordagem típica começa com um conjunto de dados e respectivas anotações emocionais geradas por ouvintes. Estes dados, sendo a forma mais comum os sinais áudio, são primeiro processados por algoritmos computacionais para extracção de informação sobre os mesmos (e.g., batidas por minuto ou métricas de energia). De seguida, o conjunto de características extraídas é analisado por algoritmos de aprendizagem computacional, identificando padrões que associam as mesmas às diferentes emoções associadas. O resultado final é um modelo que utiliza as regras aprendidas para identificar a emoção numa nova música ainda desconhecida.

Embora vários investigadores tenham abordado o tema, consideramos que existem dois problemas principais que se mantêm em aberto e contribuem para a falta de progresso nesta área. Primeiro, faltam conjuntos de dados de qualidade, tamanho considerável e livre acesso que sejam adoptados como testes-padrão deste ramo de investigação e assim facilitem a comparação de trabalhos. Para além disso, e não menos importante, são necessários novos algoritmos computacionais capazes de extrair do sinal áudio características musicais emocionalmente relevantes. Na base desta ideia, está o facto de a grande maioria dos trabalhos anteriores ser mais focada na componente de classificação computacional, limitando-se durante a extracção de características a utilizar algoritmos criados para outros problemas (e.g., reconhecimento de fala). Este trabalho tem como principal objectivo o de atacar estes problemas.

A extracção de características emocionalmente relevantes a partir de sinais áudio requer um conhecimento sólido em diversas áreas. Assim, este trabalho começou com uma revisão da literatura nas áreas da música e da emoção. Estas serviram de base para perceber os diferentes paradigmas na classificação de emoções, as várias componentes e dimensões musicais e identificar as relações que são conhecidas entre dimensões musicais e respostas emocionais específicas. De seguida, foram analisados vários dos algoritmos computacionais existentes para extracção de características de áudio, associando cada um destes com uma das oito dimensões musicais possíveis: melodia, harmonia, ritmo, dinâmica, timbre (ou tom da cor), técnicas de expressividade, textura e forma. Como resultado, verificámos que dimensões como a textura e forma musical ou técnicas de expressividade são apontadas como relevantes emocionalmente mas poucos são os algoritmos que tentam capturar alguma desta informação.

De forma a mitigar esta lacuna, foram propostos vários algoritmos para extrair características musicais. Estas começam por utilizar trabalho anterior, transformando o sinal áudio numa estimativa das notas que representam a melodia principal. Através

destas, são extraídas diversas métricas, e.g., presença de *glissando*, informação sobre articulação, variações de dinâmica, entre outras. Para avaliar a influência destas no reconhecimento emocional, foi criado um conjunto de dados de 900 excertos musicais anotados em quatro classes (quadrantes) e devidamente balanceados.

Os resultados experimentais demonstram que a adição das características propostas melhora a classificação de forma estatisticamente significativa. Além disso, foi também medida a influência das várias características, levando a uma melhor compreensão de como as diferentes dimensões musicais influenciam estados emocionais específicos. Estas permitem traçar alguns caminhos para investigação futura, uma vez que o problema do reconhecimento emocional em música está longe de estar resolvido.

ACKNOWLEDGEMENTS

"Gratitude is not only the greatest of virtues but the parent of all others."

Marcus Tullius Cicero (January 106 BC – December 43 BC)

This work would have been impossible without the aid and support of several people over the last years. Among those, three stand out due to their continuous inspiration and support during this adventure and to them I dedicate this work: my mother Lurdes, my girlfriend and now wife Joana, and my supervisor and friend Rui.

To my mom, for the unconditional love, education and sacrifices that led me to this point, for the interesting and long-lasting conversations and lastly but not least important, the delicious meals: I owe it all to you! To my wife Joana, for all the love and affection (and infinite hugs). Thank you for all our shared memories throughout these years, as well as the patience and support even when I was more absent. I love you! I am also profoundly grateful to my supervisor, Professor Rui Pedro Paiva, who guided me during this long journey, sharing his technical knowledge but also his contagious joy and in the process becoming a true friend.

I would like to thank the Center for Informatics and Systems of the University of Coimbra (CISUC), where this work was carried out, for the logistical and financial support. I would also like to thank to the CISUC researchers and staff, especially Ricardo Malheiro and Bruno Rocha, with whom I shared ideas and work. A very special gratitude goes out to my PhD colleagues and friends, particularly to Nuno and Maryam for the research tips and ideas, as well as the random conversations about life in general.

Finally, just like a marathon, a work of this kind requires motivation, resilience and perseverance to overcome many unforeseen challenges. This was only possible with the support of my father, sister, close family and friends which helped ensure my mental sanity. Heartfelt thanks go to my grandmother Maria de São José for keeping our family reunions alive and always ensuring the excessive amounts of food. I am very grateful to my closest friends, the gang who has always been there, and always will be, for the lunches at Sereia do Mondego, super fun road trips and so on: Sam, Carla (asuka), Cioga, Cláudio (Gordo!), Sérgio (kid), Daniel (miúdo) and Carlota (and her mother Rita). I am also grateful to my pharmacist friends for the lunches, dinners and trips around the world. To my Spanish friend Ana for the long distance chats, my training buddy Laura, as well as Bia, João and even Maçãs for some of the most memorable adventures as a PhD student.

CONTENTS

LIST OF FIGURES

List of Tables

List of Algorithms

ABBREVIATIONS

AMC	Audio Mood Classification	(defined on page 4)
ANEW	Affective Norms for English Words	(page 168)
APG	Automatic Playlist Generation	(page 239)
ASR	Average Silence Ratio	(page 90)
AV	Arousal and Valence	(page 30)
AVD	Arousal, Valence and Dominance	(page 122)
BOW	Bag of Words	(page 141)
BPM	Beats per Minute	(page 76)
CAL500	Computer Audition Lab 500 Dataset	(page 107)
CAL500exp	Computer Audition Lab 500 Expansion Dataset	(page 113)
CCRF	Continuous Conditional Random Fields	(page 151)
CD	Crescendo and Decrescendo	(page 188)
CNN	Convolutional Neural Networks	(page 157)
CNSL	Chroma Note Space Length	(page 185)
dB	deciBel	(page 74)
DBN	Deep Belief Networks	(page 150)
DEAM	MediaEval Database for Emotional Analysis in Music	(page 118)
DEAP	Dataset for Emotion Analysis using Physiological and Audiovisual Signals	(page 111)
deepGP	Deep Gaussian Process	(page 157)
DWCH	Daubechies Wavelets Coefficient Histograms	(page 145)
EEG	Electroencephalogram	(page 111)
ERB	Equivalent Rectangular Bandwidth	(page 87)
f0	Fundamental Frequency	(page 70)
FD	Fractal Dimension	(page 198)
FFS	Forward Feature Selection	(page 128)
FFT	Fast Fourier Transform	(page 74)
FKNN	Fuzzy K-Nearest Neighbors	(page 138)

GAD	Greek Audio Dataset	(page 114)
GC	Glissando Coverage	(page 194)
GD	Glissando Duration	(page 194)
GDIR	Glissando Direction	(page 195)
GE	Glissando Extent	(page 194)
GEMS	Geneva Emotional Music Scale	(page 29)
GI	General Inquirer	(page 140)
GMD	Greek Music Dataset	(page 115)
GMM	Gaussian Mixture Models	(page 132)
GNGR	Glissando to Non-Glissando Ratio	(page 195)
GP	Glissando Presence	(page 194)
GS	Glissando Slope	(page 194)
GWAP	Game with a Purpose	(page 108)
HCDF	Harmonic Change Detection Function	(page 73)
HFC	High-Frequency Content	(page 86)
HFVC	High-Frequency Vibrato Coverage	(page 197)
HINDR	High Intensity Notes Duration Ratio	(page 188)
HINR	High Intensity Notes Ratio	(page 187)
hop	Hop-size	(page 183)
HPCP	Harmonic Pitch Class Profile	(page 71)
HWPS	Harmonically Wrapped Peak Similarity	(page 78)
KL	Karhunen-Loève	(page 127)
LENDR	Legato Notes Duration Ratio	(page 192)
LINDR	Low Intensity Notes Duration Ratio	(page 187)
LINR	Low Intensity Notes Ratio	(page 187)
LJ2M	LiveJournal Two-million Post dataset	(page 113)
LNDR	Long Notes Duration Ratio	(page 189)
LNR	Long Notes Ratio	(page 189)
LPCC	Linear Predictive Coding Coefficients	(page 88)
LR	Legato Ratio	(page 192)
LSA	Latent Semantic Analysis	(page 141)
MDS	Multidimensional Scaling	(page 115)

MER	Music Emotion Recognition	(page 3)
MEVD	Music Emotion Variation Detection	(page 113)
MFCC	Mel-Frequency Cepstral Coefficient	(page 7)
MINDR	Medium Intensity Notes Duration Ratio	(page 187)
MINR	Medium Intensity Notes Ratio	(page 187)
MIR	Music Information Retrieval	(page 2)
MIREX	Music Information Retrieval Evaluation eXchange	(page 4)
ML	Musical Layers	(page 190)
MLD	Musical Layers Distribution	(page 190)
MLNDR	Medium Length Notes Duration Ratio	(page 189)
MLNR	Medium Length Notes Ratio	(page 189)
MLR	Multiple Linear Regression	(page 143)
MMRD	Music Mood Rating Dataverse	(page 115)
MNN	MIDI Note Number	(page 185)
MTurk	Amazon Mechanical Turk	(page 109)
MTurk240	MoodSwings dataset	(page 108)
ND	Note Duration	(page 189)
NI	Note Intensity	(page 187)
NS	Note Smoothness	(page 186)
NSL	Note Space Length	(page 185)
OBSC	Octave-Based Spectral Contrast	(page 149)
OTR	Other Transitions Ratio	(page 192)
OTNDR	Other Transitions Notes Duration Ratio	(page 192)
PCA	Principal Component Analysis	(page 127)
POS	Part of Speech	(page 141)
Q1 to Q4	Quadrant 1, 2, 3 or 4	(page 32)
RBF	Radial Basis Function	(page 130)
RMLT	Ratio of Musical Layers Transitions	(page 190)
RMS	Root-Mean-Square	(page 79)
RNDT	Ratios of Note Duration Transitions	(page 190)
RNN	Recurrent Neural Networks	(page 157)

SBS	Stepwise Backward Selection	(page 128)
SCF	Spectral Crest Factor	(page 85)
SFM	Spectral Flatness Measure	(page 84)
SNDR	Short Notes Duration Ratio	(page 189)
SNR	Short Notes Ratio	(page 189)
SR	Staccato Ratio	(page 192)
SSD	Statistical Spectrum Descriptors	(page 149)
STFT	Short-Time Fourier Transform	(page 83)
STNDR	Staccato Notes Duration Ratio	(page 192)
SVM	Support Vector Machines	(page 130)
TEINR	Transitions to Equal Intensity Notes Ratio	(page 188)
TELNR	Transitions to Equal Length Notes Ratio	(page 190)
TEPNR	Transitions to Equal Pitch Notes Ratio	(page 186)
TF.IDF	Term Frequency-Inverse Document Frequency	(page 141)
THINR	Transitions to Higher Intensity Notes Ratio	(page 188)
THPNR	Transitions to Higher Pitch Notes Ratio	(page 186)
TLINR	Transitions to Lower Intensity Notes Ratio	(page 188)
TLNR	Transitions to Longer Notes Ratio	(page 190)
TLPNR	Transitions to Lower Pitch Notes Ratio	(page 186)
TSNR	Transitions to Shorter Notes Ratio	(page 190)
VAT	Voice Analysis Toolbox	(page 198)
VC	Vibrato Coverage	(page 197)
VD	Vibrato Duration	(page 197)
VE	Vibrato Extent	(page 197)
VNBF	Vibrato Notes Base Frequency	(page 197)
VNVR	Vibrato to Non-Vibrato Ratio	(page 197)
VP	Vibrato Presence	(page 196)
VR	Vibrato Rate	(page 196)
YangAV	Yang et al. Arousal and Valence dataset	(page 107)
ZCR	Zero Crossing Rate	(page 83)

Chapter 1

INTRODUCTION

Music is engrained in our history since the beginning of humankind, always present in our lives, serving a myriad of purposes both socially and individually. It has been present in fields as diverse as religion, sports, health or war. Across different religions, it has the power to help believers pray and experience the transcendental; in sports and during physical exercise, it is used to motivate athletes and improve performance, or mark a specific exercise cadence; connected to the celebration of life events, from birthdays to weddings and funerals or in specific occasions such as Halloween, Thanksgiving, Bar Mitzvah or Christmas Eve; in the entertainment and health fields, music helps listeners to relax, creating pleasurable moments and changing our mood; even in war scenarios, for instance with percussion instruments to guide marching soldiers or instigate fear and tension in the enemy. Music is a universal form of communication used all over the world, across civilizations and over distinct epochs, conveying emotions and perceptions to the listeners, which may vary between cultures and civilizations.

Nowadays, music is all around us, in advertisement, at our homes and cars via television and radio, at the gym, elevators and supermarkets. Unsurprisingly, it plays a big role in the world economy. The music distribution industry has received a tremendous impulse as a result of technological innovations in the last decades. Factors like the widespread access to the Internet, bandwidth increasing in domestic and mobile accesses or the generalized use of compact audio formats, such as mp3, and streaming services, such as Spotify and YouTube, have contributed to that boom.

As an indication of this fact, in the USA, the music industry runs billions of US dollars per year. As an example, in 2005 it was estimated that Apple iTunes sold approximately 1.25 million songs everyday (TechWhack.com, 2005), achieving a total of over 6 billion songs sold by the end of 2008 (Schonfeld, 2009). At the time, over 10 million songs were available in the iTunes library (Schonfeld, 2009). In 2011 this number rose to 28 million, with the same service also providing videos and applications. In the first quarter of 2011 alone, iTunes store's revenues totaled a record sum of nearly US$1.4 billion (Dilger, 2011), surpassing the barrier of 16 billion songs sold by October, 2011 (Melanson, 2011). Also in Portugal, music shows sold 30 million euro in tickets, in 2007

(Lusa, 2009). In 2017, digital music revenues rose to US$7.8 billion, now accounting for 50% of the global music revenues, while the global recorded music market grew by 5.9%. According to the International Federation of the Phonographic Industry, this growth in revenues has been fueled "through ongoing investment, not only in artists, but also in the systems supporting digital platforms, which has allowed for the licensing of over 40 million tracks across hundreds of services" (IFPI, 2017). In the last years, we have witnessed the widespread massification of very fast wired and wireless broadband, where users with smartphones are always connected. As a result, the digital distribution paradigm has been moving from buying and downloading songs to streaming directly from massive online services, such as Spotify, Apple Music, Google Play Music, Amazon Music Unlimited and even YouTube. As an example, in the second quarter of 2018 Spotify hit 83 million premium subscribers and a total of €1.27 billion in revenue (Schneider, 2018), while Apple Music surpassed the 50 million subscribers (including free trials) (Variety, 2018) becoming a US$10 billion business (Duggan, 2018).

This frenetic growth in music supply and demand uncovered the need for more powerful methods for the automatic retrieval of songs in a given context from such enormous databases. In fact, any large music database, or, generically speaking, any multimedia database, is only really useful if users can find what they are seeking in an efficient manner. Furthermore, it is also important that the organization of such data can be performed as objectively and efficiently as possible (Paiva, 2006, p. 4). Currently, exploring such music repositories is hindered by the available search methods, mostly based on manually added metadata information (e.g., artist, title, genre, album, year), which needs to be already known by the user. Additionally, the user can also rely on social-based information, such as music tags defined by other users (e.g., in platforms such as Last.FM[1], an online social music service), or by recommendations from friends, radio, and so on (e.g., user created playlists or thematic radios in Spotify).

Hence, the necessity of new, more advanced tools, providing new capabilities for easily searching and browsing large music collections based on the needs of specific individuals lead to the emergence of Music Information Retrieval (MIR) as a key research field (Schedl, Gómez, & Urbano, 2014). MIR is a relatively recent research area that is gaining greater and greater awareness due to the present mentioned challenges. Several universities, research institutions and companies (e.g., Phillips, Sony, Gracenote, Fraunhofer, Echo Nest and Spotify) worldwide are investing on this field worldwide. Some popular commercial applications are already available today, e.g., Shazam[2], based on music identification technologies.

However, we must bear in mind that "music's preeminent functions are social and

[1] https://www.last.fm

[2] http://www.shazam.com/music/web/about.html

psychological", and so "the most useful retrieval indexes are those that facilitate searching in conformity with such social and psychological functions. Typically, such indexes will focus on stylistic, mood, and similarity information" (Huron, 2000). In this direction, studies on music information behavior have identified emotional content of music as "important criterion used by people in music seeking and organization" (Y.-H. Yang & Chen, 2011a, p. 2), an idea supported by 28.2% of the participants surveyed in (Lee & Downie, 2004). Moreover, users on Last.FM social music tagging website use emotions as the third most frequent tag (after genre and locale) (Lamere, 2008). Such needs opened the door to the appearance of Music Emotion Recognition (MER) as a sub-area of MIR. This thesis aims, therefore, to offer novel contributions to address the current MER problems, as will be presented in the following sections.

In this opening chapter, we present the motivation, objectives, results and main contributions of this research work, as well as the overall organization of the thesis. This chapter is structured as described in the following paragraphs.

Section 1.1. Motivation and Scope

We begin by introducing the motivation and scope of this work. The problem of music emotion recognition is presented and some of its research areas are described. The relevance of applications of emotion recognition in music is then discussed.

Section 1.2. Objectives and Approaches

In the second section, we describe our main objectives and briefly introduce the employed approaches.

Section 1.3. Results and Contributions

Next, the main contributions accomplished with this work are summarized. The publications that resulted from this work are also listed and briefly described.

Section 1.4. Thesis Outline

Finally, we end this chapter with the structure of the document by briefly resuming each chapter.

1.1. Motivation and Scope

As abovementioned, in recent years the growth rate of the electronic music delivery industry has been tremendous. It is expected that this frenetic growth will continue in the next years at an even higher rate as we move to a more global world each day. This demands more advanced, flexible and user-friendly search mechanisms, adapted to the requirements of individual users. In particular, methods relying on the emotional content of music play a significant role, providing much more refined methods for a better browsing and filtering of such databases.

In the music information retrieval field, several works have already approached music emotion synthesis and emotion analysis. From these, the first dealt with MIDI or symbolic representations (Katayose, Imai, & Inokuchi, 1988), while nowadays most tackle the problem of emotion detection in audio music signals. Presently, there is already a significant corpus of research works on different perspectives of MER, e.g., classification of song excerpts (Feng, Zhuang, & Pan, 2003; Laurier & Herrera, 2007; Y.-H. Yang, Lin, Su, & Chen, 2008), emotion variation detection (L. Lu, Liu, & Zhang, 2006), automatic playlist generation (Flexer, Schnitzer, Gasser, & Widmer, 2008), exploitation of lyrical information (Malheiro, Panda, Gomes, & Paiva, 2018), cross-cultural and cross-dataset works (Hu & Yang, 2017) and multimodal approaches (Panda, Malheiro, Rocha, Oliveira, & Paiva, 2013).

Although this field has received increasing attention in recent years, limitations can be found and several problems are still open, since it still is a fairly recent research topic. Namely, the lack of a consensual and public dataset and the need to further exploit emotionally-relevant acoustic features. Most of the attention in recent studies has been on different perspectives, datasets and improved machine learning techniques while applying already existent audio features developed in other contexts, such as speech recognition or music genre classification. Particularly, we believe that features specifically suited to emotion detection are needed to narrow the so-called semantic gap (Celma, Herrera, & Serra, 2006) and their lack hinders the progress of research on MER. Moreover, reality shows that the state-of-the-art solutions are still unable to accurately solve simple problems, such as classification with few emotion classes (e.g., four to five). This is supported by both existing studies (Y. E. Kim et al., 2010; X. Yang, Dong, & Li, 2017) and the small improvements observed in the 2007-2017 Music Information Retrieval Evaluation eXchange (MIREX) Audio Mood Classification (AMC) task[3] results, an annual comparison of MER algorithms. There, the best algorithm achieved 69.8% accuracy in a task comprising 5 categories[4]. Moreover, this score has remained stable for several

[3] http://www.music-ir.org/mirex/

[4] http://www.music-ir.org/mirex/wiki/2017:MIREX2017_Results. Moreover, as it will be discussed in Section 5.1, several limitations have been identified in this dataset.

years, which calls for methods that help breaking this so-called "glass ceiling" (Celma et al., 2006).

Besides its usefulness in the music distribution industry, the range of applications of music emotion recognition is wide and varied:

- Emotion-based playlist generators and music selectors. Such tools could give the possibility for users exercising or instructors to choose what kind of tracks they would like to listen to, ranging from high tempo, fast songs to calm and relaxing songs, to be used in meditation sessions;
- Advertisement, television and music industry. These areas could make use of these capacities to find songs that match a desired emotional context, instigating fear, anger, joy or captivating the attention;
- Call center waiting music. Call centers that tend to have clients waiting in line listening to the same classical music excerpt over and over could now automatically pick some more recent songs from alternative genres that would adjust to the objective of maintaining the costumer happily waiting;
- Gaming industry. MER mechanisms would permit searching for the right sound to apply in specific moments to increase tension, mark a moment of happiness, anger, revenge and other similar emotions frequently present in games;
- Health informatics. MER may also find applications in the clinical field, such as the motivation to compliance to sport activities prescribed by physicians, as well as stress management.
- Personal use. Any regular person who, after an exhausting day wants some relaxing music, songs that will cheer him/her up after some sad events.

1.2. Objectives and Approaches

Overall, this study addresses the automatic analysis and classification of emotional content in musical pieces. Despite the broad range of opportunities of such a system, a satisfactory solution to the problem is yet to be discovered, probably with several years of research work ahead. During the last decade, the MIR field has gained substantial research interest. Still, research in music emotion recognition received less attention in part due to the subjectivity and ambiguity of the field.

Previous work on the analysis and classification of emotion in music was based on solutions used successfully in other MIR areas, namely on other classification problems, such as genre classification. Over the years, several authors proposed solutions using classification and regression approaches (e.g., (Aljanaki, Yang, & Soleymani, 2017; Feng et al., 2003; Hu & Yang, 2014, 2017; L. Lu et al., 2006; Y.-H. Yang, Lin, Su, et al.,

2008)). Most of these differed in the audio features and machine learning techniques employed, testing combinations proposed before in distinct fields such has music genre classification or speech recognition. Few have tackled what we believe to be, presently, one core problem: proposing novel and adequate emotion-related features and related strategies. As a consequence, the improvements in MER over the last years have been low. The best performing emotion classification system in MIREX'2007 achieved 61.7%[5] accuracy. In 2010, the best algorithm attained 64.2%[6] accuracy and in 2011 topped at 69.5%[7]. Until 2017 the best performing solution to this challenge obtained 69.8%[8], representing an improvement of 8.1% in ten years, with virtually no improvement in the last six. Furthermore, in approaches based on continuous (rather than discrete) models of emotions, a similar ceiling is present, with weak results particularly the "low accuracy of valence", which is still "an unsolved problem in MER" (X. Yang et al., 2017).

Our main objective is then to offer contributions to narrow the semantic gap between the audio signal (and the usually extracted low-level audio features) and the high-level cognitive/perceptual features, in this particular case perceived emotion.

We start by presenting a thorough review of the literature. This comprehensive review serves to: a) understand what emotions are and how can emotions be classified according to psychology research; b) recognize the main dimensions that relate and define music, according to musicology research; c) understand what relations have been found to exist between these musical dimensions and emotions; d) make a survey of the existent audio features implemented in state-of-the-art audio frameworks and how they relate to existent musical dimensions; e) summarize the existent MER approaches, understanding how they compare and their capabilities and limitations.

Several of the analyzed studies identified relations between musical attributes and emotions states. Some of these musical characteristics associated with emotion are: articulation, dynamics, harmony, interval, loudness, melody, mode, musical form, pitch, rhythm, timbre, timing, tonality and vibrato (Friberg, 2008; Gabrielsson & Lindström, 2001; Juslin & Laukka, 2004; Juslin & Timmers, 2011; Laurier, Lartillot, Eerola, & Toiviainen, 2009). Some illustrative examples are: 1) major modes are frequently related to emotional states such as happiness or solemnity, whereas minor modes are often associated with sadness or anger (Gabrielsson & Lindström, 2001; Lindström, 2006); 2) simple, consonant, harmonies are usually happy, pleasant or relaxed. On the contrary, complex, dissonant, harmonies relate to emotions such as excitement, tension or sadness, as they create instability in a musical motion (Juslin & Laukka, 2004). However,

[5] http://www.music-ir.org/mirex/abstracts/2007/MIREX2007_overall_results.pdf

[6] http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results

[7] http://www.music-ir.org/nema_out/mirex2011/results/act/mood_report/summary.html

[8] http://www.music-ir.org/nema_out/mirex2017/results/act/mood_report/summary.html

by analyzing the existent audio features we concluded that:

- The majority of computational extractors available nowadays are low-level, related with tone color or timbre, capturing information about the frequency spectrum;
- Most of these have been created to solve other problems, such as Mel-Frequency Cepstral Coefficients (MFCCs) to speech recognition;
- A lower number of features are related with the remaining musical dimensions, where expressive techniques, musical texture and musical form are the most underrepresented.

Based on the identified problems we propose a set of novel, higher-level, emotionally-relevant audio features to improve the MER field. As noted, many of the existent features are low-level. However, we naturally rely on higher-level cues such as melodic lines, notes and scores to assess musical dimensions such as harmony, melody, articulation or texture. Thus, the first step consists in deriving MIDI notes from the existing audio waveform. While such a task is still an open research problem, we believe that estimating musical attributes such as predominant melody lines, even if imperfect, gives important information currently missing from MER. To this end we built on previous works by Salomon et al. (2012) and Dressler (2016) to estimate predominant fundamental frequencies and saliences. The resulting pitch trajectories are then segmented into individual MIDI notes based on previous work by Paiva el al. (2006).

From the obtained notes and multiple contours estimation we then propose several features that cover various musical dimensions. Namely, melody (e.g., register distribution information), dynamics (e.g., crescendo and decrescendo), rhythm (e.g., changes in note durations), musical texture (e.g., information on the number of musical layers) and expressive techniques (e.g., measuring articulation or vibrato).

Besides using the audio signal, we also explore additional sources of information to improve MER. Empirically, we know that the message transmitted by the singer (lyrics) and how it is being transmitted (voice characteristics) may also be relevant. While the voice information is already in the original audio signal, some studies suggest that "using singing voices alone may be effective for separating the "calm" from the "sad" emotion, but this effectiveness is lost when the voices are mixed with accompanying music" and "source separation can effectively improve the performance" (X. Yang et al., 2017). Thus, we extract audio features from the voice-only signal, obtained using source separations techniques (Z.-C. Fan, Jang, & Lu, 2016), which, although still imperfect, may contribute to improve the situation.

As for lyrical content, several authors exploited it using natural language processing (e.g., (Hu & Downie, 2010a; Malheiro et al., 2018; Y.-H. Yang, Lin, Cheng, et al., 2008)), in some cases obtaining better emotion classification results when compared to audio

signals. Hence, we explore this knowledge combining all these approaches into a multimodal solution to the problem.

1.3. Results and Contributions

This work encompasses a number of contributions to improve the state-of-the-art in the MER research field, namely:

- Creation of an audio dataset containing 900 song clips, and proposal of a semi-automatic methodology, annotated following the Russell's circumplex model quadrants and enriched with relevant metadata (e.g., artist, title, year, album, genre and additional emotion information, besides the main quadrant-emotion) which can be used to explore new MER perspectives (e.g., multi-label classification, regression, as discussed in Section 4.1);
- A review of the existing musical elements, organized in eight major dimensions (melody, harmony, rhythm, dynamics, tone color or timbre, expressive techniques, musical texture and musical form) and the associations that have been found in the literature between these elements and emotional responses.
- A knowledge-base regarding audio features available in state-of-the-art audio frameworks and their relation with the eight employed musical dimensions;
- Novel emotionally-relevant audio features related with the most underrepresented musical dimensions (namely, expressivity and musical texture), which added to the existing achieved a 9% improvement in F1-Score when compared to a similar number of standard-only features;
- The uncovering of possible relations between musical elements (and features) and emotion responses, namely the weight of specific features and musical dimensions to each emotion quadrant, the possible influence of voice acoustics to valence (in low arousal situations) or the importance of expressive techniques and texture elements, as discussed in Section 4.6;

These contributions were published in the following journal and conference articles. Two publication metrics are reported when available: the impact factor in the publication year (if available), according to Clarivate Analytics / Thomson Reuters[9], and the quartile ranking (1 to 4) for the most relevant research area according to Scimago[10].

[9] http://jcr.incites.thomsonreuters.com/JCRLandingPageAction.action

[10] https://www.scimagojr.com/

Journal Papers

(P1) Panda, R., Malheiro, R., & Paiva, R. P. (2018). Novel Audio Features for Music Emotion Recognition. *IEEE Transactions on Affective Computing* (accepted for publication).
DOI: http://doi.org/10.1109/TAFFC.2018.2820691
Impact Factor (2017): 4.585 (not yet available for 2018)
Quartile (2017): Q1 (Human-Computer Interaction)

(P2) Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2018). Emotionally-Relevant Features for Classification and Regression of Music Lyrics. *IEEE Transactions on Affective Computing*, *9*(2), 240–254.
DOI: http://doi.org/10.1109/TAFFC.2016.2598569
Impact Factor (2017): 4.585 (not yet available for 2018)
Quartile (2017): Q1 (Human-Computer Interaction)

(P3) Panda, R., Rocha, B., & Paiva, R. P. (2015). Music Emotion Recognition with Standard and Melodic Audio Features. *Applied Artificial Intelligence*, *29*(4), 313–334.
DOI: http://doi.org/10.1080/08839514.2015.1016389
Impact Factor (2015): 0.54
Quartile (2015): Q3 (Artificial Intelligence)

Conference Papers

(P4) Panda, R., Malheiro, R., & Paiva, R. P. (2018). Musical Texture and Expressivity Features for Music Emotion Recognition. In *19th International Society for Music Information Retrieval Conference – ISMIR 2018*. Paris, France.
DOI: n.a.

(P5) Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Classification and Regression of Music Lyrics: Emotionally-Significant Features. In *8th International Conference on Knowledge Discovery and Information Retrieval – KDIR 2016*. Porto, Portugal.
DOI: http://doi.org/10.5220/0006037400450055

(P6) Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Bi-modal music emotion recognition: Novel lyrical features and dataset. In *9th International Workshop on Music and Machine Learning – MML 2016 – in conjunction with*

the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2016. Riva del Garda, Italy.
DOI: n.a.

(P7) Panda, R., Malheiro, R., Rocha, B., Oliveira, A., & Paiva, R. P. (2013). Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. In *10th International Symposium on Computer Music Multidisciplinary Research – CMMR 2013* (pp. 570–582). Marseille, France.
DOI: n.a.

(P8) Panda, R., Rocha, B., & Paiva, R. P. (2013). Dimensional music emotion recognition: Combining standard and melodic audio features. In *10th International Symposium on Computer Music Multidisciplinary Research – CMMR 2013* (pp. 583–593).
DOI: n.a.

(P9) Rocha, B., Panda, R., & Paiva, R. P. (2013). Music Emotion Recognition: The Importance of Melodic Features. In *6th International Workshop on Music and Machine Learning – MML 2013 – in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2013*. Prague, Czech Republic.
DOI: n.a.

(P10) Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2013). Music Emotion Recognition from Lyrics: A Comparative Study. In *6th International Workshop on Music and Machine Learning – MML 2013 – in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2013*. Prague, Czech Republic.
DOI: n.a.

(P11) Panda, R., & Paiva, R. P. (2012). MIREX 2012: Mood Classification Tasks Submission. In *8th Music Information Retrieval Exchange – MIREX 2012, as part of the 13th International Society for Music Information Retrieval Conference – ISMIR 2012*. Porto, Portugal (invited short paper, no peer-review).
DOI: n.a.

(P12) Panda, R., & Paiva, R. P. (2012). Music Emotion Classification: Dataset Acquisition and Comparative Analysis. In *15th International Conference on*

Digital Audio Effects – DAFx 12. York, UK.
DOI: n.a.

(P13) Panda, R., & Paiva, R. P. (2012). Music Emotion Classification: Analysis of a Classifier Ensemble Approach. In *5th International Workshop on Music and Machine Learning – MML 2012 – in conjunction with the 19th International Conference on Machine Learning – ICML 2012*. Edinburgh, UK.
DOI: n.a.

(P14) Panda, R., & Paiva, R. P. (2011). Automatic Creation of Mood Playlists in the Thayer Plane: A Methodology and a Comparative Study. In *8th Sound and Music Computing Conference – SMC 2011*. Padova, Italy.
DOI: http://doi.org/10.5281/zenodo.849887

(P15) Panda, R., & Paiva, R. P. (2011). Using Support Vector Machines for Automatic Mood Tracking in Audio Music. In *130th Audio Engineering Society Convention – AES 130*. London, UK.
DOI: n.a.

(P16) Cardoso, L., Panda, R., & Paiva, R. P. (2011). MOODetector: A Prototype Software Tool for Mood-based Playlist Generation. In *3º Simpósio de Informática – INForum 2011*. Coimbra, Portugal (national conference).
DOI: n.a.

1.4. Thesis Outline

Chapter 1: Introduction

In this introductory chapter, we present the motivation, objectives and main contributions of this research work.

Chapter 2: Music and Emotion

Music emotion recognition is an interdisciplinary field resorting to knowledge from distinct areas such as computer science, artificial intelligence, music psychology and musicology. This chapter surveys the existent knowledge regarding music and emotion and the possible relations between them.

Section 2.1 serves as introduction, discussing the definitions of emotion and its relation with music. Following this, Section 2.2 introduces the most known and well-

accepted emotion classification models, a main research area in psychology. In Section 2.3 we go from emotion to the music field, describing the typical musical elements and organizing them in eight typical musical dimensions encompassing the various musical characteristics. Finally, Section 2.4 combines music and emotion by introducing the relations between them that have previously been identified by researchers.

Chapter 3: Music Emotion Recognition Literature Review

The overview of the music and emotion fields is followed by an in-depth review of the MER state-of-the-art.

Section 3.1 presents a summary of the computer frameworks used to analyze and extract musical features from audio. In addition, it comprehensively describes the audio extractors available in state-of-the-art frameworks, relating then with the musical dimensions identified in Chapter 2.

Following, Section 3.2 critically reviews the most relevant MER works to date. This section begins with a general explanation of the typical MER system based on audio features, which comprises three distinct parts: 1) dataset acquisition, reviewing the major datasets existent in the field; 2) feature extraction and selection; and finally 3) classification and evaluation.

After the general explanation, the Section ends with a historic contextualization of the progress achieved in the MER field by describing some of the most relevant works over the last three decades.

This review reinforced our understanding that MER research has been focused on different classification or regression strategies and exploring different MER perspectives, while neglecting research in novel features that better capture emotional information in audio music.

Chapter 4: A Novel System for Music Emotion Recognition: New Dataset and Audio Features

Building on the gathered knowledge, Chapter 4 presents the research work carried out to address the identified problems by proposing novel emotionally-relevant audio features and a dataset.

The first step in this direction is the construction of a sizeable dataset used to validate our work. Section 4.1 describes the entire procedure, which includes data gathering from the AllMusic platform[11], transformation from emotion tags (categorical view) to Russell quadrants (dimensional view), filtering and validation.

[11] A comprehensive online music guide available at https://www.allmusic.com/

Section 4.2 describes the novel proposed features. To this end, we link the musical dimensions organized in Section 2.3, their relations to emotional responses approached in Section 2.4, and the current panorama in standard audio features of Section 3.1. This approach allowed us to identify the emotionally relevant musical dimensions that lack computational extractors. These include musical texture, expressive techniques such as articulation or vibrato and others, for which audio features are proposed.

A typical MER classification strategy is then followed, described in Sections 4.3 and 4.4, to assess the performance of the novel features when compared to the existent features.

Finally, the classification results obtained in each of the problems tested are discussed in Section 4.5, showing that emotion recognition is significantly improved with the addition of novel features. To conclude, the most influential features to each problem are discussed in Section 4.6, uncovering some interesting relations. Namely, the influence of the novel musical texture features in quadrants classification, the high number of features related with expressive techniques that contribute to valence classification, or how specific feature characteristics seem to be more related with each quadrant (e.g., Q1 with rhythm, Q2 with dissonance, and how the voice-only signal seems important for the remaining two quadrants).

Chapter 5: Other Experiments

In addition to the main contributions described in previous chapters, a number of additional experiments were conducted. These served to test different ideas, improve on existing approaches and build a solid foundation in MER. Some of these consisted in the construction of other datasets, bi-modal approaches exploring audio and lyrics or the analysis of existent works and datasets.

Chapter 6: Conclusions and Perspectives

To close, we summarize the main conclusions of this thesis and mention potential directions for future research based on the difficulties identified or faced.

Bibliography

The entire set of references used and cited in this thesis is listed under this chapter.

Appendices

In Appendix A, we provide an extended description of various musical characteristics composing each of the eight musical dimensions introduced in Section 2.3.

In Appendix B, the list of emotion tags used in our novel dataset are listed. This information is organized by quadrants, providing the number of songs for each emotion tag and quadrant, as well as the total values.

Finally, Appendix C complements Section 4.6, providing additional information regarding the relevance of the analyzed features for MER. This includes graphical representations of the feature weights by musical dimension as well as the best features for organized by musical dimension for each MER problem.

Chapter 2

Music and Emotion

In this chapter, essential knowledge about musical dimensions, emotions and the relations between them is discussed, building the foundations needed for novel contributions to the music emotion recognition (MER) field.

Section 2.1. What is Emotion?

With that in mind, we first explore the definitions of emotion from a scientific perspective and how emotion in music can be regarded: as the emotion expressed by the performer, the emotion perceived by listeners, or the emotion felt by the listener (induced).

Section 2.2. Emotion Taxonomies

After understanding what emotion is, we delve on how it can be classified. Here, we explore the two main views, categorical or dimensional, as well as some alternative approaches and evaluate how adequate they are for MER research.

Section 2.3. Musical Dimensions

Next, we review the major musical elements, organizing them into eight dimensions: melody, harmony, rhythm, dynamics, tone color (or timbre), expressive techniques, musical texture and musical form.

Section 2.4. Relations between Music and Emotions

Finally, we connect the dots between music and emotions from a music psychology perspective. As a result, we build a well-grounded knowledge base on the musical elements associated with specific emotional responses.

.

2.1. What is Emotion?

The word "emotion" has been a scientific term and research subject in the psychology field only since the 19th century, in part due to the influential article "What Is an Emotion?" by William James (1884). Nowadays, more than a century later, scientists are yet to find a consensual answer to James' question, with some researchers even sustaining that the term is "ambiguous and has no status in science", and thus should be dropped (Izard, 2010b, pp. 367–368).

Nowadays, Merriam-Webster dictionaries[12] define emotion as:

1. **a** *(obsolete)* : *disturbance*
   b : *excitement*
2. **a** : the affective aspect of consciousness : *feeling*
   b : a state of feeling
   c : a conscious mental reaction (such as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body

To better understand emotion and the discussion around its definition we should know its history. The word "emotion" was introduced in the English language around the 16th century. The first references in English literature were the Montaigne essays translated from French. In them, the translator apologized for the introduction of various "uncouth terms" from French, which included the word "emotion" (Montaigne, 1603, p. A5f.). At the time and until the 18th century, emotion denoted "physical disturbance and bodily movement", which could be "commotion among a group of people (as in the phrase "public emotion"), or a physical agitation of anything at all, from the weather, or a tree, to the human body" (Dixon, 2012).

According to the etymology of the word "emotion"[13], it originated from the assimilation of Latin words "ex-" – meaning out, and "movere" – to move, resulting in "emovere", or move out, remove, agitate. It was imported into Old French as "émouvoir" – to stir up, excite, and later "émotion", finally appearing in English (Figure 2.1).

At that time, scholars used several distinct words such as "passions", "affections", "sentiments" or "appetites" to refer to what we mean today as emotions. This need for distinction stemmed back from ancient debates between Stoicism and Catholicism, especially by the desire of theologians Augustine of Hippo and Thomas Aquinas to provide an alternative to the moral philosophy of the Greek and Roman Stoics (Dixon, 2003).

[12] https://www.merriam-webster.com/dictionary/emotion

[13] https://www.etymonline.com/word/emotion

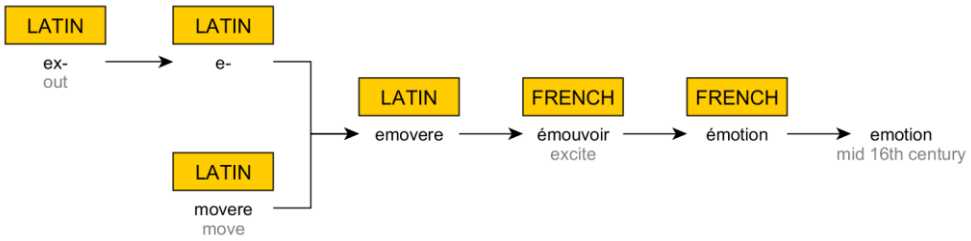


Figure 2.1: Origin of the word "emotion" in the English language.

Historically, Stoics had famously avoided all kinds of passions, considering them as "diseases of the soul, from which the wise man could be cured by the application of calm reason" (Dixon, 2012). Thus, a Stoic aimed to maintain his composure and peace of mind, his *apatheia*, while still enjoying milder positive feelings (Sorabji, 2002). The Christian theologians had a mixed opinion about this. While they considered that passions were indeed evil, conflicting with reason and leading to sin, they did not support that *apatheia* was the goal. As Augustine stated "someone who no longer trembled from fear or suffered from sorrow would not have won true peace, but would rather have lost all humanity" (Augustine (426 A.D.), 1871, bk. XIV.9)[14]. Therefore, the need for distinction between the troubling emotions – passions, lusts, desires – which Christians should avoid (caused by senses), and the more virtuous affections such as love and compassion, which should be pursued (caused by our will).

Up until the 18th century, all the new ideas about sentiments supported the distinction between the two categories (DeJean, 1997; Dixon, 2003). However, in the 19th century, the Edinburgh professor of moral philosophy Thomas Brown, in his lectures published in 1820, resumed all the "appetites," "passions," and "affections" categories in a single theoretical category in mental science: the "emotions" (Dixon, 2003, p. 109). About the definition of "emotion", Brown stated: "Perhaps, if any definition of them be possible, they may be defined to be vivid feelings, arising immediately from the consideration of objects, perceived, or remembered, or imagined, or from other prior emotions." (Brown, 2010, pp. 145–146).

The wide category was later explicitly defined in the first modern book of psychology as: "Emotion is the name here used to comprehend all that is understood by feelings, states of feeling, pleasures, pains, passions, sentiments, affections" (Bain, 1859, p. 3). Following researchers compiled hundreds of discrete feeling states that were now included into the category (McCosh, 1880).

Another key figure in the definition of emotion was Edinburgh physician and phi-

[14] The original work was published in Latin by Aurelius Augustine, also known as Saint Augustine of Hippo in 426 A.D. The citation is taken from the 1871 translation by Marcus Dods.

losopher Charles Bell, which was the first to give a constitutive role to bodily movements. Bell considered "emotions" as movements of the mind, where organs such as the heart or lungs were not only the "expression" but could have a role in causing the emotions, defining emotions as: "certain changes or affections of the mind, as grief, joy, or astonishment," which could become visible through "outward signs" on the face or body (Bell, 1824, pp. 18, 20). Other notorious thinkers such as Darwin supported the idea, stating that: "Most of our emotions are so closely connected with their expression, that they hardly exist if the body remains passive" (Darwin, 1872, p. 239).

Founded on this knowledge, it is clear that when William James famously asked "What is an emotion?" (James, 1884) he was not entering the centuries old discussion, but in search of a definition to the psychological category that was initiated only decades before. In his own answer, he defined emotions as "vivid mental feelings of visceral changes brought about directly by the perception of some object in the world" (Dixon, 2012).

In the following years, James's theory was highly criticized, accused of, among other inconsistencies, being unable to differentiate between different emotions or between emotions and non-emotions, giving excessive weight to body related emotion components and neglecting the cognitive factors in emotion generation (Dixon, 2003; Ellsworth, 1994; Feinstein, 1970). This led the author to write a new statement on his view (James, 1994)[15], that basically negated his initial theory.

Consequently, as summarized by Dixon: "by the 1890s, although the idea that "emotion" as the name of a psychological category had become entrenched, the nascent psychological community had neither an agreed definition of the extent of the category, nor a shared idea of the fundamental characteristics of the states that fell within it." (Dixon, 2012). This new discipline and term, which comprised a category covering much of our mental life, was by this time widely adopted and, understandably, a source of discussion, since the connections between the mind (thought) and the body (feeling) were still uncertain.

The criticism regarding the definition of "emotion" among researchers continued until today (Izard, 2010b, 2010a), showing that, as Brown put it, while "every person understands what is meant by an emotion" (Brown, 2010, p. 145), it is very hard to define it consensually.

In an attempt to achieve a unanimous description of emotion, Izard's surveyed leading contemporary emotion scientists and experts, summarizing the most commonly cited features in one sentence as:

> *"Emotion consists of neural circuits (that are at least partially dedicated), response systems, and a feeling state/process that motivates and organizes*

[15] The reference is a reprint of the original work by James in 1894, which is currently unavailable.

cognition and action." (Izard, 2010b, p. 367).

In addition, although not unanimous, a longer and more complete definition provided by Kleinginna et al. describes emotion as:

> *"Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as perceptually relevant effects, appraisals, labelling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-oriented, and adaptive." (Kleinginna & Kleinginna, 1981).*

Finally, it is important to highlight that while the terms emotion and mood are close and have been used interchangeably in the Music Information Retrieval field, they are not equivalent.

The definition of mood and its relation to emotion given by Amado-Boccara et al. helps understanding the differences:

> *"The conception of mood in cognitive psychology is derived from the analysis of emotion. While emotion is an instantaneous perception of a feeling, mood is considered as a group of persisting feelings associated with evaluative and cognitive states which influence all the future evaluations, feelings and actions" (Amado-Boccara, Donnet, & Olié, 1972).*

In other words, moods differ from emotions in that the former are less specific, less intense, and less likely to be triggered by a particular stimulus or event. To illustrate this, we can be sent into a mood from the happiness of seeing an old friend to the anger of discovering betrayal by a partner. We may also just fall into a mood.

In this work the term "emotion" is used preferably, as from our understanding it is more accurate to our problem.

2.1.1. Emotion Context and Subjectivity

Over the last centuries researchers have tried to explain emotion by relating it to single specific elements of human nature (e.g., subjective experiences, psychophysiological changes or human behavior), as described in the previous section. Nowadays it is said that emotions involve five different elements, all required simultaneously (Scherer, 2005): cognitive processes, physiological changes, subjective experiences and expressive

and instrumental behaviors. These components are briefly described in Table 2.1.

Emotion function	*Organismic subsystem and major substrata*	*Emotion component*
Evaluation of objects and events	Information processing (CNS)	Cognitive component (appraisal)
System regulation	Support (CNS, NES, ANS)	Neurophysiological component (bodily symptoms)
Preparation and direction of action	Executive (CNS)	Motivational component (action tendencies)
Communication of reaction and behavioral intention	Action (SNS)	Motor expression component (facial and vocal expression)
Monitoring of internal state and organism-environment interaction	Monitor (CNS)	Subjective feeling component (emotional experience)

Note: CNS = central nervous system; NES = neuro-endocrine system; ANS = autonomic nervous system; SNS = somatic nervous system.

Table 2.1: Emotion components and functions according to Scherer (2005).

While previous studies have shown that basic emotions are universal, experienced across different backgrounds and cultures (Ekman, 1971; Ekman et al., 1987), we also know that experienced emotion can be highly subjective (hence the subjective feeling component). For instance, while we have words to describe basic emotions such as happiness or sadness, these experiences are much more multi-dimensional, since they can vary greatly between different persons and experiences. Are our sad experiences always the same? In addition, many of our emotional experiences are not pure. Instead, it is common to have situations where various emotions are mixed. Experiences such as having a child might be marked by a wide spectrum of emotions from joy and happiness to anxiety or fear, occurring at the same time or alternately.

Context, memory and culture also contribute to these subjective experiences, since these factors influence how emotions affect us. As an example, a specific song may be associated with distinct life events of different persons and thus elicit different, even opposite emotional experiences. Social factors can also enhance emotions, as is the case of music in many concerts or religious ceremonies, but also other events such as rallies, where there is a "contagious" effect rising emotions intensity.

Regarding music, several factors from background, gender, personality and musical

training (Abeles & Chung, 1996), musical taste and musical memory (Hargreaves & North, 1997) all contribute to this issue, explaining why often it is hard to have a consensus on the emotions present in specific songs.

2.1.2. Emotion Types in Music: Expressed, Perceived and Induced

It is known that music transmits emotions, and this is considered its primary purpose (Cooke, 1959). Thus, this is the main reason why we engage with it (Juslin & Laukka, 2004). This process is typically divided into three distinct parts (Gabrielsson, 2001a; Pannese, Rappaz, & Grandjean, 2016):

- Expression – pertains to the expressed emotion, as the emotion that the composer or performer aimed to transmit with the musical piece.
- Perception – concerns the emotion the listener identifies when listening to a song, which may be different from what the composer attempted to express and what the listener feels in response to it.
- Induction – relates to the emotion that is felt (evoked in) by the listener is response to the song.

While the expressed emotion is easier to grasp, created for instance by "a composer adopting certain metaphoric or stylistic devices in order that the score may express certain emotional qualities that reflect the composer's own emotional state" (Pannese et al., 2016), the relation between perceived and inducted emotions has been subject of discussion among researchers. The source of this is the rather complex relation between music and emotions, as demonstrated by the so-called paradox of negative emotion, where music generally characterized as conveying negative emotions (e.g., sadness, depression, anger) is often judged as enjoyable (Pannese et al., 2016). For instance, listeners exposed to sad music often appear to "lack the beliefs that typically go with sadness" (Davies, 2003, pp. 185–186). This separation between the emotion identified by the listener and his emotional response has been the main reason suggesting a separation of music emotion into perception and induction. The former is a "sonic-based phenomenon, tightly linked to auditory perception, and consisting in the listener's attribution of emotional quality to music" (Pannese et al., 2016). It tends to have a high inter-subjective agreement, where different listeners are likely to agree on the identified emotion independently of musical training (Heinlein, 1928), intelligence (Hevner, 1935) or culture (Fritz et al., 2009). The later, induced emotion, is an individual rather than collective phenomenon, regarding the personal emotional experience by the listener while listening to the song (Scherer, 2004). It is often more related with cultural[16], contextual and

[16] This does not invalidate the fact that emotion perception may also be influenced by culture.

cognitive responses as well as individual preferences (Gabrielsson, 2001b; Rentfrow & McDonald, 2010). It is also more controversial: some authors state that "the body of research that purports to support direct induction of emotion by music is recent and unconvincing" (Konečni, 2008, p. 115).

This suggests that the process of emotions induction by music is indirect, combining numerous perceptual and cognitive aspects. Several hypothesis have been proposed to explain it, one of the most well-known states that it is caused by six complementary (i.e., non-mutually exclusive) mechanisms unrelated to music (i.e., pertain to general cognition), ranging from physiological processes such as brain stem reflexes, to cognitive functions such as music expectancy (Juslin & Västfjäll, 2008). Additional mechanisms have been proposed, such as semantic association (Fritz & Koelsch, 2008; Steinbeis & Koelsch, 2008) and others. These build on the conceptual metaphor (CM) theory, which regards metaphor as a process of mapping from a source domain to a target domain (Lakoff & Johnson, 1980a, 1980b), such as related with time (Epstein, 1995), space (Bonde, 2007) and movement (Johnson & Larson, 2003) (e.g., sonic features such as pitch or intervals are often conceptualized in reference to space (Parkinson, Kohler, Sievers, & Wheatley, 2012)). Other hypothesis in the same direction is the multilayered conceptualization of musical expression of emotions whereby a set of basic emotions (i.e., common across cultures) interacts with (context-dependent) additional layers enabling the expression and perception of complex emotions (Juslin, 2013), with metaphors "acting as the hinge between language, emotion, and aesthetic response" (Pannese et al., 2016), as illustrated in Figure 2.2.

This work is focused on emotion perception in music since it is the part that is more intersubjective[17] and largely driven by physical properties and physiological responses, depending less on personal factors.

[17] Intersubjectivity has been used in social science to refer to agreement.

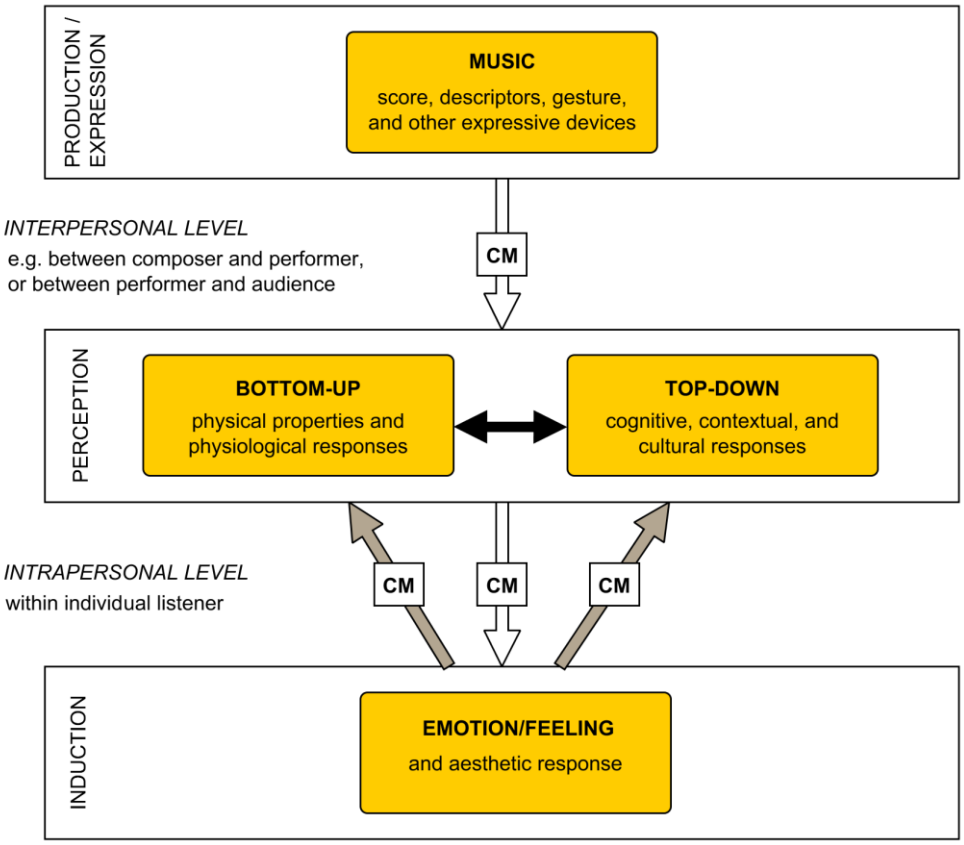


Figure 2.2: Role of conceptual metaphor (CM) in emotion response to music. In the upper part (interpersonal level), CM offers an interface for shared understanding of music in terms of time, space, motion, gesture, and others between performer and listeners. At the intrapersonal level, it enables the transition between emotion perception and emotion induction (adapted from (Pannese et al., 2016)).

2.2. Emotion Taxonomies

Psychology researchers have discussed for long on how emotions can be represented and classified. Several aspects make this task difficult. First its subjective nature, where emotions are regarded as diffuse reactions, that vary from person to person, from moment to moment and also across cultures. Furthermore, there are many different words, across different languages, employed to describe emotional states, some of which are direct synonyms, while others represent small variations. Different persons may have different perceptions of the same stimulus and often use some of these different words to describe

similar experiences. Understandably, there is not one standard, widely accepted, taxonomy for emotions. Therefore, correctly studying and understanding the existent emotion taxonomies and adequately choosing the one that best fits our needs is one important foundation to this work.

Several models have been proposed over the last century by authors in the psychology field. These models can be grouped into two major approaches: the discrete (or categorical) emotion models or the dimensional models of emotion.

2.2.1. Categorical Emotion Models

Categorical models of emotion, also known as discrete models, use words or groups of words to describe an emotion. Several distinct models exist in this category. Some of such models have a stronger psychological and physiological foundation, namely the basic emotions theory, while others were proposed as a more domain-specific solution to the music emotion field, as is the Hevner's adjective clock.

Basic Emotions

In research fields such as psychiatry and neuroscience, the dominant theory of emotions states that humans have a discrete and limited set of basic emotions which is universal and innate (Ekman, 1992; Panksepp, 1998; Tomkins, 1962, 1963). Accordingly, "each of these emotions is independent of the others in its behavioral, psychological, and physiological manifestations, and each arises from activation within unique neural pathways of the central nervous system" (Posner, Russell, & Peterson, 2005). Regarding its origins, each of these emotions has been shaped by evolution and is connected to goal-relevant events (Johnson-Laird & Oatley, 1992).

This approach of basic emotions was derived mostly from research with animals, using neural stimulation and subsequent behavior observation, as well as the opposite (i.e., neural observation after specific behavior) (Panksepp, 1998). However, such approaches have been subject to criticism since specific affective behaviors are not always sufficient or necessary to have specific emotional states (Kagan, 2003). For instance, it is possible to have anxiety without behavior changes, or in the opposite direction, smiling may be attained without obvious changes in the emotional state. Furthermore, it is questionable whether such approach is mapping neural systems related primarily to affective behaviors rather than subjective feelings. Thus, such approach should be replicated with human studies, which has proved intangible (Berridge, 2003).

Some authors have also studied the basic emotions theory in humans using facial expressions, assuming that facial expressions, more specifically patterns of facial innervation and musculature are specific and distinct in each basic emotion (Ekman, 1992;

Ekman, Levenson, & Friesen, 1983). Using this view, Ekman derived a set of six basic emotions: happiness, sadness, fear, disgust, anger, and surprise, as illustrated in Figure 2.3. These emotions are considered the basis from which all the others are built on. From a biological perspective, as abovementioned, this idea is manifested in the belief that there are neurophysiological and anatomical substrates corresponding to the basic emotions. From a psychological perspective, basic emotions are often held to be the primitive building blocks of other, non-basic emotions, which can all be derived from them.

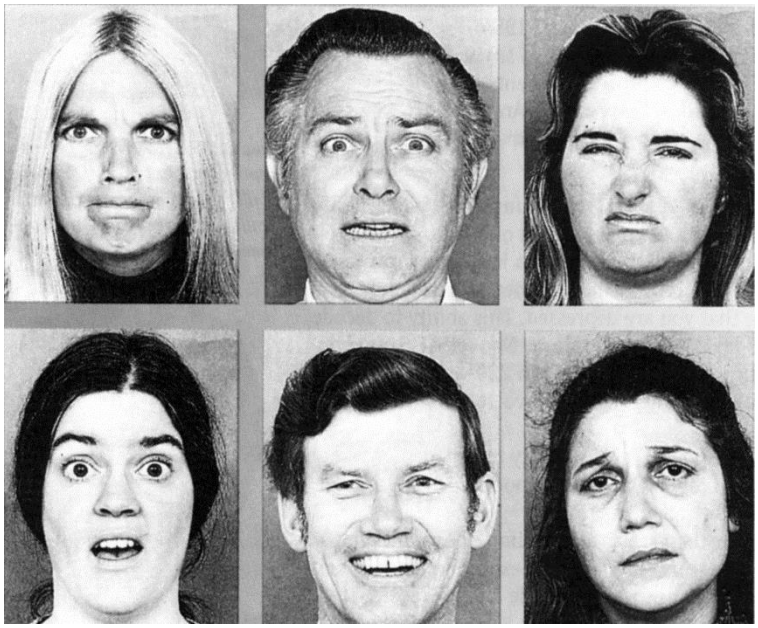


Figure 2.3: Facial expressions representing basic emotions (top, from left to right: anger, fear, disgust; bottom: surprise, happiness and sadness) from (Ekman & Friesen, 2003).

Even though the idea of basic emotions has received support from several studies over the last decades (Panksepp, 1992), other researchers have raised issues with it, stating that "there is no coherent nontrivial notion of basic emotions as the elementary psychological primitives in terms of which other emotions can be explained" (Ortony & Turner, 1990). In addition, regarding the connection between facial expressions and affective states, basic emotions have not been found to be associated with specific patterns of autonomic activation (Cacioppo, Berntson, Larsen, Poehlmann, & Ito, 2000). Moreover, based on the lack of agreement on the name and number of emotion categories representing the basic emotions "suggests that these may be based on linguistic and cultural taxonomies, rather than on emotions themselves" (Zentner & Eerola, 2010).

To the point, basic emotion theory has been primarily based on studying both behavior and expressive manifestations of emotions. Still, a strong corpus of research has suggested that emotions arise from cognitive interpretations of core physiological experiments, instead of being based on a direct map of discrete emotions and the central neural system (Cacioppo et al., 2000; Russell, 2003). Even so, the idea has been adopted in MER research, most likely due to the use of specific words, offering integrity across different studies, and their frequent use in the neuroscience field, related with physiological responses. In some of these studies, part of the original emotions have been replaced with more musically related terms (e.g., changing disgust and surprise[18] to tenderness and peacefulness) (Gabrielsson & Juslin, 1996; Vieillard et al., 2008).

Domain-Specific Approaches

Over time, researchers in the music emotion field have also proposed other categorical models, which were developed specifically to capture emotions in music, e.g., (Hevner, 1936; Hu & Downie, 2007). Proponents of these ideas contested the assumption that emotions in music and the remaining emotions were identical, arguing that models such as the basic emotions theory were not devised to capture music emotions but instead focus on a limited and very specific set of emotions related to our species evolution and survival process (e.g., anger, fear, shame, guilt). On the other hand, emotions evoked by music are expected to be "of a more contemplative kind" (Zentner & Eerola, 2010). This vision has been discussed since the 19th century, where Gurney stated that "the prime characteristic of Music, the *alpha* and *omega* of its essential effect: namely, its perpetual production in us of an emotional excitement of a very intense kind, which yet cannot be defined under any known head of emotion" (Gurney, 1880, p. 120).

While the domain-specific approaches typically contain descriptors that are musically more plausible, their origin is less substantiated. There, the choice of labels is dependent on authors' particular views, as opposed to more scientifically-based psychological models of emotion.

Hevner's Adjective Circle

A widely known discrete domain-specific model of music emotion is Hevner's adjective circle or clock (Hevner, 1936). Kate Hevner is best known for her research in music psychology, being one of the first to do research on the subject of emotions in music. In her research, she concluded that music and emotions are intimately connected, with music often carrying emotional meaning in it. As a result, she proposed a grouped list of adjectives (emotions), instead of using single words (Figure 2.4).

[18] Although the authors of the study removed "surprise", it can be argued that surprise can be both perceived and induced by music and thus the decision is not unanimous.

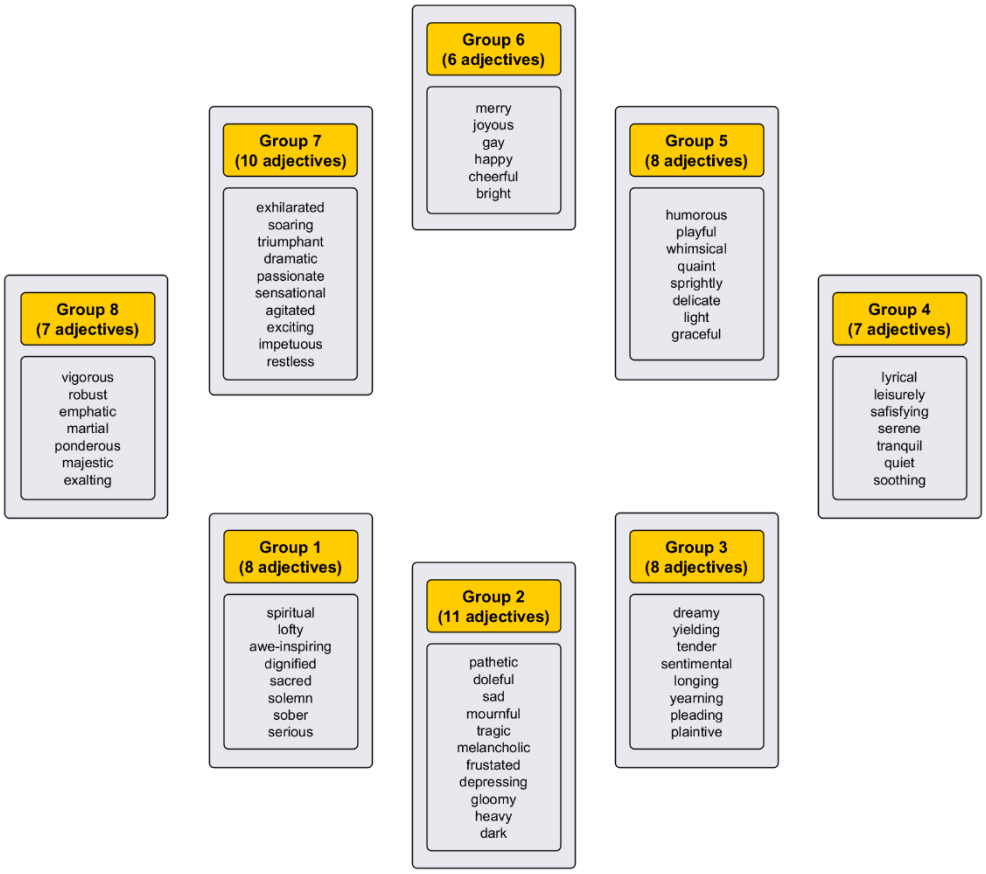


Figure 2.4: Hevner's adjective circle[19].

Hevner's list is comprises 67 different adjectives, organized into eight different groups, with 6 to 11 emotions each, in a circular way. These groups, or clusters, contain adjectives with close meaning, used to describe the related emotional states. Neighboring clusters are emotionally close, deviating slightly until reaching contrasting emotional states in the opposite position (e.g., group 2 opposes group 6).

One of the obvious problems with Hevner's circle is the unbalanced number of terms in each of the eight groups, which may reduce the probability of selection for groups with fewer words. Additionally, it was built specifically for classical music and some of the terms may be infrequently associated with music nowadays (e.g., gay), hindering its usage. Several authors proposed updates to Hevner's adjectives circle, adding new terms and reorganizing the clusters (Campbell, 1942; Farnsworth, 1954; Schubert,

[19] The word "melancholy" was replaced with "melancholic" to maintain consistency.

2003; K. B. Watson, 1942).

MIREX Mood Classification Task Taxonomy

A more recent domain-specific model of emotion has been proposed by Hu et al. (Hu & Downie, 2007) and adopted in the MIREX Mood Classification Task. MIREX[20], acronym of Music Information Retrieval Evaluation eXchange, is an annual comparison of state-of-the-art MIR algorithms held in conjunction with the ISMIR (International Society for Music Information Retrieval) conference[21]. One of the available tasks, Audio Music Mood Classification, consists in a train/test challenge using a private dataset annotated with the abovementioned categorical model.

The MIREX emotion taxonomy was directly derived from songs' metadata provided by the AllMusic service[22], "a popular music database that provides professional reviews and metadata for albums, songs and artists" (Hu & Downie, 2007). At the time, a total of 179 mood tags, "adjectives that describe the sound and feel of a song, album, or overall body of work" were available, "created and assigned to music works by professional editors" (Hu & Downie, 2007).

The taxonomy was built using a three-step process. First, the mood similarity was measured. To this end, the mood tags associated with less than 50 songs and 50 albums were removed, generating a subset of 40 mood tags, related with 2748 albums and 3260 songs. Next, these 40 moods were used to create a 40 x 40 matrix, where each cell contains the number of songs related with the respective moods pair. A second matrix using the number of albums was also derived. The similarity between each pair of moods was then measured by computing the Pearson's correlation between the two rows corresponding to the selected pair. In the second step, the similarity data was clustered using agglomerative hierarchical clustering with Ward's criterion (Berkhin, 2006), leading to two cluster sets, one regarding song moods and the other album moods. Finally, the clustering results were analyzed and the authors identified 29 mood tags "consistently grouped into 5 clusters at a similar distance level", as presented in Table 2.2.

Although used annually to compare advances in the MER field, this MIREX taxonomy (and dataset) suffers from several limitations. To begin with, it lacks support from any psychology studies. It is purely data-driven, based on the annotations by AllMusic experts, but few details are provided about the process, which does not allow for a critical analysis of the annotation process. Moreover, there is semantic overlap (ambiguity) between clusters 2 and 4, and acoustic overlap (based on the analysis of the MIREX dataset)

[20] http://www.music-ir.org/mirex/

[21] http://www.ismir.net/

[22] https://www.allmusic.com/about

between clusters 1 and 5 (Laurier & Herrera, 2007). For illustration, the word fun (cluster 2) and humorous (cluster 4) share the synonym amusing. As for songs from clusters 1 and 5, there are acoustic similarities: both tend to be energetic, loud, and many use electric guitar (Laurier & Herrera, 2007).

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
Rowdy	Amiable / Good natured	Literate	Witty	Volatile
Rousing	Sweet	Bittersweet	Whimsical	Visceral
Boisterous	Fun	Autumnal	Wry	Aggressive
Passionate	Rollicking	Brooding	Campy	Tense / Anxious
	Cheerful	Poignant	Quirky	Intense
			Silly	

Table 2.2: Emotion taxonomy used in the MIREX Audio Mood Classification task.

In addition, according to our experiments, the updated AllMusic data seems to suggest higher inter-cluster similarity than expected. To assess this, we obtained the mood tag data, which has been revised from 179 to 289 adjectives. Furthermore, each AllMusic mood tag also contains a list of similar moods. As an example, the tag fun[23] is similar to boisterous, humorous, quirky, rollicking, rowdy, silly, whimsical, and witty. Examining this supplementary similarity information shows that some clusters (e.g., cluster 1 and 2) have more extra-cluster than intra-cluster similarities, as detailed in Chapter 5.

Geneva Emotional Music Scale

The Geneva Emotional Music Scale (GEMS) is a domain-specific model specifically designed to musically evoked emotions (Zentner, Grandjean, & Scherer, 2008). It was proposed to overcome the limitations of previous models, namely the "failure to distinguish felt from perceived emotion, vague criteria for selection of affect terms, lack of methodological rigor, and the absence of contextualization of the musical emotions within the broader context of emotion research" (Zentner & Eerola, 2010, p. 203).

To address this, four studies were carried out by the authors. The initial two studies were used to gather a compilation of music-relevant emotion terms. To this end, 92 psychology students from Geneva participated in the first study with the objective of creating a "comprehensive list of words genuinely suited to describe experienced or felt emotion" (Zentner et al., 2008). After processing the results, a total of 146 affect terms were selected from the original list of 515 terms.

[23] https://www.allmusic.com/mood/fun-xa0000001006

The second study aim was “to examine which of these terms would be actually relevant in relation to music”. To this end, 262 undergraduate psychology students were asked for their music genre preference between five possible genres: classical, jazz, pop/rock, Latin American, and techno. Next, they were instructed to rate the 146 affect terms regarding: 1) how often they felt a given emotion when listening to their favorite genre; 2) how often they perceived a given emotion when listening to their favorite genre; and finally 3) how often they experienced such emotions in their extra musical everyday life. The participants were instructed to consider only “pure music, without text or lyrics”. As a result, a total of 89 emotion descriptors were identified as musically relevant. This value was later reduced to 66 terms “more than just occasionally experienced or perceived” in several music genres.

One of the interesting findings of the second study was a considerable variability across genre and between perceived and induced emotions, with induced emotions being more positive. As a possible justification, the authors state that “as people move into a mental state in which self-interest and threats from the real world are no longer relevant, negative emotions lose their scope.” (Zentner et al., 2008).

The third and fourth studies’ main objective was to examine whether the 66 emotions induced by music could be differentiated into several sub-units. To this end, 801 questionnaires were filled out by listeners at a music festival comprising various musical genres. There, the subjects were asked to rate the affect terms according to the emotions experienced. Later, confirmatory factor analyses of the gathered data consisting of ratings of emotions evoked by various genres of music were carried out to derive the GEMS model.

The full GEMS contains 45 terms that proved to be consistently chosen, which were also grouped into 9 different categories. These nine emotional scales in turn condense into three "superfactors". In addition to the full scale (GEMS-45), two shorter Scales, the GEMS-25 and the GEMS-9 have also been developed.

Recently, some researchers have disputed the conclusions of this work, where the authors proposed GEMS scale has a more adequate instrument to measure musical emotion than previous arousal and valence (AV) and basic emotion models. Namely, Aljanaki (2016) pointed out several issues: the small size of the original experiment (only 16 musical pieces), the overrepresentation of one genre (only classical music) and the unconventionality of the questions regarding the AV model.

To conclude, a major hurdle in employing categorical models in MER applications is the usefulness of its final result. By classifying two songs with the same adjective or cluster, it is impossible to discriminate between them, even though one might be slightly different than another (e.g., more intense). Moreover, given the very large music databases nowadays available to users, employing such emotion models will typically result

in some (e.g., 5 to 10) still large subsets of songs (e.g., in a one hundred thousand song database, using 10 balanced emotion categories will result in 10 groups of ten thousand songs) that in many contexts may still be too large to be practical for users.

2.2.2. Dimensional Emotion Models

Over the years, researchers have observed that subjects have difficulties describing their emotions, which should not be the case according to the idea that emotions are discrete and isolated from each other (Saarni, 1999). Instead, emotions seem to overlap, just like the color spectrum, without discrete boundaries between them (Russell & Fehr, 1994). As an example, people typically describe as feeling not one but multiple positive emotions at the same time (D. Watson & Clark, 1992). This issue has led to the proposal of dimensional models of emotion, which view emotional experiences as "a continuum of highly interrelated and often ambiguous states" (Posner et al., 2005). In these, a multidimensional space is used, mapping different emotional states to locations in that space.

The most notable dimensional approaches are two-dimensional, a number that is supported by extensive research of the intercorrelations between emotional experiences (Larsen & Diener, 1992). These two dimensions are found in a wide number of models, which have been conceptualized in different terms: the widely known arousal and valence (Russell, 1980), but also tension and energy (Thayer, 1989), approach and withdrawal (Lang, Bradley, & Cuthbert, 1998), or dimensions of positive and negative affect (D. Watson, Wiese, Vaidya, & Tellegen, 1999). Supporters of this idea suggest that emotional states arise from the combination of two distinct neurophysiological systems: one for arousal and other for valence (Russell, 2003).

Taking fear as an example, it arises from a combination of high arousal and negative valence stimuluses to the central nervous system. These patterns of neurophysiological activity are then interpreted by our cognitive system resulting in our personal, subjective experience of fear (Russell, 2003). Emotions are thus "the end product of a complex interaction between cognitions, likely occurring primarily in neocortical structures, and neurophysiological changes related to the valence and arousal systems, which presumably are subserved largely by subcortical structures" (Posner et al., 2005).

Russell's Circumplex Model of Emotion

In contrast to the idea of independent neural systems to each basic emotion, Russell (1980) proposed that each emotional state sprouts from two independent neurophysiologic systems. In his study, the two proposed dimensions are valence (pleasure-displeasure) and activity or arousal (aroused-not aroused). The result, illustrated in Figure 2.5,

is a two-dimensional plane forming four different quadrants, which can be roughly defined as: 1) exuberance, referring to happy and energetic emotions (Q1); 2) anxiety, representing frantic and energetic ones (Q2); 3) depression, referring to melancholic and sad emotions (Q3); and 4) contentment, representing calm and positive emotions (Q4). An important characteristic of this model is that emotions are placed far away from the center. Otherwise, cases where both arousal and valence have neutral values do not represent clear, identifiable emotions.

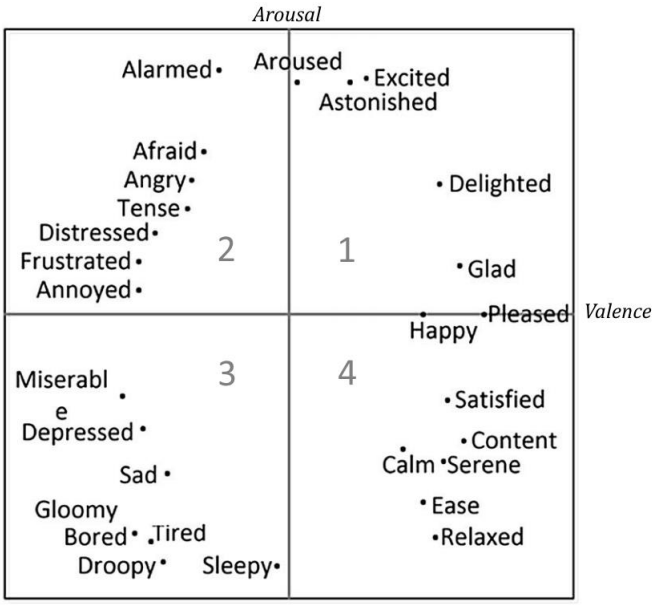


Figure 2.5: Russell's circumplex model of emotion.

The circumplex model has received wide support by several studies (e.g., (Barrett & Russell, 1999; Russell, 1983), reviewed in detail in (Posner et al., 2005)), and been adopted in MER research as the standard dimensional model of emotion.

As stated, some authors proposed alternative two-dimensional models using different labels in each axis. Two of the most recognized models rotate the Russell's circumplex model by 45 degrees: the Positive and Negative Affect Schedule (PANAS) (D. Watson, Clark, & Tellegen, 1988; D. Watson & Tellegen, 1985) and the Thayer's model of emotion (Thayer, 1989), as shown in Figure 2.6. The first uses Positive Affective (PA) dimension, representing high arousal and positive valence, and Negative Affective (NA) dimension. The second uses energetic arousal (EA) and tense arousal (TA), where Thayer suggests that "emotions are represented by components of two biological arousal systems, one which people find energizing, and the other which people describe as producing tension" (energetic arousal – readiness for vigorous action versus tense arousal –

preparatory-emergency system) (Thayer, 1989).

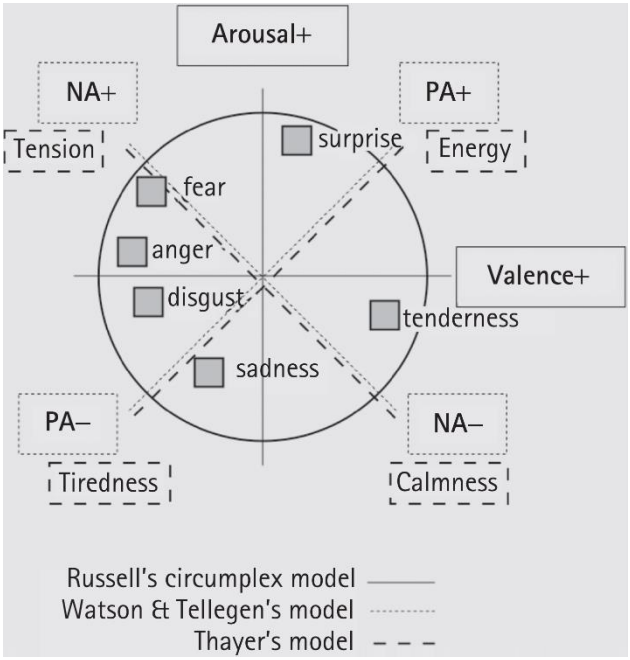


Figure 2.6: Russell's, Watson et al. and Thayer's two-dimensional models of emotions (from (Zentner & Eerola, 2010, p. 199)).

Regarding the two-dimensional approach, researchers have noted that it fails to account for all variance in music-mediated emotions (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; G. L. Collier, 2007; Ilie & Thompson, 2006), by placing distinct emotions close in the 2D space (e.g., anger and fear are both high in arousal and negative in valence) (Scherer, Johnstone, & Klasmeyer, 2003). Several authors proposed a third dimension to help reduce this problem, namely: potency, intensity, dominance or interest (Gabrielsson & Juslin, 2003).

Schimmack & Grob model of emotion

One of the most well regarded three-dimensional models of emotion was proposed by Schimmack and Grob (Schimmack & Grob, 2000). There, the authors combined the Russell's and Thayer's models, obtaining valence, tense arousal and energetic arousal dimensions (illustrated in Figure 2.7). The reasoning behind this proposal is the fact that the two arousal (or activation) dimensions are actually controlled by independent physiological systems (D. Watson et al., 1999), which can be independently stimulated, even in opposite directions (Gold, MacLeod, Thomson, Frier, & Deary, 1995). Although

this model seems better suited to reduce the two-dimensional variance problem (Schimmack & Reisenzein, 2002), it is yet to be adopted in MER studies.

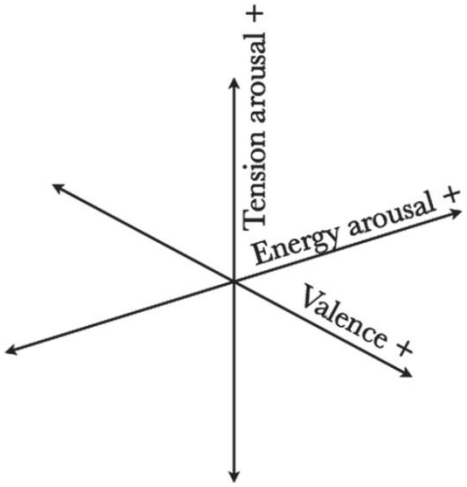


Figure 2.7: Schimmack & Grob three-dimensional model of emotion (from (Eerola & Vuoskoski, 2011)).

Tellegen-Watson-Clark's Model

An additional taxonomy using a third variable is the Tellegen-Watson-Clark model of emotion (1999), depicted in

Figure 2.8. This model extends the previous dimensional models, emphasizing the value of a hierarchical perspective by integrating existing models of emotional expressivity.

In it, a three-level hierarchy incorporates at the highest level a happiness versus unhappiness dimension, an independent positive affect (PA) versus negative affect (NA) dimension at the second order level below it, and discrete expressivity factors of joy, sadness, hostility, guilt/shame, fear emotions at the base.

The key to this hierarchical structure is the recognition that the general bipolar factor of happiness and independent dimensions of PA and NA are better viewed as different levels of abstraction within a hierarchical model, rather than as competing models at the same level of abstraction. At the highest level of this model, the general bipolar factor of happiness accounts for the tendency for PA and NA to be moderately negatively correlated. Therefore, the hierarchical model of affect accounted for both the bipolarity of pleasantness-unpleasantness and the independence of PA and NA, solving a debate that occupied the literature for decades (Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2011).

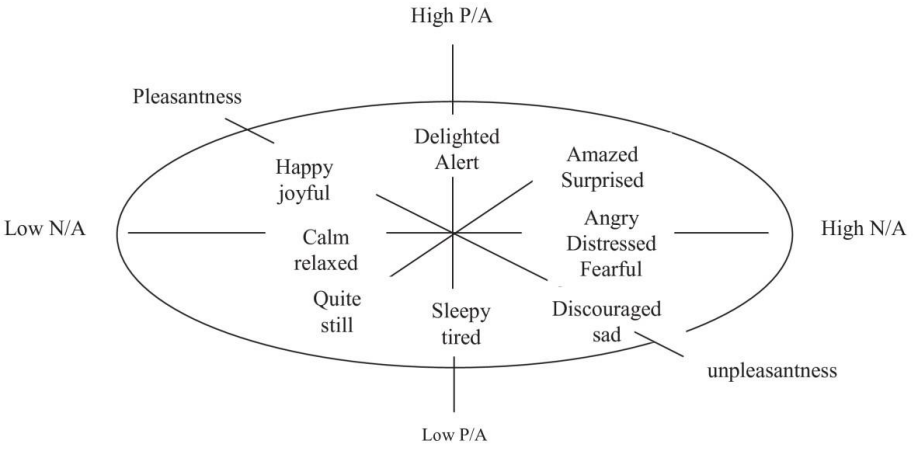


Figure 2.8: Tellegen-Watson-Clark model of emotion (Trohidis et al., 2011).

Although three-dimensional models reduce the lack of differentiation in emotions that are close neighbors in the valence-activation space, a problem often criticized in two-dimensional models, one of its major disadvantages is the high complexity, making it unpractical for non-academic use.

Concluding, the benefit of dimensional models is the reduced ambiguity when compared with the categorical approach, since emotions are positioned in different parts of an N-dimensional emotional plane. The dimensional model is sometimes further divided into discrete and continuous.

Continuous models view the emotion plane as a continuous space where each point denotes a different emotional state. As a result, the ambiguity related with emotion states is removed. Given the higher complexity introduced by continuous models to MER applications, both in terms of machine learning techniques and annotation process, some authors opt to use a discrete view of the dimensional models, considering parts of the plane as representing specific emotions. Taking the Russell's circumplex model, a simple example is viewing it as four distinct emotions, defined by each of its four quadrants: 1) happiness or enthusiasm; 2) anxiety or anger; 3) depression or sadness; 4) contentment or relaxed.

Moreover, some authors have studied both dimensional and categorical models of emotions simultaneously, in order to "clarify their mutual relationship and applicability to music and emotions" (Eerola, Lartillot, & Toiviainen, 2009). There, 360 audio excerpts were rated using both approaches: a categorical model with five of the basic emotions: happiness, sadness, tenderness, anger and fear); as well as a three-dimensional

model with valence, activity and tension. The results show that at least two of the emotion dimensions correlated highly: tension and valence (highly correlated) and activity and tension (moderately correlated). On the other hand, valence and activity (the two dimensions of Russell's circumplex model did not present such a relation. Considering this, the authors applied ridge regression, a variation of linear regression better at analyzing multiple regression data that suffer from multicollinearity, to predict the dimensional ratings from the categorical ratings and vice versa. The results presented by Eerola et al. (2009) suggest that the basic emotion model "can more accurately explain the results obtained with the three-dimensional model than contrariwise". Moreover, it was also verified that the two-dimensional models can explain the results obtained with the basic emotion model "virtually as accurately as the three-dimensional model, with the exception of anger and tenderness" (which still only had minor differences).

A brief summary of the previously discussed emotion models is presented in Table 2.3.

Emotion Model	*Type*	*Summary*
Ekman's Basic Emotions	Categorical	6 words, considered the basic emotions
Hevner's adjective clock	Categorical	8 clusters, 67 adjectives
MIREX Mood Taxonomy	Categorical	5 clusters, 29 mood tags
Geneva Emotional Music Scales	Categorical	45 terms, 9 categories, 3 superfactors
Russell's circumplex model of emotion	Dimensional	2D using arousal and valence, ∞ emotional states (continuous view)
Schimmack & Grob model	Dimensional	3D using valence, energetic arousal and tense arousal, ∞ emotional states
Tellegen-Watson-Clark model of emotion	Dimensional	3-level hierarchy with happiness vs affect vs discrete factors

Table 2.3: Comparison of the reviewed emotion models.

2.2.3. Selecting an Emotion Taxonomy for MER

Given the lack of consensus regarding emotion taxonomies, with several models available, each following distinct principles, an obvious question arises: which emotion taxonomy should be used for MER?

There is no simple answer to this question, as demonstrated by the numerous studies in MER using each of the approaches, from categorical, domain-specific models, to dimensional ones. Some studies have attempted to answer this question by asking listeners to rate emotions induced by music clips using either a domain-specific model of emotion, the basic emotions or a dimensional model (Zentner et al., 2008).

Generally, the results of the study demonstrated a clear preference from listeners to use the emotion terms of the selected domain-specific model (Geneva Emotional Music Scales, known as GEMS) to describe the emotions felt. Additionally, the ratings using the domain-specific model increased agreement across listeners and provided better discrimination between the musical excerpts (Zentner et al., 2008). Still, while the domain-specific model performed better for induced emotions, the results were not so convincing for perceived emotions. Inversely, the basic emotions model achieved better results for ratings for perceived emotion than for induced emotion. In brief, while the study suggests domain-specific models as better suited to rate induced emotions, "the kind of model that provides the best fit for perceived musical emotions remains unclear" (Zentner & Eerola, 2010).

Selecting an emotion taxonomy for a novel MER study is thus a complex process, which should consider the current state-of-the-art in the field and the study objectives. While dimensional models reduce or eliminate the ambiguity and are better aligned with our biology, we humans prefer to use discrete labels reasoned by our cognitive system when talking about emotions. Moreover, the current results from MER studies using simpler categorical models (e.g., MIREX AMC task results) are still average and dimensional models typically pose an even harder computational problem and are less understood by listeners since concepts such as valence or arousal are not intuitive for the average user.

2.3. Musical Dimensions

To better understand how music and emotion relate, we first need to have a deeper understanding of the fundamental musical dimensions and their organization, as described in this section.

Musical dimensions are usually organized into four to eight different categories (depending on the author, e.g., (Meyer, 1973; Owen, 2000)), each representing a core concept. In this section, based on the cited works, we use an eight category organization based on the literature to briefly describe the main musical features: melody, harmony, rhythm, dynamics, tone color (or timbre), expressive techniques, musical texture and musical form.

The organization of these dimensions is not strict. Many of the musical features are

somehow interconnected and may interact and touch other dimensions. Thus, it can be argued that some of them could be placed in different musical categories. In any case, through this organization, we are able to better understand: i) where features related to emotion belong; ii) for these musical features, which of them can be extracted from audio signals with the existing algorithms; iii) and thus, which categories may lack computational models to extract musical features relevant to emotion.

It is important to note that this section does not offer an exhaustive review on music theory. Its aim is to catalog the main elements and characteristics of music in order to better understand which ones might be relevant to emotion recognition and not yet explored by existent computational algorithms and audio features.

2.3.1. Melody

Melody can be defined as a horizontal succession of pitches (perceptual correlate of fundamental frequency) or musical tones, perceived by listeners as a single musical line. Johann Philipp Kirnberger, a student of Bach, defined melody as "the true goal of music [...]. All the parts of harmony have as their ultimate purpose only beautiful melody" (Forte, 1979, p. 203).

Name	*Description*
Melodic arrangement	How melodies are placed in the piece, e.g., in sequence or counter-melodies.
Melodic movement and contour	Pitch directions in a melody (patterns of notes) and shapes or contours formed by them.
Pitch	The sound perception, which can be "higher" or "lower", related to frequency.
Pitch Range	The pitch distance from highest to lowest (or notes extent) used in a melody, which can be narrow or wide.
Register	The "height" of a sound, generally classified in high, middle or low.
Melodic features	Features added to the melody in order to enrich or connect it to others (e.g., melodic riffs, repetitions).

Table 2.4: Summary of the melodic attributes of music.

Melodies can be classified according to their motion (intervals between pitches) as conjunct (smooth) or disjunct (disjointedly ragged or jumpy). Their main components

are pitch (which can be definite or indefinite), pitch range, register, melodic contour or shape, melodic movement and melodic arrangement.

A summary of the attributes related to melody are presented in Table 2.4. We offer a more detailed, yet concise, description in Appendix A. For a thorough analysis, we recommend (Benward & Saker, 2008; Laitz, 2007).

2.3.2. Harmony

If melody is said to be the horizontal part of music, harmony refers to its "vertical" aspect. That is, the sound produced by the combination of various pitches (notes or tones) in chords. The word "harmony" originates from the Greek language, meaning "agreement, concord of sounds" or "combination of tones pleasing to the ear"[24].

Analyzing the harmony of a song involves the study of chords, made of several notes played simultaneously, and of chord progressions, which are the sequences of chords arranged together (illustrated in Figure 2.9).

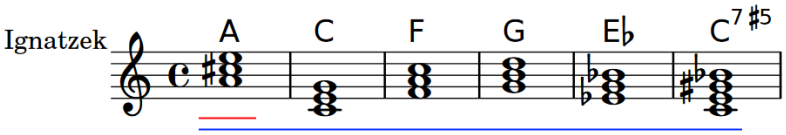


Figure 2.9: A 3-note chord in red and a chord progression of 6 chords in blue. Chord names are displayed at the top using the Jazz notation system proposed by Klaus Ignatzek.

In Western music, the key of a musical piece is the scale (i.e., group of pitches, featuring a tonic note and corresponding chords, which provides a sense of rest) that forms the basis of the composition. The key may be in the major or minor mode. Music lacking a tonal center or key is said to be atonal (as opposed to tonal). In addition to the traditional major and minor scales of tonal music (i.e., seven-note diatonic scales), more unusual ones exist such as the old church modes (e.g., Gregorian mode) and the chromatic scale, a twelve-note scale featuring all semitones.

A summary of the harmonic characteristics is presented in Table 2.5. We offer a more detailed, yet concise, description in Appendix A. For a thorough analysis, we recommend (Benward & Saker, 2008; Laitz, 2007).

[24] https://www.etymonline.com/word/harmony

Name	*Description*
Harmonic rhythm or harmonic tempo	Rate at which the chords change in relation to the rate of notes.
Harmonic progression	Succession of musical chords, or chord changes, that helps to indicate where the melody should go.
Modulation	The process of changing the key center (tonal or tonic center) in a musical piece.
Harmonic perception	Relative harshness of a sound. Consonant sounds are pleasing, made of smooth-sounding harmonic combinations. Dissonant are the opposite.

Table 2.5: Summary of the harmonic characteristics of music.

2.3.3. Rhythm

Rhythm represents the element of "time" in music, the patterns of long and short sounds and silences found in music. In its most general sense, rhythm (originating from the Greek word *rhythmos*, derived from *rhein*, "to flow") is an ordered alternation of contrasting elements[25].

Several important aspects are part of rhythm, namely, tempo, duration or meter. Some of the most relevant are briefly described in Table 2.6. We offer a more detailed, yet concise, description in Appendix A. For a thorough analysis, we recommend (Benward & Saker, 2008; Laitz, 2007).

Name	*Description*
Rhythm types	Can be simple or complex, regular or irregular.
Note values and rests	Indicates a note length or duration. Some examples: semibreve, quaver, semiquaver.
Rhythmic devices	Rhythmic devices give a piece of music its shape and often indicate its genre (e.g., riff, repetition, syncopation are common in rock).
Rhythmic layers	Grouping of performing media (instruments) in a musical piece (e.g., instrumental groups and vocals).
Duration	How long a sound (or silence) lasts.

[25] https://www.britannica.com/art/rhythm-music

Beat	The underlying, regular pulse in a piece of music. Can be strong/definite or weak/indefinite.
Metre	The grouping of beats in a piece of music that we hear as an organized succession of rhythmic pulses.
Tempo	The speed of the beat, such as fast or slow or becoming faster or slower.

Table 2.6: Summary of the rhythmic attributes.

2.3.4. Dynamics

Dynamics represents the variation in loudness or softness of notes in a musical piece. All musical aspects relating to the relative loudness (or quietness) of music fall under the general element of dynamics. Important aspects include the relative softness and loudness of sound, change of loudness (contrast), and the emphasis on individual sounds (accent).

The dynamics markings in a musical score are always relative, e.g., indicating that a specific passage should be played louder, but not defining an exact level of loudness. Changes in dynamics are used by musicians to create interest and communicate with the audience.

Several important features are part of dynamics, namely, dynamic levels, accents and dynamic changes. A summary of the dynamics attributes is presented in Table 2.7. We offer a more detailed, yet concise, description in Appendix A. For a thorough analysis, we recommend (Benward & Saker, 2008; Laitz, 2007).

Name	*Description*
Dynamic levels	The loudness levels in a musical piece (e.g., forte, piano).
Accents and changes in dynamic levels	Gradual changes in dynamics (e.g., crescendo for gradually getting louder). Accents are an emphasis for specific notes and sounds (e.g., sforzando meaning playing a note with sudden emphasis).

Table 2.7: Summary of the attributes related with dynamics.

It can be argued that elements such as accents can be classified as articulation mechanisms and thus should be placed in the expression techniques dimension (see Section 2.3.6). As stressed before, many of the musical attributes touch several musical dimensions and thus this organization reflects our view.

2.3.5. Tone Color or Timbre

Tone color, also known as timbre, refers to the perceived sound quality (properties) of a sound (e.g., a musical note). It is the tone color of a sound that allows the listener to distinguish between different sources, such as two different instruments playing similar notes, differentiate human voices or even distinguish instruments of the same family, such as a trumpet from a saxophone.

Name	*Description*
Instrument materials	Material and shape of an instrument influences its sound (e.g., wood, metal, vocal).
Playing methods	Method used to produce a sound from the instrument (e.g., pluck, hit, blow).
Instruments' and voices' types	Classification of the source producing the sound (e.g., strings or percussion for western instruments; or soprano or tenor for voices).
Combinations and types of sounds	Acoustic (non-electric) or electronic instruments, combined in different musical groups (e.g., bands, orchestras, Jazz trio or choirs).

Table 2.8: Summary of the elements influencing tone color.

In an analogous way to the color used by an artist, sound can be said to have a spectrum of tone colors. Each instrument has a distinct tone color in this spectrum, which the composer uses and combines, creating contrasts and new colors (combination of instruments) to enhance his musical piece, just like an artist painting a scene.

Some authors have tried to decompose tone color into distinct components (e.g., (Erickson, 1975)) but results are not consensual, since even tone color itself is still not fully agreed upon, with some researchers classifying it as "the psychoacoustician's multi-dimensional waste-basket category for everything that cannot be labeled pitch or loudness." (McAdams & Bregman, 1979). Nonetheless, two of its main components are harmonics and the sound envelope, which are explained in Appendix A.

A summary of the attributes contributing to tone color are presented in Table 2.8.

2.3.6. Expressive Techniques

Expressive techniques refer to the way a performer plays a musical piece, specifically the

techniques used by him/her to create the musical detail that articulates a style or interpretation of a style. As stated in Section 2.3.4, expressive techniques are combined with dynamics to give "soul" to a piece of music.

Name	*Description*
Tempo (changes)	Tempo and its changes can also affect the expressive quality of music. Can get faster or slower, gradually or immediately.
Stylistic indications	Terms to indicate the style in which a piece is to be performed, such as legato (smoothly, connected notes) or rubato (with freedom).
Articulation	The way in which specific parts or notes in a piece of music are played (e.g., staccato for short detached notes or slur for two notes played without separation).
Ornamentation	Decoration of notes with special features (such as glissando or trills) to add interest and expressive qualities.
Instrumental, vocal and electronic techniques	Techniques to produce different sounds to express a specific style (e.g., vibrato, tremolo), or electronic enhancements (e.g., vocoders).

Table 2.9: List of expressive techniques attributes described.

Over the centuries, several expressive techniques have been created. These can be techniques related with instruments or vocals, ornamentations, changes in tempo or specific techniques articulating consecutive notes together.

These expressive techniques, together with all the other musical elements, contribute greatly to define musical styles. From western classical music, to one of the mainstream international genres (e.g., rock, pop, rap or metal) or world music such as Indian ragas, African zouk or Portuguese *fado*, specific styles have specific, typical expressive techniques and distinct instruments.

Several important features are part of expressive techniques, namely, tempo changes, stylistic indications, articulation, ornamentation, as well as instrumental, vocal and electronic techniques. A summary of the expressive techniques attributes is presented in Table 2.9. We offer a more detailed, yet concise, description in Appendix A. For a thorough analysis, we recommend (Benward & Saker, 2008; Laitz, 2007).

2.3.7. Musical Texture

Musical texture refers to the way the rhythmic, melodic and harmonic information produced by musical instruments and voices is combined in a musical composition. It is thus related to the combination and relations between the musical lines or layers (one or more instruments with the same role) or a song.

Texture can be described based on its density, from thin to thick, and range, from narrow to wide. As an example, a song played by a solo guitar will have a single layer and thus a thin texture, while a musical piece for orchestra, with several melodic, harmonic and rhythmic lines, will have a thick texture. A single musical line or layer can have several performers following the same melody. The range of a texture is rated based on the distance between the lowest and highest tones (Benward & Saker, 2008, p. 146).

Name	*Description*
Number of layers, density and range	Number of musical lines (e.g., single melodic line, melodic with accompaniment, multiple melodic, non-melodic), their density (thin or think) and range (narrow to wide).
Texture Types	Different combinations of layers such as monophonic (single layer); homophonic (two or more layers, with one prominent melody); polyphonic (two or more independent melodies).

Table 2.10: Summary of the musical texture attributes.

Additionally, the musical texture can also be classified according to its type, based on the number of existent layers and their relations. Some common types are monophonic, homophonic and polyphonic. As with other musical dimensions, texture and other elements sometimes overlap. For instance, if the number of layers increases (thicker texture), usually a corresponding increase in dynamics is expected.

A summary of the attributes of musical texture are presented in Table 2.10. We offer a more detailed, yet concise, description in Appendix A. For a thorough analysis, we recommend (Benward & Saker, 2008; Laitz, 2007).

2.3.8. Musical Form

Musical form or musical structure refers to the overall structure of a musical piece, and describes the layout of a composition as divided into sections (Brandt, 2011). These sections are usually identified by changes in rhythm and texture. If the rhythm and texture remain constant, the listener tends to perceive the excerpt as a single section. On

the other hand, a marked change in rhythm or texture is normally perceived as a point of contrast – a boundary, from which the piece passes into a new section.

The organization of a musical piece can be scrutinized at various levels. At a lower level, the passages can be combined in different basic musical forms such as the strophic (repeating sections), binary and ternary forms, but also thirty-two-bar form, verse-chorus form, and others.

Name	*Description*
Song elements	Different sections composing the musical piece (e.g., introduction, verse, chorus or bridge).
Organization levels	Music form can be roughly divided into three levels designated as passage (lowest, related to musical phrases and paragraphs), piece (related to the entire piece) and cycle (large compositions).
Basic musical forms	The combination of different sections can be organized into several forms. Some examples are: through-composed (no repeated sections), strophic form (the opposite, also called verse-repeating) or binary form (repetition of two contrasting sections).

Table 2.11: Summary of features contributing to musical form.

Several distinct elements are present in a song, each with a different function and position. The most common two are the *verse*, normally containing different sets of lyrics, and *chorus*, which usually repeats the same melodic and lyrical verses. Other common sections exist such as *intros*, in the beginning of a song, *bridges*, connecting verses and chorus or *outros*, at the end of a song. In specific genres such as pop/rock or blues it is not uncommon to also have a *solo* section, where a melodic line is played (sometimes even improvised). These song elements are further described in the Appendix A.

A summary of the musical form attributes is presented in Table 2.11.

2.4. Relations between Music and Emotions

As explained earlier, music has been with us since our prehistoric times[26], serving as a language to express our emotions. This is regarded as music's primary purpose (Cooke, 1959) and the "ultimate reason why humans engage with it" (Pannese et al., 2016).

[26] Music is said to exist for at least 50,000 years, invented in Africa before our species left the continent for the first time, evolving to become a fundamental constituent of human life (Wallin, Merker, & Brown, 1999).

In our path to advance the music emotion recognition field, we reviewed the history and definitions behind emotion and explored how emotions can be classified. Next, we reviewed the music theory literature, in order to better understand what is music and its key elements. In this section we bridge the previous sections, by discussing the theoretical knowledge that connects music dimensions and emotion.

The relations between music and emotions have been debated since millennia ago, with associations between modes and emotions found in ancient texts, from Indian, Middle Eastern (e.g., Persian), and far eastern (e.g., Japanese) traditions (Pannese et al., 2016). Natya Shastra (Nātya Śāstra), an ancient Sanskrit Hindu text describing performance arts, estimated to have been written somewhere between 500 B.C. and 500 A.D. (Dace, 1963, p. 249) suggests elements such as modes and musical forms as able to express particular emotions.

A similar view of modes as suited to represent particular emotions has been suggested in ancient Greece by Plato (Plato (375 B.C.), 1969, fol. 3.398b-3.398e)[27], which also advocated that "good rhythm wait upon good disposition, [...] the truly good and fair disposition of the character and the mind" (Plato (375 B.C.), 1969, fol. 3.400e)[28]. In addition, Plato considers harmony as capable of moving the listener, arguing that both "rhythm and harmony find their way to the inmost soul and take strongest hold upon it" (Plato (375 B.C.), 1969, fol. 3.401d)[29]. Aristotle, arguably Plato's most famous student, supported the same ideas, stating that "rhythms and melodies contain representations of anger and mildness, and also of courage and temperance" (Aristotle (IV c B.C.), 1944, fol. 8.1340a)[30], while different harmonies could range from relaxing to "violently exciting and emotional" (Aristotle (IV c B.C.), 1944, fol. 8.1342a-8.1342b)[31, 32].

Scientific studies focusing on the relations between music and emotions started more than a century ago. One of these early examples is a study by Kate Hevner, where the author evaluates the influence of several musical factors such as rhythm, pitch, harmony, melody, tempo and mode to each of the eight emotion clusters earlier proposed by her (Hevner, 1937).

Up to this day, this research problem is still far from completely solved. Still, several contemporary research works had already identified possible correlations or in some cases causal associations between specific musical elements and emotions. One of the most widely accepted is mode: major modes are frequently related to emotional states

[27] http://data.perseus.org/citations/urn:cts:greekLit:tlg0059.tlg030.perseus-eng1:3.398

[28] http://data.perseus.org/citations/urn:cts:greekLit:tlg0059.tlg030.perseus-eng1:3.400e

[29] http://data.perseus.org/citations/urn:cts:greekLit:tlg0059.tlg030.perseus-eng1:3.401d

[30] http://data.perseus.org/citations/urn:cts:greekLit:tlg0086.tlg035.perseus-eng1:8.1340a

[31] http://data.perseus.org/citations/urn:cts:greekLit:tlg0086.tlg035.perseus-eng1:8.1342a

[32] http://data.perseus.org/citations/urn:cts:greekLit:tlg0086.tlg035.perseus-eng1:8.1342b

such as happiness or solemnity, whereas minor modes are often associated with sadness or anger (Gabrielsson & Lindström, 2011); simple, consonant, harmonies are usually happy, pleasant or relaxed. On the contrary, complex, dissonant, harmonies relate to emotions such as excitement, tension or sadness, as they create instability in a musical piece (Laurier et al., 2009). Many other musical elements have been related to emotion, namely: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality, rhythm, mode, loudness, vibrato or musical form (Friberg, 2008; Laurier, 2011; Laurier et al., 2009; Meyers, 2007).

Over the decades, several associations have been identified, linking specific emotional responses to the musical elements described in Section 2.3. The next sections detail the most relevant findings in this area. For some musical elements, the research can be quite contradicting in its findings, which can be caused by many factors – from different research methodologies[33] to differences in the scope of the studies (e.g., induced or perceived emotion, significant differences in methodologies, population, and others). This is also caused by the complexity associated with music emotion research and indicates that further research is needed.

Most of the associations described below pertain to music emotion perception or transmission, since most studies tackled that problem (e.g., by asking listeners which emotions they identified). Still, some studies do not clearly state whether their findings concern perceived or induced emotion.

2.4.1. Melody and Emotion

Given its central role in a musical piece, being (one of) the most memorable elements in a song, associations between melodic cues and emotions are expected and suggested since the Plato days, as discussed earlier.

Some of the strongest associations found are the wider melodic ranges (pitch ranges) and energetic emotions such as joy (Balkwill & Thompson, 1999) or fear (Krumhansl, 1997), while narrow ranges are associated with lower arousal emotions such as sadness, melancholy or tranquility (Gundlach, 1935). Other melodic elements such as ascending versus descending melodic contours have been studied and related to several emotions (Gerardi & Gerken, 1995; Hevner, 1936). However, some of these are disputed in other studies, arguing that the relation is more complex and involves interactions with other elements such as rhythm and modes (Lindström, 2006). These findings have been ob-

[33] Researchers have used several distinct methods over time. As an example, Grabrielsson et al. divides the existent studies in different groups: "early studies using choice among descriptive terms; based on multivariate analyses; and later experimental studies" (Gabrielsson & Lindström, 2011, p. 387).

served in cross-cultural studies, where listeners have also associated joy with simpler melodies, and sadness with more complex ones (Balkwill & Thompson, 1999), even when exposed to unfamiliar tonal systems.

The links found between melody and emotions are summarized in Table 2.12[34].

Musical Element	*Value*	*Associated emotions*
Melodic intervals	Large	Powerful (Maher & Berlyne, 1982)
	Minor 2nd	Melancholic (Maher & Berlyne, 1982)
	Perfect 4th, major 6th, minor 7th	Carefree (Maher & Berlyne, 1982)
	Perfect 5th	Carefree (Maher & Berlyne, 1982), active (Smith & Williams, 1999)
	Octave	Carefree (Maher & Berlyne, 1982), positive and strong (Smith & Williams, 1999)
Melodic direction and contour	Ascending	Happy (W. G. Collier & Hubbard, 2001; Gerardi & Gerken, 1995), fearful, surprised, angry (Scherer & Oshinsky, 1977), tense (Krumhansl, 1996)
	Descending	Sad (Gerardi & Gerken, 1995; Scherer & Oshinsky, 1977), bored, pleasant (Scherer & Oshinsky, 1977)
Melodic movement	Stepwise motion	Dull melodies (Thompson & Robitaille, 1992)
	Intervallic leaps or skips	Exciting melodies (Thompson & Robitaille, 1992)
	Stepwise and skipwise leaps	Peaceful melodies (Thompson & Robitaille, 1992)
Pitch	High	Surprised, angry, fearful and others (Scherer & Oshinsky, 1977), happy

[34] In this and subsequent tables, the emotions described with nouns by the original authors were converted to adjectives to maintain consistence between the various works.

		(W. G. Collier & Hubbard, 2001), increased tense arousal (Ilie & Thompson, 2006)
	Low	Sad (Scherer & Oshinsky, 1977; Wedin, 1972), bored, pleasant (Scherer & Oshinsky, 1977), increased valence (Ilie & Thompson, 2006)
Pitch variation	Large	Active, happy, pleasant, surprised (Scherer & Oshinsky, 1977)
	Small	Angry, bored, disgusted, fearful (Scherer & Oshinsky, 1977)
Pitch range	Wide	Joyful (Balkwill & Thompson, 1999), fearful (Krumhansl, 1997), scary (Schimmack & Grob, 2000)
	Narrow	Sad (Balkwill & Thompson, 1999)

Table 2.12: Summary of the emotions associated with melodic elements.

2.4.2. Harmony and Emotion

Harmony, together with rhythm and melody, has been suggested as able to elicit emotions since the ancient times. Consonant harmonies are usually associated with happiness, tranquility, serenity, while dissonant complex harmonies are related with negative emotional states, as with tension and sadness, due to the instability they create in the piece (Laurier et al., 2009).

In addition, major modes have been frequently related with positive emotions (e.g., happiness), while minor modes are linked to negative ones (e.g., sadness) (Gabrielsson & Lindström, 2011). Some authors such as Cook et al. have tried to further understand this affective response to major/minor chords and resolved/unresolved chords, concluding that this emotional association is "neither due to the summation of interval effects nor simply arbitrary, learned cultural artifacts, but rather that harmony has a psychophysical basis dependent on three-tone combinations" (Cook & Fujisawa, 2006).

The relations between emotions and harmony are summarized in Table 2.13.

Musical Element	*Value*	*Associated emotions*
Harmonic perception (harmonic intervals)	Consonant (simple)	Normally associated with positive emotions, such as: happy, serene and dignified (Hevner, 1936), pleasant (Costa, Bitti, & Bonfiglioli, 2000; Wedin, 1972), tender (Lindström, 2006)
	Dissonant (complex)	Associated mostly with negative emotions: vigorous, sad (Hevner, 1936), unpleasant (Costa et al., 2000; Wedin, 1972), tense (Krumhansl, 1996), fearful (Krumhansl, 1997), angry (Lindström, 2006)
	High-pitched	Happy, more active / powerful (Costa et al., 2000; Maher, 1980)
	Low-pitched	Sad, less powerful (Costa et al., 2000; Maher, 1980)
Harmony (mode)	Major	Positive emotions, e.g., happy (Krumhansl, 1997; Lindström, 2006; Scherer & Oshinsky, 1977; Wedin, 1972), serene (Costa, Fine, & Bitti, 2004), tender (Lindström, 2006)
	Minor	Negative emotions, e.g., sad (Gagnon & Peretz, 2003; Krumhansl, 1997; Lindström, 2006; Wedin, 1972), disgusted and angry (Scherer & Oshinsky, 1977)
Harmony (tonality)	Tonal	Present in joyful, dull or peaceful melodies (Thompson & Robitaille, 1992), pleasant (Costa et al., 2004)
	Atonal	Present in angry melodies (Thompson & Robitaille, 1992)
	Using chromatic scales	Present in sad and angry melodies (Thompson & Robitaille, 1992)

Table 2.13: Summary of the relations between harmony and emotions.

2.4.3. Rhythm and Emotion

Rhythm, together with melody and harmony, is one of the dimensions most associated to the emotional expression in music. In fact, some authors consider it the most important one (e.g., (Gagnon & Peretz, 2003; Hevner, 1937; Juslin, 1997)). Rhythm elements, such as the augmentation of tempo (from 90 to 150 bpms), has been shown to increase happiness and surprise measures (i.e., induce) (Fernández-Sotos, Fernández-Caballero, & Latorre, 2016), while decreasing sadness, as illustrated in the Figure 2.10.

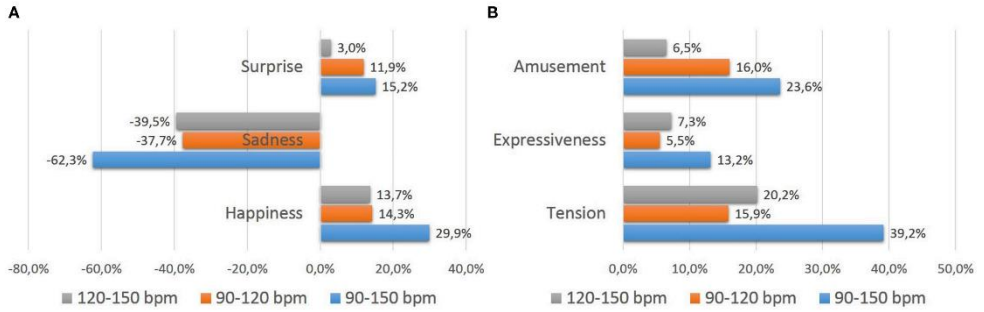


Figure 2.10: Influence of different tempo values in emotional measures. Changes in (A) basic emotions, (B) descriptive scales. From (Fernández-Sotos et al., 2016, p. 9).

In the study, the authors used two groups of words to study different emotion types: 3 "basic emotions" where users reported what they felt (i.e., induced emotion) in a scale of 1 to 8; and 4 "descriptive words" (tension, expressiveness, amusement and attractiveness) to classify (i.e., perceived emotion) the musical piece in a scale of 1 to 5.

In addition to tempo, the rhythmic unit of a piece have also been shown to influence the emotional message of a song. As an example, variations "of the rhythm of the melody without altering the musical line, harmonics or beat" (Fernández-Sotos et al., 2016), such as changes from whole and half notes (theme) to eighth or sixteenth, as well syncopated notes, were associated with specific emotions. A representation of these findings, showing how specific rhythmic elements influence emotions in the Russell's circumplex plane is provided in the Figure 2.11. Similar studies have supported the idea that rhythm is somehow influencing the emotional information in music (e.g., (Plewa & Kostek, 2012)).

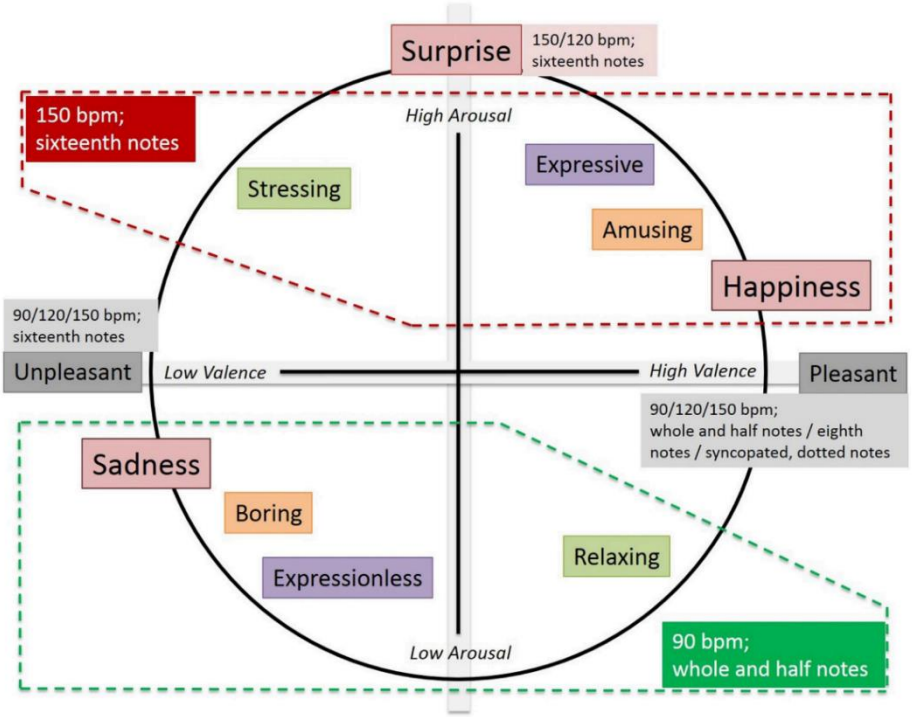


Figure 2.11: Influence of different rhythmic attributes to distinct emotional states in the Russell's circumplex model. From (Fernández-Sotos et al., 2016).

Table 2.14 summarizes the associations between rhythm and emotion and associated studies, based on the reviews present in (Gabrielsson & Lindström, 2011; Juslin & Laukka, 2004; Juslin & Timmers, 2011), as well as the other mentioned papers.

Musical Element	*Value*	*Associated emotions*
Rests	After tonal closure (a sequence which starts and ends in the same key)	Lower tension (Margulis, 2007)
	After no tonal closure	Higher tension than observed if after tonal closure (Margulis, 2007)
Rhythm Types	Regular/smooth	Happy, glad, serious, dignified, peaceful, majestic (Gundlach, 1935; K. B. Watson, 1942)
	Irregular/rough	Amusing, uneasy (Gundlach, 1935; K. B. Watson,

		1942)
	Complex	Angry (Lindström, 2006; Thompson & Robitaille, 1992)
	Varied	Joyful (Thompson & Robitaille, 1992)
	Firm	Dignified, vigorous, sad, exciting[35] (Hevner, 1936), sad (Wedin, 1972)
	Flowing/fluent	Happy, dreamy, graceful, serene (Hevner, 1936), gay (Wedin, 1972)
Tempo	Fast	Several, among which: happy, graceful, vigorous (Hevner, 1937), pleasant, happy (Rigg, 1940), pleasant (Wedin, 1972), active, angry, fearful, energy arousal and tension arousal (Ilie & Thompson, 2006), high arousal e.g., happy, stressful, amusing (Fernández-Sotos et al., 2016)
	Slow	Several, among which: serene, dreamy, dignified (Hevner, 1937), serious, sad (Rigg, 1940), tranquil, sentimental, dignified (Gundlach, 1935), sad (Balkwill & Thompson, 1999; Gagnon & Peretz, 2003; Hevner, 1937; Juslin, 1997; Scherer & Oshinsky, 1977; K. B. Watson, 1942; Wedin, 1972), peaceful (Balkwill & Thompson, 1999)
Tempo and Note Values	High tempo (150bpm) and sixteenth notes	High arousal: happy, amusing, expressive, stressful (Fernández-Sotos et al., 2016)
	Moderate to fast tempo (120 or 150bpm) and sixteenth notes	Surprised (Fernández-Sotos et al., 2016)
	Slow to moderate tempo (90bpms) and whole and half notes	Sad, boring, relaxing, expressionless (Fernández-Sotos et al., 2016)

Table 2.14: Elements of rhythm associated with emotion.

[35] Sometimes opposite emotions are associated to the same musical element, even in the same study, as found here. In this specific case, Hevner used 142 listeners to associate types of rhythm (firm or flowing) to 8 emotion clusters. Both "sad" and "exciting" clusters were related with firm rhythm, although the associated weight was lower than the remaining two clusters (dignified and vigorous).

2.4.4. Dynamics and Emotion

The influence of dynamics, namely loudness and loudness variations, in music emotions (both induced and perceived) have been studied by some researchers, some of which relate them with specific emotion states. Empirically, an association of loud music (high intensity) with powerful and intense emotions such as joy, anger or tension seems logical. In contrast, soft music is mostly linked to calm, serene or sad music. Such associations have been verified by several researchers (Gundlach, 1935; Ilie & Thompson, 2006; Juslin, 1997; K. B. Watson, 1942). Variations in loudness over a musical piece have also been studied. Namely, larger variations are usually more negative (K. B. Watson, 1942), while smaller variations are more positive (Scherer & Oshinsky, 1977).

Known associations between dynamics and emotion according to the reviewed studies are summarized in Table 2.15.

Musical Element	*Value*	*Associated emotions*
Dynamic levels	High/Loud	Excited (K. B. Watson, 1942), triumphant (Gundlach, 1935), strong/powerful (Kleinen, 1968), tense (Krumhansl, 1996), angry (Juslin, 1997), energy arousal and tension arousal (Ilie & Thompson, 2006)
	Low/Soft	Melancholic (Gundlach, 1935), peaceful (K. B. Watson, 1942), solemn (Wedin, 1972), fearful, tender, sad (Juslin, 1997), lower intensity, higher valence (Ilie & Thompson, 2006)
Accents and changes in dynamic levels	Large	Fearful (Scherer & Oshinsky, 1977)
	Small	Happy, pleasing, active (Scherer & Oshinsky, 1977)
	Rapid variations	Playful, pleading (K. B. Watson, 1942), fearful (Krumhansl, 1997)
	No changes	Sad, peaceful, dignified, happy (K. B. Watson, 1942)
	Crescendo, *decrescendo*, *accelerando*, *ritardando*	Said to be useful to describe perceptual and emotional processes (Langer, 1957, fig. 183)

Table 2.15: Elements of dynamics associated with emotion.

2.4.5. Tone Color and Emotion

Tone color or timbre is usually related to lower level elements and properties of the sound itself (such as amplitude and spectrum) essential to differentiate instruments and voices. Supported by this, it is sometimes incorrectly assumed that such musical dimension is too abstract and thus not relatable with a high-level concept as is emotion perception.

Several sound properties have been associated with emotional states. A rounder amplitude envelope is related with negative emotions such as disgust, sadness or fear (Juslin, 1997; Scherer & Oshinsky, 1977), while a sharper one gives rise to positive emotions such as happiness or surprise (Scherer & Oshinsky, 1977), with some authors also linking it to fear (Juslin, 1997). The number of harmonics has also been studied, where a lower number is associated with boredom, happiness or sadness (Scherer & Oshinsky, 1977), while a high number of harmonics is usually related with higher stress emotions such as anger, disgust, fear or surprise (Scherer & Oshinsky, 1977).

The tone color of specific instruments has also been suspected to carry emotional expression cues. In fact, composers and movie and marketing directors select specific instruments to express distinct emotions. This idea has been supported by studies such as (Eerola, Ferrer, & Alluri, 2012; B. Wu, Horner, & Lee, 2014b). In this respect, Hailstone et al. state that "timbre (instrument identity) independently affects the perception of emotions in music after controlling for other acoustic, cognitive, and performance factors" (Hailstone et al., 2009). These works highlight the importance of spectral centroid (brightness) as a "significant component in music emotion". Moreover, spectral centroid deviation, spectral shape, attack time and even/odd harmonic ratio were all considered relevant (B. Wu, Horner, et al., 2014b).

A summary of the relations found in the literature linking tone color and emotions is presented in Table 2.16.

Musical Element	*Value*	*Associated emotions*
Amplitude envelope	Round	Disgusted, bored, potent (Scherer & Oshinsky, 1977), fear, sadness (Juslin, 1997)
	Sharp	Pleasant, happy, surprised, active (Scherer & Oshinsky, 1977), angry (Juslin, 1997).
Spectral envelope (number of harmonics)	Low	Bored, happy, pleasant, sad (Scherer & Oshinsky, 1977)

	High	Active, angry, disgusted, fearful, potent, surprised (Scherer & Oshinsky, 1977)
Spectral characteristics such as spectral centroid, spectral centroid deviation and even/odd harmonics ratio	Positive correlation	Positive emotions: happy, heroic, comic, joyful (B. Wu, Horner, & Lee, 2014a; B. Wu, Horner, et al., 2014b)
	Negative correlation	Negative emotions: sad, scary, shy, depressed (B. Wu, Horner, et al., 2014a, 2014b)

Table 2.16: Summary of the relations between tone color and emotions.

2.4.6. Expressive Techniques and Emotion

Expressive techniques in music encompass several ornaments and features that are used by composers to enrich their pieces, as well as the performers, which try to express their emotions at that specific moment. Both parts have been studied and related with specific emotional states. As an example, staccato articulation is normally associated with higher intensity and energetic emotions (Wedin, 1972), mostly negative as with fear and anger (Juslin, 1997). On the other hand, legato is associated with softness (Wedin, 1972) and sadness (Juslin, 1997). Similar research has been conducted regarding vibratos and emotion expression, observing that "singing an emotional passage influences acoustic features of vibrato when compared with isolated, sustained vowels" (Dromey, Holmes, Hopkin, & Tanner, 2015). To assess this, classical singers were asked to sing passages of their preference[36] containing both high and low levels of emotion. The analysis of the recordings shows significant changes in vibrato characteristics such as frequency modulation rate and extent.

Regarding emotion expression by the performer, some studies highlighted that artists typically use different ornaments, such as accentuating specific notes considered happy, whereas not doing the same for sadness (Lindström, 1999). In addition, Timmers et al. (2007) studied the usage by flute and violin performers of specific ornamentations such as trills, turns, *mordente*, *arpeggio* and others, when they intended to express one of four specific affect terms (happiness, sadness, anger and love), and how these emotions were perceived by listeners. The accuracy between intended versus rated emotions was lowest for happiness. The performers employed more complex ornamentations for angry and the least complex for sadness.

[36] The authors did not set any emotion to be sang, instead singers were free to "identify a passage that they judged to be emotionally expressive"

The relations between emotions and expressive techniques are summarized in Table 2.17.

Musical Element	*Value*	*Associated emotions*
Articulation	Legato	Soft (Wedin, 1972), tender, sad (Juslin, 1997)
	Staccato	Intense, energetic, active (Wedin, 1972), fearful, angry (Juslin, 1997)
Ornamentation[37]	Single appoggiatura	[positive] Flute: lovely, sad [negative] Flute: happy, angry
	Double appoggiatura	[negative] Violin: sad
	Trill	[positive] Flute: angry [negative] Flute: lovely, sad
	Turn	[positive] Violin: happy
	Mordent	No significant correlation was observed.
	Slide	No significant correlation was observed.
	Arpeggio	[positive] Flute: angry; [negative] Flute: lovely, sad;
	Substitute	[positive] Violin: sad
Vibrato	Higher frequency modulation (FM) rate + higher FM extent + lower modulation variability	Observed when classical singers sang "more emotional passages"[38] (as opposed to neutral songs) (Dromey et al., 2015).
	Higher mean fundamental frequency + higher mean intensity	Observed in "more emotional passages" (Dromey et al., 2015).

Table 2.17: Elements of expressive techniques associated with emotion.

[37] From (Timmers & Ashley, 2007), showing only results based on listeners ratings, where significant correlations ($p<0.05$) were observed. The indicated associations can be either positive or negative correlated.

[38] As explained earlier, no specific emotions were selected, instead subjects were asked to sing "emotional passages" of their preference and the voice signals were analyzed.

2.4.7. Musical Texture and Emotion

Fewer studies have been conducted regarding musical texture and emotions and of these some contain contradicting results. In one of the oldest studies, the authors evaluated the emotional differences between monophonic (melody only) and homophonic textures (melody with block chords accompaniment) by children aged three to twelve. In that study, the unaccompanied version (monophonic) was rated as more positive (Kastner & Crowder, 1990). A similar result was observed by Webster et al., where non-harmonized melodies were considered happier (Webster & Weir, 2005). However, further studies trying to replicate Kastner et al.'s findings observed exactly the opposite result. There, not only children but also adult subjects considered monophonic sounds as less happy than accompanied ones (Gregory, Worrall, & Sarge, 1996; McCulloch, 1999). A possible explanation to this contradicting results are the different versions of "dense textures" used in each (Broze, Paul, Allen, & Guarna, 2014), where very basic/simple chords and a single instrument were used in the studies observing negative emotions, while the others used more complex (and thus, with higher density) accompaniments taken from published songbooks. These differences may influence greatly other musical dimensions (e.g., harmony) making it harder to correctly compare the results.

Polyphonic textures, containing several voices, have also been explored recently, suggesting that music with a higher number of voices is perceived as more positive. Such musical excerpts were rated as "sounding more happy, less sad, less lonely, and more proud" (Broze et al., 2014).

Although further studies are required to better understand exactly how musical texture influences emotion, the existent ones have demonstrated that it can indeed influence emotion in music either directly or by interacting with other features such as tempo and mode (Broze et al., 2014). Table 2.18 summarizes the associations found between musical texture and emotions.

Musical Element	*Value*	*Associated emotions*
Texture type	Monophonic	More positive (Kastner & Crowder, 1990) and happier (Webster & Weir, 2005) than homophonic
	Homophonic	Happier (Gregory et al., 1996; McCulloch, 1999) than monophonic.
Number of layers and density	Music with higher number of voices (polyphonic)	"more happy, less sad, less lonely, and more proud" (Broze et al., 2014)

Table 2.18: Summary of the relations between musical texture and emotions.

2.4.8. Musical Form and Emotion

Similarly to musical texture, few studies have investigated possible relations between musical form and emotion. From these, it seems that forms with lower complexity are associated with positive emotions (Imberty, 1979) such as relaxation, joy or peace (Balkwill & Thompson, 1999). On contrary, higher complexity forms usually result in more negative emotions such as sadness (Balkwill & Thompson, 1999), which can be higher in arousal (e.g., aggressive) or lower (e.g., melancholy) depending on the dynamism (high or low, respectively) (Imberty, 1979).

Some researchers have explored the relation between emotions and form by changing the order of sections (in classical music) but no relevant results were obtained (Konečni & Karno, 1994; Tillmann & Bigand, 1996). The few associations found between musical form and emotions are presented in Table 2.19.

Musical Element	*Value*	*Associated emotions*
Form complexity	Low	Positive emotions (Imberty, 1979), Joy, peace, relaxation (Balkwill & Thompson, 1999)
	High	Sadness (Balkwill & Thompson, 1999)
	High complexity and low dynamism	Depression, melancholy (Imberty, 1979)
	High complexity and high dynamism	Aggressiveness, anxiety (Imberty, 1979)

Table 2.19: Summary of the relations between musical form and emotions.

2.4.9. Interactions between Musical Dimensions

As described in the previous sections, each musical element may influence distinct emotional expressions. Based on this, it is clear that the emotional content in music is not defined exclusively by a single element but is built by the merging and interaction of several factors. Beyond studying associations concerning musical dimensions and emotions independently, these interactions between several musical dimensions and the associated emotional responses have also been studied and reviewed (e.g., (Gabrielsson & Juslin, 2003; Schubert, 1999)). Such works unveil interesting indirect relations and interactions regarding the variation of specific elements and the corresponding emotional changes, as well as possible interactions between elements, resulting in different emotional states. One example is the interaction between tempo and mode (Schubert, 1999),

illustrated in Figure 2.12 – high tempo and minor mode results in only high arousal, while the same high tempo, but with major mode, results in high arousal and positive valence.

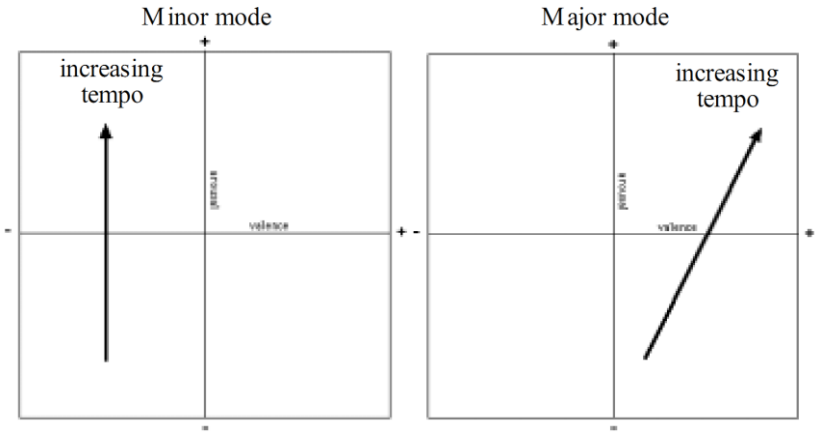


Figure 2.12: Hypothesized interaction between tempo and mode (Schubert, 1999, p. 390).

Several other authors have studied possible interactions, such as mode and tempo (Gagnon & Peretz, 2003; Rigg, 1940), the influence of pitch height, intensity and tempo in valence (Ilie & Thompson, 2006), the influence of rhythm, melodic contour and melodic progression in happy music (Lindström, 2006) or interactions between tempo, texture and mode (Webster & Weir, 2005).

Emotion conveyed by music has frequently been associated with both tempo and mode (major-minor). Gagnon et al. examined this and their possible interaction by presenting different versions of various melodies to volunteers and asking whether the melodies "sounded happy or sad" (Gagnon & Peretz, 2003). In the first experiment, the melodies were manipulated ensuring that either mode or tempo in isolation were altered. Next, both were manipulated simultaneously towards the same emotion (and the opposite). Results showed that both tempo and mode can indeed influence emotions in isolation, with fast tempo and major mode being associated with happy emotions (and the opposite with sad emotions). Moreover, their combination resulted in "significantly more extreme ratings". Regarding the experiment with diverging tempo and mode, the observed emotion ratings were less pronounced, with the authors concluding that "in the presence of conflicting information, subjects tend to rely more on tempo information than on mode in their judgements". Interactions between tempo, mode and texture has been studied by other researchers, as discussed in (Webster & Weir, 2005).

Ilie et al. studied the consequences of manipulating intensity, rate, and pitch height in music and speech, by asking participants to rate 64 music and 64 speech excerpts in

terms of valence, tension arousal and energy arousal (Ilie & Thompson, 2006). In addition to the results observed in isolation, the authors identified several interactions as being statistical significant. Namely, for music stimuli, the valence ratings suggested that "the effects of intensity were greater when pitch height was high than when it was low", with lower valence being reported to loud-high pitched music. Moreover, fast-tempo music excerpts were judged as more pleasant when the pitch height was low than when it was high. Regarding speech, the authors reported an interaction between intensity and pitch height, where "soft high-pitched voices were perceived as more pleasant than any other pitch height and intensity combination". In addition to valence, a three-way interaction was observed in energy arousal speech ratings, relating rate, pitch height, and intensity.

Regarding rhythm, melodic contour and melodic progression, Lindström identified five two-way interactions, with the more significant being between melodic contour and direction (Lindström, 2006). In addition, a complex three-way interaction between rhythm, contour and direction was also observed in the participants' ratings of stability – instability.

Chapter 3

Music Emotion Recognition Literature Review

The creation of new knowledge requires a deep understanding about what has been done in a specific field and which questions remain unanswered. The previous chapter explored the basic theoretical knowledge needed to understand the music and emotion components of the MER field.

This chapter delves into the technical part of the field, analyzing how the existing musical dimensions have been captured computationally, exploring the most relevant works and strategies of music emotion recognition, the existing ground truth and current limitations.

Section 3.1. Standard Computational Audio Features

With that in mind, we first explore the available computational algorithms that have been proposed over the years in the Music Information Retrieval (MIR) field and used in MER. The gathered knowledge is organized by musical dimensions, as previously described in Chapter 2, to better identify possible gaps in the area.

Section 3.2. Music Emotion Recognition Approaches

Next, a comprehensive review of MER history is presented. To this end, we start with a generic overview, explaining how the typical MER approach is organized, from ground-truth collection and existent datasets to feature extraction and emotion recognition. Finally, we build on this knowledge to present the historical timeline of the MER field in the last three decades.

3.1. Standard Computational Audio Features

In general terms, a feature is a notable or characteristic part of something. Features help to distinguish one thing from another, by providing the essential descriptive primitives by which individual objects or works may be identified (Huron, 2001).

In musical terms, features may be characteristic of a musical work, of a movement, of a composer, of a very specific musical dimension, of a genre, and so forth. As Huron states, "what constitutes a feature depends on the scope of our gaze" (Huron, 2001). For illustration, features can be employed to represent any aspect that is relevant to the identification of a song, from the chords, to abstract statistics regarding physical aspects of the sound wave, rhythm information and others. Summing it up, the goal of feature extraction is to reduce the information of songs to descriptors that can fully describe them (X. Yang et al., 2017).

Over the last decades, several algorithms have been proposed to extract information from audio signals. These features have been developed to solve a myriad of problems, from speech recognition, to content-based retrieval, indexing, and fingerprinting. Nowadays, most of these are implemented in state-of-the-art audio frameworks, commonly used by most MIR studies. Hereafter, we term such features "standard computational audio features".

3.1.1. State-of-the-Art Audio Frameworks

Audio features represent information extracted from an audio signal and are the basis of diverse research fields such as music emotion recognition, digital audio effects, music fingerprinting and similarity measures, among others. Although different, these problems rely heavily on common audio features, such as zero crossing rate or signal energy.

Over the years, a range of audio frameworks[39] has been developed, implementing many of the audio features proposed in the literature. The available frameworks vary greatly in many aspects, from user-friendliness to computational efficiency or the number of implemented algorithms. Some are aimed to research, requiring specific environments (e.g., MATLAB), while others are designed with performance in mind, more suited to be used in the industry. Below we introduce some of the most commonly used in the MER research field, for an in-depth review see (Moffat, Ronan, & Reiss, 2015).

Marsyas

Marsyas[40] (Tzanetakis, 2002), an acronym of Music Analysis, Retrieval and Synthesis for Audio Signals, is an open-source audio framework created by George Tzanetakis and other researchers. It was developed for audio processing with specific emphasis on MIR applications and used for a variety of projects in both academia and industry[41]. It is one

[39] Computational tools and libraries used to extract audio features.

[40] http://marsyas.info/

[41] http://marsyas.info/about/projects.html

of the most computationally efficient frameworks available, in part due to the fact of being written in highly optimized C++ code, as opposed to more academic alternatives using MATLAB or interpreted languages.

In addition to the library, some command line tools are also provided, capable of, amongst other things, extracting features and training classifiers. The native integration with Qt, a multi-platform development framework, provides the foundations to create full, portable, applications with graphical user interfaces.

Although the number of features extracted by Marsyas is high, this is in part due to various statistical moments used to summarize features which output time series. As concluded in previous works (Panda & Paiva, 2011a), the framework lacks some features that have been identified as relevant for MER. Other problems are related with the lack of detail in some of the documentation, complex application programming interface and syntax to build and control the audio processing networks.

MIR Toolbox

MIR Toolbox[42] (Lartillot & Toiviainen, 2007) is a MATLAB framework implementing several algorithms specific to the extraction of musical features. The software is very modular and complex algorithms are built of smaller more elementary functions and mechanisms. This approach allows for the usage and combination of these minimal blocks independently as required.

The framework provides several graphical representations to export and graphically visualize the extracted information. Alternatively, these can be omitted to use in a batch scripting approach. A great number of both low and high-level features is available, some of which proved interesting to MER problems in previous works (Panda & Paiva, 2011a; Y.-H. Yang & Hu, 2012). The framework has also been used by some of the best performing algorithms in MIREX mood classification contests (e.g., (Panda & Paiva, 2012a; J.-C. Wang, Lo, Jeng, & Wang, 2010). One of its advantages is the quality of the documentation provided, especially when compared with other frameworks. Since it is built on top of MathWorks' MATLAB and MathWorks' Signal Processing Toolbox, this is its biggest drawback, as both are very resource intensive commercial products, hindering its usage on scenarios where lots of data needs to be processed.

PsySound

Currently in version 3, PsySound3[43] is a MATLAB toolbox for the analysis of sound recordings (Cabrera, Ferguson, & Schubert, 2007). Its aim is to provide precise analysis

[42] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox

[43] http://psysound.wikidot.com/

using standard acoustical measurements, as well as implementations of psychoacoustical and musical models, such as loudness, sharpness, roughness, fluctuation strength, pitch, rhythm and Interaural Cross Correlation (IACC). Compared to the alternatives, this framework extracts a smaller set of features, while being slower and much more computer- and memory-intensive.

Although cited in some literature (e.g., (Y.-H. Yang, Lin, Su, et al., 2008)) as having several features relevant to emotion recognition, older studies used a previous release of PsySound (version 2), available only for Mac PowerPC architecture computers. Since then, the program was rewritten in MATLAB, resulting in PsySound3. The current version, which has not been updated in the last years[44], still contains several issues and lacks stability, which makes the most relevant features identified in older studies (Y.-H. Yang, Lin, Su, et al., 2008) hard or impossible to replicate. Even with the existent issues and lacking comprehensive documentation and usage instructions, the framework is quite relevant since it implements several features that might be unavailable in alternative frameworks.

jMIR

jMIR[45] is an open-source Java software package for music information retrieval research (Mckay, 2010). Its development started in 2005 at McGill University, with the initial goal of providing a framework to eliminate the effort in calculating features from audio signals, providing a wide range of analysis algorithms suitable to MIR tasks by default. To complement this and decrease the slope of the learning curve, the application also provides an easy to use graphical user interface, making feature selection and audio processing straightforward. Since it was written in Java, the framework is portable across several operating systems, although having a lower performance when compared to solutions such as Marsyas.

Nowadays, jMIR includes not only feature extractors from several sources (e.g., audio, lyrics, symbolic files), but also other tools, namely, machine learning algorithms, heuristic error checkers and metadata mining and analysis tools. To this end, the suite is composed of several components, among which:

- jAudio: extracting low and high-level features from audio
- jLyrics: mining lyrics from the web and extracting textual features
- jSymbolic: extracting high-level features from symbolic music encodings
- jWebMiner: extracting cultural features from the internet

[44] https://github.com/densilcabrera/psysound3

[45] https://sourceforge.net/projects/jmir

Essentia

Essentia[46] is an open-source C++ library for audio analysis and audio-based music information retrieval (Bogdanov et al., 2013). It has been designed with a focus on robustness, performance and optimality of the provided algorithms both in terms of memory and speed, making it ideal for industrial solutions. It is also cross-platform, fully supporting Linux and Mac OS X, and partially supporting a longer range of platforms such as Windows, mobile (iOS and Android) or web (JavaScript). In addition, the library can be used in Python (via python wrappers) and contains several examples, command line tools and extensions. As a result, it has been used in several projects, from large web platforms such as freesound.org[47] (for large-scale indexing and content-based search), to Android and iOS apps such as the Freetello[48] (guitar learning app).

The library provides an extensive number of algorithms for audio input/output, signal processing and extraction of musical descriptors. In addition, Essentia is also able to estimate higher-level music descriptors such as genre, danceability, mood, dynamic complexity, voice or instrumental and others (for details see Section 3.1.10). To this end, Gaia[49] (a C++ library for similarity and classification) is used to train classification models and output results.

LibROSA

LibROSA[50] is an open-source python package for music and audio analysis (Mcfee et al., 2015). It provides implementations of several functions and algorithms necessary to create music information retrieval systems. It has been designed to be easy to use, especially for researchers familiar with MATLAB, with standardized interfaces and variable names and to achieve "backwards compatibility against existing reference implementations" as much as possible (Mcfee et al., 2015). As with other frameworks, the functions are modular and can be combined in new, different ways by researchers to explore new features.

The features available from LibROSA are mostly spectral features, with a few rhythmic descriptors also available. While the package is well documented and easy to use, its main drawback is the algorithmic efficiency, being the slowest of all the frameworks compared in (Moffat et al., 2015).

The audio frameworks described above are able to extract a high number of features,

[46] http://essentia.upf.edu/

[47] https://freesound.org/

[48] https://fretello.app/

[49] https://github.com/MTG/gaia

[50] https://librosa.github.io/

and new ones are added as they are proposed in new research. The most common features, such as the spectral related ones, are typically available across every audio framework, something that does not happen with the more complex or newer ones. Next, we catalog the audio features that have been proposed in the literature over the years and are now available in these frameworks, organizing them according to the musical dimensions to which they are closer to. In thus ways, we aim to: 1) get a better picture of the current state-of-the-art regarding audio features extraction; 2) understand how they relate with the musical dimensions and attributes that are relevant to music emotion; 3) and identify musical dimensions that are known to be relevant to MER but might be lacking in terms of computational extractors available.

Many of the features are extracted repeatedly for smaller excerpts (analysis windows) of the entire audio clip, returning series of data. These frame-level features are normally integrated using statistical moments such as mean, standard deviation, skewness and kurtosis, maximum and minimum before being used with machine learning techniques.

We selected three of the abovementioned audio frameworks to use in our research: Marsyas, MIR Toolbox and PsySound3. Several factors contributed to this selection, namely, these have been consistently used in previous MER works, were part of the best performing algorithms in MIREX AMC task over the years and together capture a very high number of audio features as detailed below and in (Moffat et al., 2015). For each of the features described in the next sections, we also indicate audio frameworks which implement such algorithms. This indication is not comprehensive but rather indicates at least one framework where the implementation is available.

3.1.2. Melody Features

In this section we describe the audio features that capture information primarily related with melody and its components, as described in Section 2.3.1.

Pitch

Pitch represents the perceived fundamental frequency of a sound. It is one of the three major auditory attributes of sounds, along with loudness and timbre. Pitch (as an audio feature) typically refers to the fundamental frequency of a monophonic sound signal and can be calculated using various different techniques (Rabiner, Cheng, Rosenberg, & McGonegal, 1976). Many frameworks, such as the MIR Toolbox or Marsyas, implement pitch extraction algorithms, returning the results either as continuous pitch curves or as discretized note events. One of the most common methods to calculate pitch, employed in Marsyas, MIR Toolbox and Essentia is the YIN algorithm (de Cheveigné & Kawahara, 2002). PsySound3 also implements Swipe and Swipe' (Sawtooth Waveform Inspired

Pitch Estimator) algorithms proposed by Camacho (Camacho, 2007).

Available in: MIR Toolbox, Psysound3, Marsyas and Essentia.

Virtual Pitch Features

Ernst Terhardt et al. proposed an algorithm to extract virtual pitch, which is concerned with the psychoacoustics and modeling of the perceived pitch (Terhardt, Stoll, & Seewann, 1982). The PsySound3 framework implements this algorithm using the parameters described by Parncutt (Parncutt, 1989, Chapter 6), extracting the following descriptors:

- Virtual Pitch Pattern – strength of virtual pitches (using Terhardt's algorithm)
- Spectral Pitch Pattern – strength of spectral pitches (using Terhardt's algorithm)
- Chroma Pattern – Chroma salience according to Parncutt (Parncutt, 1989, p. 146)
- Pure Tonalness – reflects the audibility of spectral pitches
- Complex Tonalness – reflects the audibility of virtual pitch, an estimation of how similar the spectrum is to a harmonic (or complex) tone
- Multiplicity – estimation of the number of tones simultaneously noticed in a sound (this particular descriptor can also be related with texture)
- ChordChangeLikelihood – although extractable, no details are provided in the documentation.

Available in: Psysound3.

Pitch Salience

The perception of pitch, in particular its salience is a complex idea that can be roughly explained as how noticeable (that is, strongly marked) is the pitch in a sound, and was proposed as a quick measure of tone sensation. Pure tones have an average pitch salience value close to 0 whereas sounds containing several harmonics in the spectrum have higher salience values.

Different approaches have been proposed to extract this feature. The pitch salience function calculates the salience of a signal frame given its spectral peaks. To this end, the salience of a given frequency is computed as the sum of the weighted energies found at integer multiples (harmonics) of that frequency (Salamon & Gómez, 2012).

A second approach implemented in Essentia computes the pitch salience of a spectrum, as the ratio of the highest auto correlation value of the spectrum to the non-shifted auto correlation value (Ricard, 2004). MIR Toolbox also implements this strategy and is able to output the salience, sometimes referred as "amplitude" in the manual, as described in (Lartillot, 2018, p. 145).

Available in: Essentia, MIR Toolbox.

Predominant Melody F0

Several authors have proposed algorithms to estimate the fundamental frequency, f_0, of the predominant melody in both polyphonic and monophonic music audio signals. This is still an open research problem, and most of the audio frameworks do not include polyphonic audio melody F0 extractors. Still, some of the proposed algorithms are nowadays available as separate tools. Two of these cases are the MELODIA algorithm (Salamon & Gómez, 2012), that is freely available online[51], and the Dressler's algorithm (Dressler, 2016), currently of restrict access.

Currently, the Essentia framework implements an algorithm for Melody F0 estimation based on the MELODIA algorithm. Still, the current implementation in Essentia is limited to monophonic signals. The approach is based on creation of time continuous sequences of pitch candidates grouped using auditory streaming cues, known as pitch contours. Other frameworks such as Marsyas also provide tools for monophonic pitch extraction.

Available in: Essentia[52]

Pitch content

In his PhD thesis, Tzanetakis defines a set of simple features extracted from folded[53] and unfolded pitch histograms to describe pitch information (Tzanetakis, 2002, p. 51):

- FA0: Amplitude of the maximum peak of the folded histogram. This corresponds to the most dominant pitch class of the song. For tonal music, this peak will typically correspond to the tonic or dominant chord. This peak will be higher for songs that do not have many harmonic changes;
- UP0: Period of the maximum peak of the unfolded histogram. This corresponds to the octave range of the dominant musical pitch of the song;
- FP0: Period of the maximum peak of the folded histogram. This corresponds to the main pitch class of the song;
- IPO1: Pitch interval between the two most prominent peaks of the folded histogram. This corresponds to the main tonal interval relation. For pieces with simple harmonic structure, this feature will have a value of 1 or -1, corresponding to fifth or fourth intervals (tonic-dominant);

[51] http://www.justinsalamon.com/melody-extraction.html

[52] The implementation in Essentia is adapted / limited to monophonic signals.

[53] In the folded pitch histogram all notes are mapped to a single octave.

- SUM: The overall sum of the histogram. This feature is a measure of the strength of the pitch detection.

Although the author described these features in his PhD thesis about the Marsyas framework, the current documentation seems to ignore them. Due to this we could not confirm that the framework is able to extract them.

Available in: Marsyas (unconfirmed)

3.1.3. Harmony Features

In this section we describe the audio features that capture information primarily related with harmony and its components, as described in Section 2.3.2.

Inharmonicity

Inharmonicity measures the amount of partials that are not multiples of the fundamental frequency**.** Inharmonicity influences the timbric perception of a given sound. One approach to compute this was proposed by Peeters (Peeters, 2004, p. 17) and is implemented in Essentia. MIR Toolbox framework measures the inharmonicity as the amount of energy outside the ideal harmonic series, which presupposes that there is only one fundamental frequency (Lartillot, 2018, p. 147).

Available in: MIR Toolbox, Essentia.

Chromagram

The chromagram is used to estimate the energy distribution along pitch classes. It consists of a 12-positions vector, one for each note, from A to G# (12 semitone pitch classes), with the respective intensities in each of these classes based on the spectral peaks of the waveform. It is also known as Harmonic Pitch Class Profile (HPCP).

Some extractors use variations such as vectors multiple of 12, subdividing the pitch classes. As an example, Marsyas outputs two additional values, consisting in the minimum and average intensity for the chroma A (i.e., the set of A pitches separated by N octaves).

The MIR Toolbox implements an extractor to visualize the Pearson correlation values between chromas as a color map, which can be animated if frame decomposition is used. To this end, the chromagram is projected into a self-organizing map trained with the Krumhansl-Kessler profiles (Toiviainen & Krumhansl, 2003).

Available in: MIR Toolbox, Marsyas, and Essentia.

Tuning Frequency

The tuning frequency is an estimation of the exact frequency (in Hz) on which a song is tuned, based on the audio signal. It is used as an intermediary step for HPCP calculation and key estimation, but can also be applied for classification tasks such as western vs. non-western music (Gómez, 2005).

Available in: Essentia, librosa.

Key Strength

Key strength consists in the computation of the probability of each possible key candidate to be the key of a given song (e.g., outputting scores between 0 and 1, or -1 to 1). The algorithm is based on the cross-correlation of the chromagram (Gómez, 2006a, p. 103).

Available in: MIR Toolbox, Essentia.

Key and Key Clarity

These features give a broad estimation of tonal center positions and their respective clarity. This is obtained by peak picking in the key strength curve. There, the best key (or keys) is given by the peak abscissa, while the key clarity is the key strength associated with the best keys, i.e., the key ordinate (Lartillot, 2018, p. 156).

Available in: MIR Toolbox, Essentia.

Modality

Several algorithms exist to estimate modality, i.e., major vs. minor, returning either a binary label (e.g., major / minor) or a numerical value (e.g., between -1 (minor) and 1 (major)). Some of the common strategies use the estimated strength of each key and consist of:

- The difference between the strength of the strongest major and minor keys
- The sum of all the differences between each major key and its relative minor key pair

Available in: MIR Toolbox, Essentia.

Chord Sequence

Extracting chords from an audio signal is a complex task, for which researchers have yet to propose robust solutions. The existent methods to estimate this are still experimental, based on pitch profile classes (Gómez, 2006b; Temperley, 1999). Essentia implements an algorithm based on this research, able to compute the sequence of chords in a song.

Such algorithm calculates the best matching major or minor triad and outputs the result as a string (e.g., A#, Bm, G#m, C). The existing implementation is marked as experimental and requires further work before being usable.

While MIR Toolbox cites the same works by Gómez, its documentation only mentions chromagram extraction and the associated chroma classes, not documenting any method to output the best matching triad.

Available in: Essentia.

Tonal Centroid Vector (6 dimensions)

The tonal centroid is represented as a 6-dimensional feature vector. It corresponds to a projection of the chords along circles of fifths, of minor thirds and of major thirds (Harte, Sandler, & Gasser, 2006). It is based on the Harmonic Network or Tonnetz, which is a planar representation of pitch relations, where pitch classes having close harmonic relations such as fifths, major/minor thirds have smaller Euclidean distances on the plane. By calculating the Euclidean distance between successive analysis frames of tonal centroid vectors, the algorithm detects harmonic changes such as chord boundaries from musical audio (exemplified in Figure 3.1).

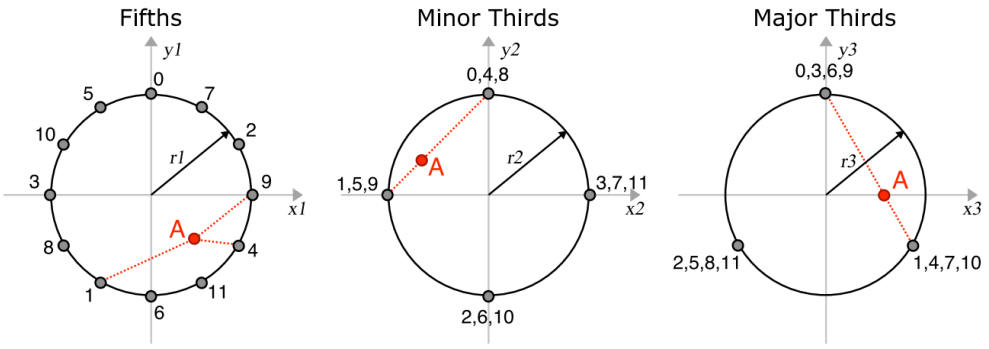


Figure 3.1: Tonal Centroid for the A major triad (pitch class 9, 1 and 4) is shown at point A (adapted from (Harte et al., 2006)).

Available in: MIR Toolbox.

Harmonic Change Detection Function

The Harmonic Change Detection Function (HCDF) is a method for detecting changes in the harmonic content of musical audio signals proposed by Harte et al. (Harte et al., 2006). It can be seen as the flux of the tonal centroid, as in the distance between the harmonic regions of successive frames (Harte et al., 2006).

Available in: MIR Toolbox.

Sharpness

Sound can be subjectively rated on a scale from dull to sharp, and sharpness algorithms attempt to model this. These algorithms are essentially weighted centroids of specific loudness. The unit of sharpness is the acum. One acum is defined as the sharpness of a band of noise centered on 1000 Hz, 1 critical-bandwidth wide, with a level of 60 decibels (dB) (sound pressure level). A 1000 Hz pure tone at 60 dB will have a similar sharpness. Psysound3 implements several algorithms for this.

Available in: Psysound3.

3.1.4. Rhythm Features

In this section we describe the audio features that capture information primarily related with rhythm and its components, as described in Section 2.3.3.

Rhythmic Fluctuation

This feature estimates the rhythm content of an audio signal. This estimation is based on spectrogram computation transformed by auditory modelling and then a spectrum estimation in each band (Pampalk, Rauber, & Merkl, 2002), in brief, the rhythmic periodicity along auditory channels. In the MIR Toolbox implementation this process can be described in two steps (Lartillot, 2018):

- First the spectrogram is computed on frames of 23 ms and half overlapping. Then the Terhardt outer ear modelling is computed, with Bark-band redistribution of the energy and estimation of the masking effects. Finally, the amplitudes are computed in deciBel (dB) scale.
- Then, a Fast Fourier Transform (FFT) is computed on each Bark band, from 0 to 10 Hz. The amplitude modulation coefficients are weighted based on the psychoacoustic model of the fluctuation strength (Fastl, 1982). The result is a matrix of the rhythmic periodicities for each different Bark band.

Available in: MIR Toolbox.

Beat Spectrum

The beat spectrum has been proposed as a measure of acoustic self-similarity as a function of time lag. It is computed from the similarity matrix, obtained by comparing the similarity between all possible pairs of frames from the original audio signal (Foote,

Cooper, & Nam, 2002).

Available in: MIR Toolbox.

Beat Location

Different beat tracking algorithms have been proposed over time. These algorithms estimate the beat locations given an input signal. A popular algorithm, proposed by Degara et al. uses statistics to estimate "the time between consecutive beat events and exploits both beat and non-beat information by explicitly modeling non-beat states" (Degara et al., 2012).

A newer approach, named multi-feature beat tracker, extends the idea of measuring the level of agreement between a committee of different beat tracking algorithms in a song-by-song basis, which attained higher results (Zapata, Davies, & Gómez, 2014). The Essentia framework implements these and other beat tracker and rhythm extractor functions. In case of the multi-feature algorithm, it combines 5 different beat trackers, taking into account the maximum mutual agreement between them. Marsyas provides the INESC-Porto Beat Tracker, a real-time/off-line tempo induction and beat tracking system based on a competing multi-agent strategy, which considers parallel hypotheses regarding tempo and beats (J. L. Oliveira, Gouyon, Martins, & Reis, 2010).

Available in: Essentia, Marsyas.

Onset Time

Another way of determining the tempo is based on first the computation of an onset detection curve, showing the successive bursts of energy corresponding to the successive pulses. Peak picking is automatically performed on the onset detection curve, to show the estimated positions of the notes. In the case of the MIR Toolbox, its onset function is able to return the onset times using any of the following options: peaks, valleys, attack phase and release phase using several algorithms as discussed in (Lartillot, 2018, p. 90)

Available in: MIR Toolbox, Essentia.

Event Density

This is a feature to estimate the "speed" of a song by estimating the average frequency of events, i.e., the number of note onsets per second (Lartillot, 2018, p. 99).

Available in: MIR Toolbox.

Average Duration of Events

The duration of events (e.g., a note) can also be estimated from its envelope. One possible approach to estimate this was proposed by Peeters (2004). It consists in detecting attack and release phases and measuring the period of time (in seconds) between them when the amplitude is at least 40% of the maximum.

Available in: MIR Toolbox.

Tempo

Several algorithms have been proposed to estimate the tempo, the speed or pace of a given musical piece, which is usually indicated in beats per minute (BPM). Tempo is typically estimated by detecting periodicities from the onset detection curve.

In case of the Essentia toolbox, this is obtained using the abovementioned Degara or Multi-feature beat tracking algorithms. The MIR Toolbox offers two approaches: 1) classical, based on detecting periodicities in a range of BPMs, and choosing the maximum periodicity score for each frame separately (Lartillot, 2010). In more complex pieces, such approach creates a tempo curve with various jumps, related with shifts in the metrical level; 2) metre-based, which tracks tempo by building a hierarchical metrical structure from the periodicities found in the event detection curve (Lartillot, Cereghetti, Eliard, & Trost, 2013). In complex cases, this approach is typically able to find coherent metrical levels leading to a continuous tempo curve.

Marsyas offers several different methods to estimate tempo using the provided tempo estimation tool[54]. According to the source code, 11 different methods are implemented, with the default being the simple tempo estimation method (STEM). This method is based on two musical rhythm properties: 1) the music signal tends to be self-similar at periodicities related to the underlying rhythmic structure; and 2) rhythmic events tend to be spaced regularly in time, as described in (Tzanetakis & Percival, 2013).

Available in: MIR Toolbox, Essentia, and Marsyas.

Tempo Change

An indicator of tempo change over time is estimated by computing the difference between successive values of the tempo curve in the MIR Toolbox. This musical descriptor is expressed independently from the choice of a metrical level by computing the ratio of tempo values between successive frames and is expressed in logarithmic scale (base 2). Thus, a value of 0 is observed when there is no change in tempo for two consecutive frames; tempo increasing gives a positive value; and tempo decreasing gives negative values (Lartillot, 2018, p. 105).

[54] https://github.com/marsyas/marsyas/blob/master/src/apps/tempo/tempo.cpp

Available in: MIR Toolbox.

Metrical Structure

This feature provides a detailed description of the hierarchical metrical structure by detecting periodicities from the onset detection curve and tracking a broad set of metrical levels (Lartillot et al., 2013). This extractor is used to calculate the metre-based tempo estimation in the MIR Toolbox.

Available in: MIR Toolbox.

Metrical Centroid and Strength

These functions provide two descriptors derived from the abovementioned metrical analysis carried out by the MIR Toolbox:

1. Dynamic metrical centroid: estimation of the metrical activity, based on the computation of the centroid of the selected metrical levels. The resulting metrical centroid curve indicates the temporal evolution of the metrical activity expressed in BPM, so that the values can be compared with the tempo values also in BPM. According to the documentation, "high BPM values for the metrical centroid indicate that more elementary metrical levels (i.e., very fast levels corresponding to very fast rhythmical values) predominate. On the contrary, low BPM values indicate that higher metrical levels (i.e., slow pulsations corresponding to whole notes, bars, etc.) predominate. If one particular level is particularly dominant, the value of the metrical centroid naturally approaches the corresponding tempo value on that particular level" (Lartillot, 2018, p. 109).
2. Dynamic metrical strength: an indicator of the clarity and strength of the pulsation. Estimates whether a "clear and strong pulsation, or even a strong metrical hierarchy is present", or if the opposite is true, where "the pulsation is somewhat hidden, unclear" (Lartillot, 2018, p. 109) or a complex mix of pulsations.

Available in: MIR Toolbox

Pulse / Rhythmic Clarity

This feature estimates the "rhythmic clarity", an indicator of the clarity and strength found in the beats estimated by tempo estimation algorithms. Distinct heuristics exist to this estimation. The most common uses the autocorrelation curve that is computed during tempo estimation (Lartillot, Eerola, Toiviainen, & Fornari, 2008).

Essentia computes an approximate metric calling it beats loudness. This feature represents the loudness computed only on the beats, that is, the spectrum energy of beats in an audio signal given their positions. The energy is computed both on the whole frequency range and for each of the specified frequency bands. To this end, the onset of a specific beat given as the input segment is detected. From the window starting on this detected onset, the spectrum is computed and the energy estimated in several bands, as defined by Scheirer (Scheirer, 1998). The same model by Scheirer is also available under MIR Toolbox.

Available in: MIR Toolbox, Essentia.

Predominant Local Pulse (PLP) Novelty Curves

Grosche and Muller (2009) introduced a novel mid-level representation for capturing dominant tempo and predominant local pulse even from music with weak non-percussive note onsets and strongly fluctuating tempo.

Instead of following the tradicional approach relying on the detection of note onsets, the authors derive a tempogram using local spectral analysis of a novelty curve. From it, the predominant tempo for each time instant is estimated, as well as "a sinusoidal kernel that best explains the local periodic nature of the novelty curve". Finally, the predominant local pulse is obtained by accumulating all the local kernels over time. The obtained curves "are robust to outliers and reveal musically meaningful periodicity information even in the case of poor onset information".

While the PLP curve does not represent high-level information such as tempo, beat level or location of onset positions, it serves as a tool, which may then “be used for tasks such as beat tracking, tempo and meter estimation”.

Available in: Essentia.

Harmonically Wrapped Peak Similarity (HWPS)

In his PhD thesis, Tzanetakis described a set of rhythmic content features calculated with recourse to the Beat Histograms (BH) of a song (Tzanetakis, 2002, p. 48) which proved useful for musical genre classification:

- A0, A1: relative amplitude (divided by the sum of amplitudes) of the first (A0), and second (A1) histogram peak;
- RA: ratio of the amplitude of the second peak divided by the amplitude of the first peak;
- P1, P2: Period of the first (P1) and second (P2) peak in BPM;
- SUM: overall sum of the histogram (indication of beat strength)

Subsequently, HWPS, a feature following similar principles has been proposed to calculate harmonicity by taking “into account spectral information in a global manner”

(Lagrange, Martins, & Tzanetakis, 2008).

Available in: Marsyas[55].

3.1.5. Dynamics Features

In this section we describe the audio features that capture information primarily related with dynamics and its components, as described in Section 2.3.4.

Root-Mean-Square (RMS) Energy

The RMS energy is used to measure the power of a signal over a window, or global energy. It roughly describes the loudness of a musical signal. The global energy of a signal can be computed by taking the root mean of the square of the amplitude, also known as root-mean-square (RMS) (Westfall, 2014).

Available in: MIR Toolbox, Marsyas, and Essentia.

Low Energy Rate

Low energy rate, also known as less-than-average energy, measures the percentage of frames with less-than-average energy (Tzanetakis & Cook, 2002). This metric estimates the temporal distribution of energy, in order to understand if this energy remains constant between frames or if some frames are more contrastive than others.

Available in: MIR Toolbox.

Sound Level

This descriptor corresponds to the power sum of the spectrum for each time window, expressed in deciBel. At a higher level, when appropriately calibrated, this represents the unweighted sound pressure level of the signal in each analysis window (Cabrera et al., 2007).

Available in: Psysound3.

Instantaneous Level, Frequency and Phase

These features consist in applying a Hilbert transform to the audio waveform

[55] Difficult to use since no instructions are provided in the official documentation, only a reference in the source code of the toolbox.

(Khvedelidze, 1990), resulting in three different outputs: the instantaneous level, instantaneous frequency and instantaneous phase. The instantaneous level can be regarded as the sound pressure level derived from the Hilbert transform.

By using a Hilbert transform, the signal envelope is extracted easily without an integrating function (the instantaneous amplitude). To this end, the input signal is phase-shifted by -90 degrees and used as the imaginary part of a complex waveform, while the original wave forms the real part[56]. The instantaneous level is obtained by transforming the magnitude of the complex waveform into the decibel scale, while the instantaneous phase is the phase of the complex waveform. Differentiating this phase with an appropriate coefficient yields the instantaneous frequency (Cabrera et al., 2007).

Available in: Psysound3.

Loudness

As described in Section 2.3.4, sound loudness is the subjective perception of the intensity of a sound. This metric is measured in sones, where a doubling in sones corresponds to a doubling of loudness. An audio signal with silence has a loudness of 0 sones, while a 1 kHz tone at 40 dB presented has a loudness of 1 sone.

Several loudness have been proposed over the years, namely:

- Leq – Equivalent sound level (Soulodre, 2004)
- LARM single-band loudness model (Skovenborg & Nielsen, 2004)
- HEIMDAL multi-band loudness model (Skovenborg & Nielsen, 2004)
- Loudness according to Steven's power law (Stevens, 1975)
- Vickers's loudness (Vickers, 2001)

An evaluation of different loudness models is available in (Skovenborg & Nielsen, 2004).

Available in: Psysound3, Essentia.

Timbral Width

The timbral width is one of six measures of timbre proposed by Malloch (1997) in a method called loudness distribution analysis. In this case, timbral width can be viewed as "a measure of the flatness of a loudness function" (Y.-H. Yang & Chen, 2011a, p. 66), or as the author puts it "a measure of the fraction of loudness that lies outside of the loudest band, relative to the total loudness" (Malloch, 1997, p. 52).

Available in: Psysound3.

[56] http://psysound.wikidot.com/system:hilbert

Volume

Volume refers roughly to the perceived "size" of the sound, or the auditory volume of pure tones. This concept was first studied by Stevens (Stevens, 1934) and, later on, Cabrera (Cabrera, 1999) developed a computational volume model for arbitrary spectra. In his work, Cabrera proposes two diotic[57] volume models. The first uses a weighted ratio between the binaural loudness and sharpness, which is the specific loudness centroid on the Bark scale. A second and better performing model uses a simpler centroid to overcome limitations in the method of sharpness calculation selected by the authors (Cabrera, 1999).

Available in: Psysound3.

Sound Balance

Maximum Amplitude Position to Total Envelope Length Ratio (MaxToTotal and MinToTotal) is a metric to understand how much the maximum amplitude (peak) in a sound envelop is off the center. To this end, the ratio between the index of the maximum (or minimum) value of the envelope of a signal and the total length of the envelope is computed. If the peak amplitude is found close to the beginning (e.g., decrescendo sounds), this ratio will be close to 0. A value of 0.5 means that the peak is close to the middle and near 1 if at the end of the sound (e.g., crescendo sounds).

Although not implemented directly in many frameworks, these metrics are simple statistics that can be easily computed from the output of existent extractors. For illustration, MIR Toolbox provides all the information needed to compute MaxToTotal or MinToTotal under its event detection function (i.e., time stamp and amplitude of the onset start, attack, decay and offset phases) with a simple ratio of the amplitude to the difference between offset and onset times.

Another metric to estimate how the sound is 'balanced' over its duration is the temporal centroid to total length ratio (TCToTotal). This is measured by computing the ratio of the temporal centroid to the total length of a signal envelope. A value near 0 is observed if the energy is concentrated at the beginning of a sound (e.g., decrescendo), close to 0.5 if the energy distribution is symmetric and closer to 1 if the energy is mostly at the end (e.g., crescendos).

Available in: Essentia, MIR Toolbox[58]

[57] Involving the simultaneous presentation of the same stimulus to each ear.

[58] As discussed, these features can be easily calculated with information from available extractors in other frameworks.

3.1.6. Tone Color Features

In this section we describe the audio features that capture information primarily related with tone color and its components, as described in Section 2.3.5.

Attack/Decay Time

As described in Section 2.3.5, one of the aspects influencing tone color is the sound envelope, which can be divided into four parts: attack, decay, sustain and release. Several descriptors can be extracted from it, mostly related with the attack phase – from the starting point of the envelope until the amplitude peak is attained.

One of these descriptors is the attack time, which consists in the estimation of temporal duration of the various attack phases in an audio signal (e.g., for a signal to rise to its peak amplitude).

Various frameworks implement extractors for this descriptor, each with slight variations. As an example, the MIR Toolbox is able to output the attack times using a linear scale or a log scale, as proposed by Krimphoff et al. (1994), as well as the time of the decay phase.

Another variation of the attack time is the log attack time, an algorithm that computes the log (base 10) of the attack time of a signal envelope. Both the MIR Toolbox and Essentia are able to extract such feature. Essentia defines specific parameters to account for noise presence (e.g., attack starts only when the signal envelope reaches 20% of its maximum value) and specificities of some instruments (e.g., attack end point is defined as 90% of its maximum value). MIR Toolbox follows the implementation proposed by Peeters et al. (2011).

MIR Toolbox is also able to compute the decay time, the temporal duration of the decay phase.

Available in: MIR Toolbox, Essentia.

Attack/Decay Slope

The attack slope is another descriptor extracted from the attack phase. It consists on the estimation of the average slope of the entire attack phase, since its start to the peak. The MIR Toolbox is also able to extract the same information from the decay phase, related to its decrease slope. Some of the common methods to compute the slope are based on the amplitude difference from the start of the attack to the amplitude peak, divided by its duration, or the average slope weighted by a Gaussian curve as proposed by Peeters (2004).

Available in: MIR Toolbox.

Attack/Decay Leap

The attack leap is a simple descriptor related to the attack phase. It consists in the estimation of the amplitude difference between the beginning (bottom) and the end (peak) of the attack phase. As with the previous features, MIR Toolbox outputs a similar descriptor related with the decay phase.

Available in: MIR Toolbox.

Zero Crossing Rate

The Zero Crossing Rate (ZCR) represents the number of times the waveform changes sign in a window (crosses the x-axis). It can be used as a simple indicator of change of frequency or noisiness. As an example, heavy metal music, due to guitar distortion and heavy percussion, will tend to have much higher zero crossing values than classical music (Tzanetakis, 2002, p. 42).

Sometimes the ZCR derivative is also computed, representing the absolute value of the window-to-window change in zero crossing rate.

Available in: MIR Toolbox, Marsyas, and Essentia.

Spectral Flux

Spectral flux is a measure of the amount of spectral change in a signal, i.e., the distance between the spectra of successive frames (Tzanetakis, 2002, p. 33). Spectral flux has also been shown by user experiments to be an important perceptual attribute in the characterization of the timbre of musical instruments (Grey, 1975).

Available in: MIR Toolbox, Marsyas, and Essentia.

Spectral Centroid

The spectral centroid is a measure of spectral shape. Information about the shape of a distribution can be obtained through the use of its (statistical) moments. The first moment, called the mean, is the geometric center (centroid) of the distribution and is a measure of central tendency for the random variable. It is defined as the center of gravity of the magnitude spectrum of the short-time Fourier Transform (STFT) (Tzanetakis, 2002, p. 32).

According to (Grey, 1975), the spectral centroid has been shown by user experiments to be an important perceptual attribute in the characterization of the timbre of musical instruments. A higher centroid value can also be an indicator of the "brightness" or "sharpness" of the sound (Lichte, 1941).

Available in: MIR Toolbox, Psysound3, Marsyas, and Essentia.

Spectral Spread

The spectral spread represents the standard deviation of the spectrum, calculated as the square root of the variance. Thus, it is a measure of the dispersion or spread of the spectrum.

Available in: MIR Toolbox, Essentia, and Psysound3.

Spectral Skewness

The third central moment of a given distribution is called the skewness and it is a measure of its symmetry.

The coefficient of skewness is computed as the ratio of the skewness to the standard deviation raised to the third power. It is many times used as an alternative to skewness, due to its better suited range, often between -3.0 and 3.0 for data from natural systems (Lartillot, 2018, p. 184).

Available in: MIR Toolbox, Essentia, and Psysound3.

Spectral Kurtosis

The exact meaning of the kurtosis has been disputed for long. The classic interpretation, which applies to symmetric and unimodal distributions (those with skewness of 0), is that kurtosis measures the "peakedness" of the distribution and the heaviness of its tail. Other interpretations, such as "lack of shoulders" (where the "shoulder" is defined loosely as the area between the peak and the tail, or more specifically as the area about one standard deviation from the mean) or "bimodality". Recently, this ambiguity has been settled, with Westfall noting that kurtosis "(...) only unambiguous interpretation is in terms of tail extremity; i.e., either existing outliers (...) or propensity to produce outliers (...)" (Westfall, 2014) . In simple terms, kurtosis captures information about existing outliers, measuring nothing about the peak.

Available in: MIR Toolbox, Essentia, and Psysound3.

Spectral Flatness

The spectral flatness indicates whether the spectrum distribution is smooth or spiky, i.e., estimates to which degree the frequencies in a spectrum are uniformly distributed (noise-like). It is computed as the ratio between the geometric mean and the arithmetic mean. Some frameworks adopt a different approach, calculating the spectral flatness in different spectral bands. Marsyas is one of these, naming it spectral flatness measure (SFM).

A high spectral flatness indicates that the spectrum has a similar amount of power in all spectral bands (flat spectrum) – this would sound similar to white noise, and the

graph of the spectrum would appear relatively flat and smooth. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands – this would typically sound like a mixture of sine waves, and the spectrum would appear "spiky".

Available in: MIR Toolbox, Essentia, and Marsyas.

Spectral Crest Factor (SCF)

Spectral crest factor (Allamanche, Hellmuth, Fröba, Kastner, & Cremer, 2001) is a measure of the "peakiness" of a spectrum and is inversely proportional to the spectral flatness measure, both proposed in the context of the MPEG-7 standard (Casey, 2002). It is commonly used to distinguish noise-like from tone-like sounds due to their different spectral shapes, where noise-like sounds have lower spectral crests. Both spectral flatness and SCF are often used together in audio fingerprinting problems (e.g., (Doets, Gisbert, & Lagendijk, 2006)).

The SCF is computed as the ratio of the maximum spectrum power and the mean spectrum power of a subband.

Available in: Marsyas.

Spectral Contrast

The octave-based spectral contrast is a feature proposed by Jiang et al. (Jiang, Lu, Zhang, Tao, & Cai, 2002) to represent the spectral characteristics of an audio signal, specifically the relative spectral distribution. According to the authors, the feature has been tested in music type classification problems, demonstrating a “better discrimination among different music types than mel-frequency cepstral coefficients (MFCC)” (Jiang et al., 2002), which is one of the features typically used in such problems.

The Essentia library implements a modified version of this algorithm, which improves its discriminative power and robustness (Akkermans, Serrà, & Herrera, 2009).

Available in Essentia, librosa.

Spectral Entropy

The spectral entropy of a signal is a measure of its spectral power distribution. It is a feature based on Shannon entropy (Shannon, 1948) from the information theory field.

The spectral entropy has been widely used in fields such as speech recognition (Shen, Hung, & Lee, 1998; Toh, Togneri, & Nordholm, 2005) and biomedical signal processing (Vakkuri et al., 2004).

Available in: MIR Toolbox, Essentia.

Spectral Rolloff

Spectral rolloff is often used as an indicator of the skewness of the frequencies present in a window. According to the (Tzanetakis, 2002, p. 33), the spectral rolloff is defined as the frequency R_t below which 85% of the magnitude distribution is concentrated. The percentage varies among authors, with 85% being the current default value for most frameworks following (Tzanetakis, 2002, p. 33), while (Pohle, Pampalk, & Widmer, 2005) propose 95%.

Available in: MIR Toolbox, Marsyas, and Essentia.

High-frequency Energy

Several algorithms have been proposed to estimate the high-frequency content in a signal. Brightness (also called high-frequency energy) is one of such algorithms implemented in MIR Toolbox. This typically consists in fixing a minimum frequency value, and measuring the amount of energy above that frequency. The result is expressed as a number between 0 and 1.

Distinct cut-off frequency values to brightness have been proposed over the years, e.g., 1500 Hz (Lartillot, 2018, p. 132), 1000 Hz (Laukka, Juslin, & Bresin, 2005) and 3000 Hz (Juslin, 2000).

The Essentia framework implements a different algorithm, named high-frequency content (HFC), to measure the amount of high-frequency energy in the audio signal (from the signal spectrum). HFC is computed by applying one of the several algorithms proposed, such as (Masri & Bateman, 1996), (Jensen & Andersen, 2003) and (Brossier, Bello, & Plumbley, 2004) to the short-time Fourier transform spectrum. For reference, Essentia library implements the three variations described above.

As with other high-energy related features, it has no perceptual base but, still, has been used in applications such as onset detection.

Available in: MIR Toolbox, Essentia.

Cepstrum (Real / Complex)

A cepstrum, derived by reversing the first four letters of "spectrum", is the result of taking the inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. There are four distinct cepstrum definitions: power cepstrum (Bogert, Healy, & Tukey, 1963), complex cepstrum (Oppenheim, 1965), real cepstrum and phase cepstrum.

Cepstrum can be viewed as measuring the rate of change in the different spectral bands and has applications in fields such as pitch analysis, echo detection and especially in human speech processing (especially the power spectrum), by providing a simple way

to separate formants (due to filtering in the vocal tract) from the vocal source (e.g., (Noll, 1967)).

Psysound3 contains analyzers to conduct both real and complex cepstral analysis for each window of an audio signal. From these, the framework is able to output various features: average power spectrum, cepstral moments, cepstrogram, liftered spectrogram, average liftered power spectrum, level and liftered spectral moments. MIR Toolbox provides only a limited set of these features.

Available in: Psysound3.

Energy in Mel/Bark/ERB Bands

In audio signal processing it is often interesting to decompose the original signal into a series of audio signals of different frequencies (i.e., low to high-frequency channels), enabling the study of each channel separately. This idea is inspired by the human cochlea, which can be regarded as a filter bank, distributing the frequencies into critical bands. Several scales have been proposed, each one using a particular range of frequencies.

The Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another (Stevens, Volkmann, & Newman, 1937). The reference point between this scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone, 40 dB above the listener's threshold.

Most audio frameworks today implement some of the various existing algorithms (Ganchev, Fakotakis, & Kokkinakis, 2005) to compute the energy in mel bands of a spectrum.

The Bark scale is a psychoacoustical scale proposed by Eberhard Zwicker in 1961 (Zwicker, 1961). It is named after Heinrich Barkhausen who proposed the first subjective measurements of loudness. The scale ranges from 1 to 24 and corresponds to the first 24 critical bands of hearing. It is related to, but somewhat less popular than, the Mel scale.

MIR Toolbox and Essentia are some of the frameworks implementing algorithms to compute the energy in Bark bands of a spectrum, by summing the power-spectrum in each bark band.

Equivalent rectangular bandwidth (ERB) is a scale from psychoacoustics, which gives an approximation to the bandwidths of the filters in human hearing, using the unrealistic but convenient simplification of modelling the filters as rectangular bandpass filters. Some frameworks such as MIR Toolbox, Essentia or Psysound3 also compute the energy in ERB bands. Other divisions have been proposed, such as (Scheirer, 1998) and (Klapuri, 1999) and are available in the abovementioned frameworks.

Available in: Essentia, MIR Toolbox, PsySound3.

Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs offer a description of the spectral shape of the sound. The frequency bands are positioned logarithmically (on the Mel scale), which approximates the response of the human auditory system more closely than the linearly-spaced frequency bands (S. Davis & Mermelstein, 1980). Then, cepstral coefficients are computed based on the Discrete Cosine Transform of the log magnitude spectrum. Typically, only the first 13 cepstral coefficients are usually returned by audio frameworks. These 13 coefficients are mostly used for speech representation but Tzanetakis states that "the first five coefficients are adequate for music representation" (Tzanetakis, 2002, p. 34).

Other authors have proposed the equivalent of MFCCs but using ERB or Bark bands. The gammatone-frequency cepstral coefficients (GFCC) is the equivalent of MFCCs but using a gammatone filterbank (ERB Bands). They have been proposed has "an auditory-based feature for robust speech recognition" (Shao, Jin, Wang, & Srinivasan, 2009). The bark-frequency cepstrum coefficients (BFCC) use the bark bands and have shown to be useful in studies regarding percussive content (Herrera, Dehamel, & Gouyon, 2003). These two alternatives are available in Essentia.

Available in: MIR Toolbox, Essentia, and Marsyas.

Linear Predictive Coding Coefficients (LPCC)

Linear predictive coding (LPC) and associated reflection coefficients (RC) are used in speech research for representing the spectral envelope of a digital speech signal in compressed form, using to this end information of a linear predictive model (Deng & O'Shaughnessy, 2003). LPCCs represent the cepstral coefficients derived from linear prediction and have been used in a wide range of speech applications, such as speech analysis, encoding and even speech emotion recognition (Ayadi, Kamel, & Karray, 2011).

Available in: Essentia, Marsyas.

LSP: Linear Spectral Pairs

Linear Spectral Pairs (LSP) or line spectral frequencies (LSF) was first introduced by Itakura (1975) as an alternative representation of linear prediction coefficients for transmission over a channel. LSPs have several properties (e.g., smaller sensitivity to quantization noise) that make them superior to direct quantization of LPCs. For this reason, LSPs are very useful in speech recognition and coding (e.g., (Zheng, Song, Li, Yu, & Wu, 1998)).

Available in: Marsyas.

Roughness (Sensory Dissonance)

Sensory dissonance, also known as roughness, is related to the beating phenomenon that occurs whenever a pair of sinusoids are close in frequency (Plomp & Levelt, 1965). In addition, Sethares proposes a method to estimate total roughness by averaging all dissonance estimates across all possible peak pairs of the spectrum (Sethares, 2005). The MIR Toolbox and Essentia implements such algorithms, while PsySound3 implements the algorithm proposed by Daniel et al. (Daniel & Weber, 1997).

Available in: MIR Toolbox, Psysound3, and Essentia.

Spectral and Tonal Dissonance

Psysound3 computes spectral and tonal dissonance features. However few information is available in the documentation. Dissonance measures the harshness or roughness of the acoustic spectrum (Cabrera et al., 2007). The dissonance generally implies a combination of notes that sound harsh or are unpleasant to people when played at the same time. PsySound3 provides two descriptions of acoustic dissonance: "spectral dissonance" which uses all Fourier components, and "tonal dissonance" which uses a peak extraction algorithm before calculating dissonance. In addition, the framework also implements the Hutchinson and Knopoff's (HK) model of acoustic dissonance (Hutchinson & Knopoff, 1978), resulting in a total of four dissonance metrics: tonal dissonance (HK and S) and spectral dissonance (HK and S). The tonal and spectral dissonance measure the dissonance among tonal components and models the degree deviating from the noisiness of the sound, respectively.

Available in: Psysound3.

Irregularity

Irregularity, also known as spectral peaks variability, is, as the name indicates, the degree of variation of the amplitude of successive peaks of the spectrum. The MIR Toolbox implements two distinct approaches to spectral peaks variability calculation.

The default approach is based on (Jensen, 1999), where the irregularity is the sum of the square of the difference in amplitude between adjoining partials.

The second approach is based on (Krimphoff et al., 1994), where the irregularity is the sum of the amplitude minus the mean of the preceding, current and next amplitude.

Available in: MIR Toolbox.

Tristimulus and even/odd-harm

Tristimulus, even-harm and odd-harm are simple spectrum features proposed in (Pollard

& Jansson, 1982), as a timbre equivalent to the color attributes in the vision. The tristimulus feature quantifies the relative energy of partial tones by three parameters that measure the energy ratio of the first partial (tristimulus1), second, third and fourth partials (tristimulus2) and the remaining (tristimulus3).

The even-harm and odd-harm features correspond to the even-harmonic and odd-harmonic energy ratio.

Available in: Essentia.

3.1.7. Expressive Techniques Features

In this section we describe the audio features that capture information primarily related with expressive techniques and its components, as described in Section 2.3.6.

Average Silence Ratio (ASR)

Average Silence Ratio is a feature proposed by Feng et al. (2003), as an estimation for articulation. In this feature a frame is considered as silence if its energy is lower than a defined threshold of the average energy in the one-second time window. ASR is thus the ratio of silence frames for the entire time window. According to the author of the feature, the "lower ASR means fewer silence frames present in musical piece, or legato in articulation, and the higher ASR means more silence frames present in musical piece, or staccato in articulation".

MIR Toolbox implements this as a variation of the low energy rate, corresponding to a RMS without the square-root.

Available in: MIR Toolbox.

3.1.8. Musical Texture Features

To the best of our knowledge, none of the features studied or found in standard audio frameworks are primarily related with musical texture.

3.1.9. Musical Form Features

Extracting musical form and structure information directly from the audio signal is harder when compared to other lower level features (e.g., spectral/timbral statistics), As a result, few computational extractors are available today. Still, some efforts have been made in this direction, with the proposal of specific higher-level strategies that combine

lower level features. One example of this is the automatic segmentation of entire songs in homogeneous parts based on timbral features and estimation of temporal discontinuities (Foote & Cooper, 2003). In this work, the audio is first segmented based on inter-frame spectral similarity. Next, spectral statistics of each segment are computed. These statistics are then used to cluster similar segments together, revealing information about the song overall structure. According to the authors, the automatic segmentation method achieved "good agreement" when compared to manual segmentation into intro, verse, chorus and bridge clusters. Still, only a song, manually segmented by the authors, was tested and the intro (appearing 4 times in the song) was automatically segmented into two different clusters.

Similarity Matrix

As described above, some approaches estimate musical structure based on the similarity between adjacent segments or frames. These similarities are often represented using an inter-frame or inter-segment similarity matrix, showing the differences between all possible pairs of frames from the input audio signal.

The similarity matrix computation uses a specific set of frame statistics (e.g., spectral features) and a distance function, to calculate the proximity between each pair of frames. As an example, MIR Toolbox is able to use MFCCs, key strength, tonal centroid, chromagram and others with one of several distance functions.

Available in: MIR Toolbox.

Novelty Curve

Based on the specific musical characteristics of each segment or frame, obtained for instance with a similarity matrix, a novelty curve can be obtained by comparing the successive frames to estimate temporal changes in the song. In this novelty curve, the probability of transitioning to a different state over time is represented by the curve peaks.

Different approaches exist to build the novelty curve, two of which are available in the MIR Toolbox. The traditional approach is to compare adjacent frames, defined as the diagonal of the similarity matrix, using cross-correlation, to an ideal template of a transition between two different states of a musical structure (ideal Gaussian checkerboard kernel) (Foote, 2000). The reasoning behind this method is that "section transitions are characterized by an abrupt change from one homogeneous acoustical content to another homogeneous acoustical content, the assumption behind the method is that boundaries in an audio signal are visualized in similarity matrices as 2-dimensional checkerboards" (Kaiser & Peeters, 2012). This approach, known as kernel-based approach, has been used in works such as (Foote & Cooper, 2003).

A newer approach has been proposed by Lartillot et al. (2013), claiming to be simpler but more powerful. In this case, the novelty value is computed as a combination of both the temporal scale of the preceding homogeneous part, as well as the amount of contrast that exists just before and after the ending of that segment.

Available in: MIR Toolbox.

3.1.10. High-Level Features

Finally, frameworks such as MIR Toolbox and the Essentia library have started to provide some experimental higher-level features, related with complex concepts such as emotion, genre, danceability, western vs non-western and others. Most, if not all, of these are actually predictors, combining classification algorithms and previously gathered data to try to label the source audio files into a fixed set of tags. A brief summary of these predictors is listed below.

Emotion

MIR Toolbox attempts to extract (predict) an emotion descriptor based on the analysis of the audio and musical contents of a given recording. The output is given in two distinct paradigms:

1. A categorical paradigm of 5 classes: happy, sad, tender, angry and fearful[59], outputting a value of 1 to 7 for each class.
2. A 3-dimensional space composed of activity (energetic arousal), valence (pleasure-displeasure continuum) and tension (tense arousal).

The classification process is based on previous work by Eerola et al. (2009) and uses multiple linear regression with the 5 best performing predictors identified.

These 5 predictors contributing to each of the 5 classes are:

- Happy: Maximum value of summarized fluctuation, Spectral spread averaged along frames, Standard deviation of the position of the maximum of the unwrapped chromagram, Key clarity (2nd output of *mirkey* function) averaged along frames, Mode averaged along frames.
- Sad: Spectral spread averaged along frames, Standard deviation of the position of the maximum of the unwrapped chromagram, Mode averaged along frames, Averaged HCDF, Averaged novelty from wrapped chromagram.
- Tender: Spectral centroid averaged along frames, Standard deviation of roughness, Key clarity (2nd output of *mirkey*) averaged along frames, Averaged HCDF, Averaged spectral novelty.

[59] Some of the classes were changed from nouns to adjectives to keep consistency.

- Angry: Roughness averaged along frames, Key clarity (2nd output of *mirkey*) averaged along frames, Entropy of the smoothed and collapsed spectrogram, averaged along frames, Averaged novelty from unwrapped chromagram.
- Fearful: Standard deviation of RMS along frames, Averaged attack time, Maximum value of summarized fluctuation, Key clarity (2nd output of *mirkey*) averaged along frames, Mode averaged along frames.

These 5 predictors contributing to each of the 3 dimensions are:

- Activity: RMS averaged along frames, Maximum value of summarized fluctuation, Spectral centroid averaged along frames, Spectral spread averaged along frames, Entropy of the smoothed and collapsed spectrogram, averaged along frames.
- Valence: Standard deviation of RMS along frames, Maximum value of summarized fluctuation, Key clarity (2nd output of *mirkey*) averaged along frames, Mode averaged along frames, Averaged spectral novelty.
- Tension: Standard deviation of RMS along frames, Maximum value of summarized fluctuation, Key clarity (2nd output of *mirkey* function) averaged along frames, Averaged HCDF, Averaged novelty from unwrapped chromagram.

The authors tested their models using a dataset of 110 15-sec clips extracted from soundtracks, reporting a dimensional prediction rate between 0.64 (valence) and 0.75 (activity) for the machine learning algorithm included in the toolbox (multiple linear regression). Regarding the basic emotions, the reported prediction rate varied between 0.38 (sad, tender) and 0.55 (fearful[60]) (Eerola et al., 2009).

Given its reliance on previously established weights, this extractor is only reliable in the MIR Toolbox version (v1.3) where it was initially "calibrated", with the newer versions outputting "distorted results" (Lartillot, 2018, p. 196).

The Essentia library implements a similar feature, classifying songs in 7 classes, 4 of them distinct emotions: happy, sad, aggressive, relaxed, acoustic, electronic, party. To this end, it contains already trained models and requires the Gaia library, described previously in Section 3.1.1, to apply similarity measures and classifications on the results of the audio analysis.

Available in: MIR Toolbox, Essentia.

[60] The authors used "scared" in the paper, but changed it to "fearful" in MIR Toolbox.

Genre, Danceability and Other Classifier Based Extractors

In a similar way to the emotion descriptor extractor (or predictor), Essentia library also includes Gaia trained models for:

- musical genre (using 4 different databases)
- ballroom music classification
- western / non-western music
- tonal / atonal
- danceability
- voice / instrumental
- gender (male / female singer)
- timbre classification

These musical descriptors work as a typical classification problem, by extracting a particular set of features from the source audio signals and feeding them to classification models trained with them in other datasets.

Available in: Essentia.

Genre-based Emotion Classifiers

In his PhD thesis, Laurier studied the problem of music emotion classification using audio signals, as well as multi-modal approaches using lyrics (Laurier, 2011). There, the author also studied the relations between genre and emotion in AllMusic data (i.e., genre-emotion pairs), and observed that some emotions were frequently associated with specific genres. This led to the idea of using automatic genre classification to improve his previous emotion classification results.

To this end, a dataset from iTunes was divided by genre into 15 classes or genres. The best performing audio emotion classification models (one per emotion) were then used to classify all the songs of this dataset. As a result, each audio file was associated with a genre annotation (from iTunes) and a set of emotion tags (obtained by classification).

By statistically analyzing the results, i.e., how frequently an emotion appears in songs of a specific genre, the author confirmed "a clear association between mood and genre, with quite intuitive and logical results". Building on this, a genre-based emotion classifier was proposed, as illustrated in Figure 3.2.

In brief, two classification models use the same low-level audio features, one to predict emotions based on these features, and a second approach which first predicts genre and then uses the genre descriptors to predict emotions. Each outputs a probability estimate, which are combined using a decision function.

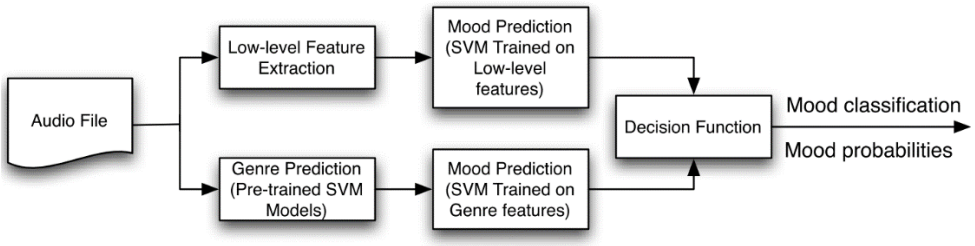


Figure 3.2: Schema of the genre-based emotion classifier proposed by Laurier (2011, p. 110).

To predict genre, four genre datasets containing a total of 32 genre tags were used to train different binary classification models, which together resulted in 32 genre models (Laurier, 2011, p. 111). Each of these classifiers output one probability estimate that represents the probability (between 0 and 1) of an audio file being of that specific genre. These 32 values are fed as genre descriptors to emotion classification models following a similar strategy for four basic emotions: happy, sad, angry and relaxed. The final emotion classification and probabilities are given by a simple weighted sum between the two probability predictions (genre-based and standard low-level based approaches).

The experimental results show statistically significant improvements in emotion classification accuracy for each of the 4 classes used, ranging from 88.51% (happy) to 99.12% (angry). Despite the very high results, it is important to note that the data is from binary classification (e.g., happy vs not happy).

One of the most interesting findings is that genre information, even if automatically extracted was relevant to MER. More importantly, genre-features were computed from the same low-level features previously used on emotion. However, they "contain expert knowledge" adding information related to "human categorization" and thus the author concludes by hypothesizing that "the more high level features we add, the better we will be able to automatically classify music in general" (Laurier, 2011, p. 114).

Danceability Estimation

As opposed to the aforementioned danceability extractor built as a pre-trained classification model, Streich et al. (2005) proposed using a low-level audio feature derived from Detrended Fluctuation Analysis (DFA) to characterize audio signals in terms of its danceability. DFA was originally proposed to be used on biomedical data (Peng et al., 1994) and consists on fractal analysis techniques to reveal correlations within data series across different time scales. Essentia implements the algorithm described in (Streich & Herrera, 2005), outputting the danceability of the audio signal in a range from 0 to 3

(higher values meaning more danceable).

Available in: Essentia.

Dynamic Complexity

In his PhD thesis, Sebastian Streich studied the automated estimation of the complexity of music based on the musical audio signal, proposing a set of complexity descriptors (Streich, 2007). The proposed algorithms focus on aspects of acoustics, rhythm, timbre, and tonality.

The Essentia library implements an extractor to estimate dynamic complexity, or whether a song contains a high dynamic range. This descriptor consists in the "average absolute deviation from the global loudness level estimate on the dB scale".[61] This extractor is related to the dynamic range and to the amount of fluctuation in loudness present in the source audio signal.

Available in: Essentia.

To conclude this section, Table 3.1 presents the number of described features per musical dimension as well as the number of these features used in our baseline model described in Chapter 4.

Musical dimension	*Number of features*	*Percentage of total*	*Features used*
Melody	5	7.9%	3
Harmony	10	15.9%	8
Rhythm	13	20.6%	10
Dynamics	8	12.7%	7
Tone Color	24	38.1%	21
Expressive Techniques	1	1.6%	1
Musical Texture	0	0.0%	0
Musical Form	2	3.2%	2
Total	*63*	*100.0%*	*52*

Table 3.1: Number of audio descriptors reviewed per musical dimension.

Please note that some of the abovementioned audio features output several metrics

[61] http://essentia.upf.edu/documentation/reference/std_DynamicComplexity.html

or generate time series of data that are later summarized, increasing the final number of descriptors. This is especially true for tone color features, where some features divide the audio signal in bands and output time-series data (e.g., MFCCs). Based on this table, we conclude that the number of available audio features is very unbalanced across musical dimensions. In fact, musical texture, expressive techniques and musical form are especially lacking, in contrast to tone color, which is, by far, the most represented category.

In our experiments, detailed in Chapter 4, the baseline model used audio features from Marsyas, MIR Toolbox and Psysound3. These audio frameworks were selected due to the high number of feature extractors provided, covering most of the described features, and because they have been extensively used in previous studies, e.g., (Aljanaki, 2016; S.-H. Chen, Lee, Hsieh, & Wang, 2015; Eerola et al., 2009; Malheiro, Panda, Gomes, & Paiva, 2016a; Y.-H. Yang, Lin, Su, et al., 2008). The remaining frameworks do not implement additional features from the dimensions identified as lacking (expressive techniques, musical texture and musical form). Of these, Essentia provides some features that are not available in the others, but was not considered since it was unknown to us at the time. We plan to extend our experiments with the few missing features in the future.

As Table 3.1 shows, the majority of the features available were used in our baseline (standard features) experiments. The few features left out were not considered for several reasons, from the lack of documentation and working extractors (e.g., pitch content in Marsyas), unavailability in the selected frameworks (e.g., spectral contrast), because they were considered an intermediate or visualization step (e.g., chromagram correlation map), or an alternative approach to capture already extracted information (e.g., beat location, which output tempo or events location).

A summary of the described features is provided in Table 3.2. The features used in our baseline experiments are marked with †.

MD	*Feature*	*Description*
Melody	Pitch†	Estimates the F0 of a sound.
	Virtual Pitch Features†	Set of features related with psychoacoustics and modelation of the perceived pitch.
	Pitch Salience†	A measure of how noticeable (that is, strongly marked) is the pitch in a sound.
	Predominant Melody F0	Estimates the F0 of the predominant melody.
	Pitch Content	Information about the pitch (e.g., dominant pitch

		class, its octave range).
Harmony	Inharmonicity†	Amount of partials that are not multiple of the fundamental frequency (F0).
	Chromagram†	Energy distribution along pitches.
	Tuning Frequency	Estimation of the exact frequency (in Hz) on which a song is tuned.
	Key Strength†	Probability of each key candidate to be the key of a given song.
	Key and Key Clarity†	Estimated tonal center positions and their respective clarity.
	Modality†	Major or minor mode estimation.
	Chord Sequence	Best matching major or minor triad.
	Tonal Centroid Vector†	6-D tonal centroid from chromagram.
	Harmonic Change Detection Function†	Flux of the tonal centroid, the distance between harmonic regions of successive frames.
	Sharpness†	Rates sound from dull to sharp.
Rhythm	Rhythmic Fluctuation†	Rhythmic periodicity along auditory channels, estimates rhythm content.
	Beat Spectrum†	Measure of acoustic self-similarity.
	Beat Location	Beat tracking, detects beat times and tempo.
	Onset Time†	Estimated starting time of the notes.
	Event Density†	Estimated note onsets per second.
	Average Duration of Events†	Average duration from attack to release.
	Tempo†	Estimated tempo of a musical piece.
	Tempo Change†	Tempo variations over time.
	Metrical Structure†	Hierarchical metrical structure information.
	Metrical Centroid and Strength†	Assessment of metrical activity and pulsation strength / clarity.
	Pulse / Rhythmic Clarity†	Strength of the estimated beats.
	PLP Novelty Curves	Mid-level rhythmic representation that may be used to estimate tempo or track beats.
	HWPS	Rhythmic content information calculated from the

		beat histograms.
Dynamics	RMS Energy†	The global energy of the signal.
	Low Energy Rate†	Percentage of frames showing less-than-average energy.
	Sound Level†	Unweighted sound pressure level of the signal.
	Instantaneous Level, Frequency and Phase†	Sound pressure estimate and others by applying Hilbert transform of the audio waveform.
	Loudness†	Subjective impression of the intensity of a sound.
	Timbral Width†	The width of the peak of the specific loudness spectrum.
	Volume†	Refers to the "size" or intensity of the sound.
	Sound Balance	How the sound is "balanced" (e.g., crescendo estimation).
Tone Color	Attack/Decay Time†	Temporal duration of the attack/decay phases.
	Attack/Decay Slope†	Gradient of the attack/decay phases.
	Attack/Decay Leap†	Attack/decay phase amplitude.
	Zero Crossing Rate†	Waveform sign-change rate.
	Spectral Flux†	Distance between successive spectral frames.
	Spectral Centroid†	1st moment (mean): indicates brightness of the sound.
	Spectral Spread†	2nd moment (variance): measures the dispersion of the spectrum.
	Spectral Skewness†	3rd moment: symmetry of the spectrum.
	Spectral Kurtosis†	4th moment: "peakedness" of the data.
	Spectral Flatness†	Smooth/spikyness of data.
	SCF†	Spectral crest factor, a measure of the "peakiness" of the spectrum.
	Spectral Contrast	The spectral contrast of a spectrum, represents the relative spectral distribution.
	Spectral Entropy†	Shannon entropy of the signal.
	Spectral Rolloff†	Measures the amount of high-frequency energy in the signal.
	High-frequency Energy†	Other metric for the amount of high-frequency energy in the signal.

	Cepstrum (Real/Complex) †	Inverse Fourier transform of the log of the spectrum.
	Energy in Mel/Bark/ERB Bands	Energy in bands using Mel or other perceptual scales of pitches.
	MFCCs†	Mel-frequency Cepstral Coefficients – measure of spectral shape.
	LPCCs†	Linear predictive coding coefficients, represent the signal envelope in compressed.
	LSP†	Linear spectral pairs.
	Roughness†	Estimation of the sensory dissonance (using the peaks of the spectrum).
	Spectral and Tonal Dissonance†	Harshness among tonal components.
	Irregularity†	Successive spectral peaks variability.
	Tristimulus and even/odd-harm	Timbre descriptors, quantifies the relative energy of partial tones and even/odd harmonics.
ET	ASR†	Average silence ratio, can be used as an assessment of articulation.
MF	Similarity Matrix†	Similarity between all possible pairs of frames.
	Novelty Curve†	Estimates temporal changes by the comparing successive frames of the similarity matrix.
High-Level	Emotion	Some audio frameworks provide high-level descriptors based on lower level features and previously trained machine learning models.
	Genre	
	Danceability	
	Genre-based emotion	Emotion prediction using genre as a feature.
	Dynamic Complexity	Musical complexity descriptors.

Table 3.2: Summary of standard computational audio features (MD: Musical Dimensions, ET: Expressive Techniques, MF: Musical Form).

3.2. Music Emotion Recognition Approaches

Music emotion recognition (MER) is a relatively recent and promising research area, part of the broader music information retrieval field (MIR). The origin of MIR is linked to the necessity to manage massive collections of digital music for "preservation, access,

research and other uses" (Futrelle & Downie, 2003) and the importance that emotional content as to this end.

Several authors have dedicated their efforts to MER in the last decades, proposing different approaches to explore the emotional content of audio music. Most of these studies can be generally described as typical (supervised) machine learning problems. For this reason, before delving deeper into the specificities of particular works, we offer a general overview of the typical approach that ties most of these solutions together.

The typical machine learning approach applied in MER has three distinct parts, as illustrated in Figure 3.3:

1. Collection of ground truth data;
2. Feature extraction;
3. Classification (training and testing).

In brief, the process consists in gathering a set of songs and respective labels that best describe their emotional content. From the audio clips, representative audio features are extracted (e.g., estimated beats per minute, etc.). Finally, these features and the emotion labels are passed to supervised machine learning algorithms, creating classification models that label new songs based on their features.

Some authors have also explored unsupervised strategies, either using only audio data (e.g., with hierarchical clustering (Bartoszewski, Kwasnicka, Markowska-Kaczmar, & Myszkowski, 2008)) or only audio annotations (e.g., k-means clustering of arousal and valence values (J. Kim, Lee, Kim, & Yoo, 2011)) in the training process. Still, these approaches proved to be less reliable in MER and thus have not been frequently adopted.

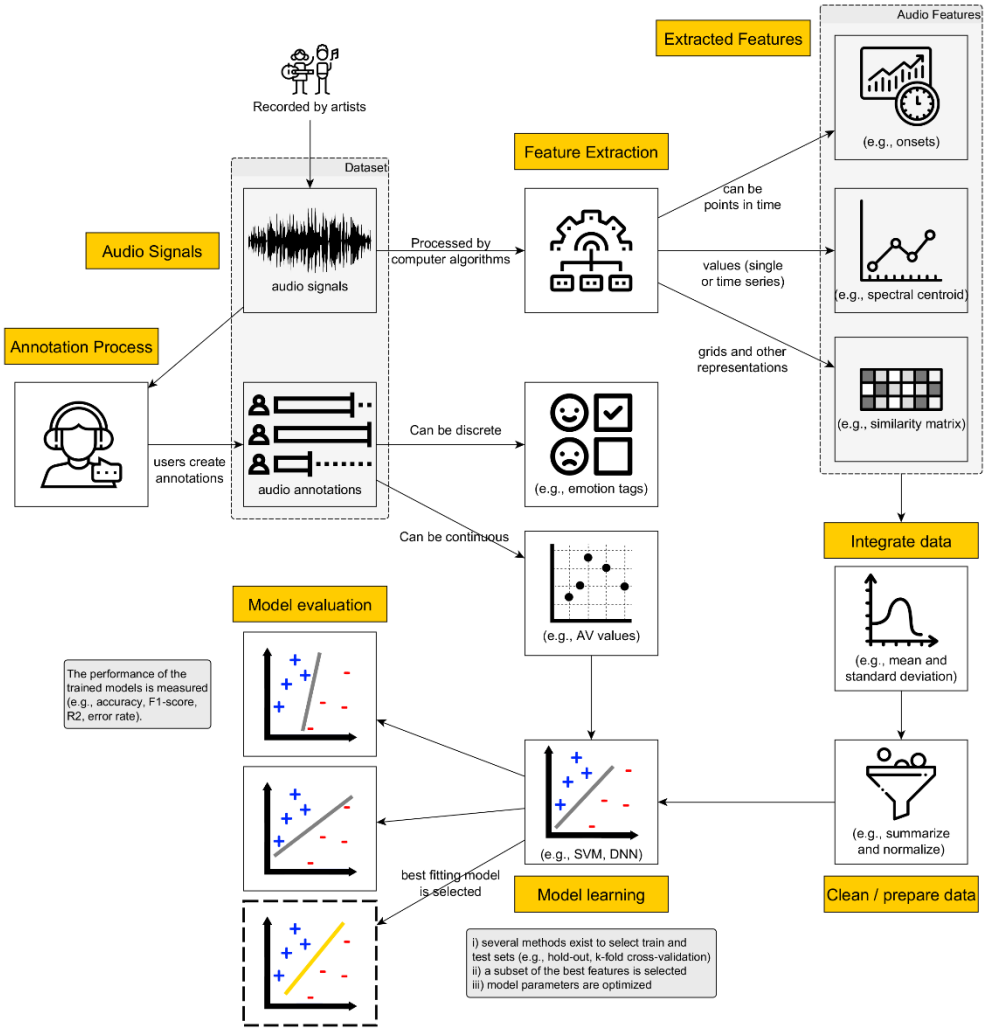


Figure 3.3: Typical supervised machine learning strategy applied in MER studies. In this example, a dimensional model is being used.[62]

3.2.1. Ground-truth Collection and Verification

The starting point of any MER study is the collection of ground-truth data, which normally consists in the annotation and dataset creation processes. This ground-truth data is said to be "the main factor that affects the results of MER, and the methods of extracting and selecting ground-truth data are the key for reducing the subjectivity of the

[62] The icons used were made by Freepik from www.flaticon.com.

results" (X. Yang et al., 2017).

Song Excerpt Selection

The first step into dataset creation is to select a set of songs (digital audio files) to be annotated. These audio files must be preprocessed, which consists in two distinct steps: data format conversion and song segmentation. Data conversion is used to homogenize the clips, converting them into a format that normally makes feature extraction less resource intensive. For this reason, this step is many times carried out only during the feature extraction process. Song segmentation consists in extracting a small segment of the original full song and poses the first problem in dataset creation. The rationale behind this is that, although full songs can be used, their emotional content fluctuates over its course and, logically, using smaller, more homogeneous clips will improve results.

The segmentation process itself is still not agreed upon, and so various studies use different lengths and selection methods. MacDorman et al. (2007) stated that the segments should be as short as possible to have a more homogeneous emotion and, thus, more consistent annotations, also enabling a more granular analysis. However, segments that are too short may be detached from its surrounding environment, "resulting in a lack of evaluation of their ecological validity" (X. Yang et al., 2017). The typical segment length applied in pop music MER studies is between 25 and 30 seconds (X. Yang et al., 2017), which is consistent with the typical duration of the chorus part of such (i.e., pop) songs (J.-C. Wang, Yang, Wang, & Jeng, 2012).

The optimal length of musical segments in classical music has been further studied by Xiao et al. (2008). To this end, 60 musical pieces were divided into segments of different lengths – 4, 8, 16 and 32 seconds - and annotated by two subjects. The classification results showed that the best performing models were based on 8 and 16-seconds segments.

It can be argued that, at least empirically, having fixed length segments might create clips with emotion fluctuations in it and that a more flexible approach (e.g., length interval) would be more reasonable. Some authors have also used segmentation methods based on other units such as musical structure, second (Soleymani, Caro, Schmidt, Sha, & Yang, 2013), lyrics (B. Wu, Zhong, Horner, & Yang, 2014) and frequency (L. Lu et al., 2006).

In addition to segment length, the segment position in the original song is of major importance. Ideally, the segments should be extracted from a representative and homogeneous part of the song. Still, this process is prone to subjectivity and influenced by the musical experience of the one selecting it. Many authors do not provide information about the criteria used to select the clips in their studies. Moreover, many rely on audio samples of online services, such as Last.FM and AllMusic, which in many cases have

non-representative segments (e.g., the first *n* seconds, or a randomly selected section).

Annotation Types

Having the segmented audio clips, the next step is to collect the ground-truth, which are annotations given by human subjects to the content of each clip. The annotations can be split into three major types: 1) categorical, where subjects select adjectives; 2) dimensional, where subjects use numerical values; and 3) based on musical features, a less used approach where emotion annotations are generated based on specific musical characteristics (e.g., Feng et al. (2003) established emotion annotation rules based on tempo and articulation cues previously found in (Juslin, 2000)).

The categorical annotations are created using a survey, where subjects select one or various adjectives that best describe the perceived emotion. They follow a specific categorical taxonomy of emotion, such as the ones described in Section 2.2.1. The main advantages are: it is easier to use by inexperienced volunteers, as the usage of adjectives with clear emotional significance are consistent with humans' subjective feelings; and lower computational complexity in the later machine learning stage when the number of categories is low. On the other hand, such approach suffers from the same problems of categorical emotion models: strong subjectivity; no distinction between songs with the same adjectives; difficulties in selecting the list of adjectives; and the accuracy of such systems is usually restricted by the number of these adjectives (Li & Ogihara, 2003).

Dimensional annotations require human subjects to rate audio clips using numerical values according to the selected dimensional taxonomy (e.g., select arousal and valence values between -1 and 1). This process requires a much higher level of knowledge and time from the subjects as the models are complex and users have difficulties translating the identified emotions to a point in space. Still, dimensional annotations are of increased usefulness for most MER applications.

A third type of annotations has been used sparely in MER works and consists in requiring subjects to annotate the perceived musical characteristics and subsequently generates the emotion annotations based on the correlations between musical features and emotions. A prime example is the work by Feng et al. (2003), which built on Juslin findings (2000). In it, Juslin asked performers to play specific musical pieces trying to express distinct emotions. Then, participants rated these recordings in terms of emotions. By correlating the data, the author identified two musical characteristics (tempo and articulation) as responsible for the emotional transfer from performers to listeners and derived a set of rules. Feng used these rules, which relate tempo (high or low) and articulation (staccato or legato) to emotions (i.e., happy, sad, angry, fearful), to generate emotional annotations.

Even though this method does not require participants to rate emotions directly, and thus "avoids a certain degree of subjectivity" (X. Yang et al., 2017), it introduces

new problems related with the annotation of musical features. Moreover, it is tied to fixed sets of audio features that have been preselected and identified as relevant and for this reason it has rarely been adopted.

Annotation Collecting and Filtering

After segmenting the audio files and selecting the annotation type to be used, the next step is to collect the annotations. As discussed, the annotation collection process is complex and prone to errors that may compromise the ground-truth quality. To leverage this, common strategies are to: collect multiple annotations for each audio clip to decrease bias and assess subjectivity; remove low quality annotations, which can have many causes, from software errors to subjects not understanding the requested task (Raykar et al., 2010); reduce the number of annotations requested per subject and allow skipping of segments to avoid fatigue, bias and random annotations on unclear segments; increase the usability of the survey software, e.g., by using gamification strategies (e.g., MoodSwings (Y. E. Kim, Schmidt, & Emelle, 2008) and TagATune (Law, Ahn, Dannenberg, & Crawford, 2007)) to increase the interest of subjects.

Given the amount of resources needed to fulfil such a task in a controlled lab, some researchers have explored alternative ideas, such as using online crowdsourcing services as the Amazon Mechanical Turk (Speck, Schmidt, Morton, & Kim, 2011), using data from online services and databases providing samples and metadata (Hu, Downie, Laurier, Bay, & Ehmann, 2008) or data from social media platforms such as Last.FM (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011).

Although these methods decrease immensely the resources needed, they generally increase annotations' noise and some additional steps are needed to filter the data. Some of the common strategies are to: force subjects to pass a preliminary evaluation to verify that the problem was correctly understood or have some control test cases among the data to be annotated and eliminate annotations from annotators that failed these cases (Soleymani et al., 2013); have a lighter second annotation process, serving to validate the gathered online annotations (Panda, Malheiro, & Paiva, 2018); or use statistical noise filtering techniques to eliminate outliers, such as proposed by Speck et al. (2011).

One last particularity with MER datasets and probably its major problem is that, contrary to other information retrieval fields, the music files are usually under very restrictive copyright laws. This issue limits the public distribution of datasets to their annotations and extracted features and is one of the major problems in MER research. Although some authors assume that audio samples under 30 seconds can be shared as "fair use" without copyright obligation, the subject is complex[63] and many datasets remain private. Due to this fact, numerous studies had to invest limited resources to build

[63] https://smallbusiness.chron.com/copyright-laws-30-seconds-music-61149.html

and use private datasets, making it harder to compare to their peers work.

3.2.2. Audio Datasets for MER

Researchers have been addressing MER for decades and, even if in many studies private datasets and ground-truth have been used, for reasons previously explained, there is a raising awareness that dataset quality and the replication and comparison of studies is an issue in MER. For this reason, some authors have dedicated efforts to produce public datasets or make available their data. In the next paragraphs we analyze these and briefly discuss their merits and limitations.

MIREX Audio Music Classification (AMC) Task Dataset – 2007

Although not being a public dataset, the MIREX AMC Task dataset is a landmark in MER research. Its origin is associated to the comparison and replicability problem in MER studies, caused by researchers being forced to build and use private datasets. To mitigate this, a music emotion classification task was proposed to the Music Information Retrieval Evaluation eXchange (MIREX) in 2007. Nowadays, the task allows researchers to submit music emotion classification algorithms, which are tested against the private dataset.

The dataset consists of a collection of 600 30-second audio clips in 22.05 kHz mono WAV format and labeled by human judges. The employed taxonomy was derived from data, as previously described in detail in Section 2.2.1 and is divided into five clusters with a total of 29 adjectives (see Table 2.2).

Being private, the dataset is only available to the MIREX task leaders, who evaluate the algorithms submitted during the annual contest and thus cannot be freely used by researchers. Moreover, several limitations have been identified, which have been extensively described in Section 2.2.1. Namely:

1. The emotion model lacks support from psychology;
2. It contains a semantic overlap (ambiguity) between clusters 2 and 4;
3. It contains an acoustic overlap (based on analysis of MIREX dataset) between clusters 1 and 5;
4. Our experiments in Section 5.1.2 suggest a higher inter-cluster similarity than expected.

Computer Audition Lab 500 (CAL500) dataset – 2007

The CAL500[64] is a 500 popular music songs dataset, each of which has been labeled with multiple adjectives by at least 3 subjects (Turnbull, Barrington, Torres, & Lanckriet, 2007). Each song in the collection was recorded by a different Western artist. The annotation process uses a categorical model, consisting of a set of 174 labels, organized into six categories: emotion, genre, instrument, song, usage and vocal. The tags are said to be "manually generated under controlled experimental conditions" (Sanden & Zhang, 2011) and therefore believed to be of high quality. This process was performed by 66 paid undergraduate students, with headphones in a laboratory.

The annotation process started with 135 tags, 18 of which were related to emotion, found by Skowronek et al. (2006) to be both important and easy to identify. These emotion-related concepts were rated by 66 undergraduate students on a scale from one to three (e.g., "happy" could be "not happy", "neutral" and "happy"). The 135 concepts were later transformed in 237 binary classes by transforming all the ternary concepts into two individual labels (e.g., the "happy" concept rated with -1, 0 or 1, was transformed into two binary classes "not happy" and "happy", which could be true or false). The final annotation vector is obtained by calculating the level of agreement of subjects (i.e., semantic weights vector). In addition, binary labels are also created for labels with agreement of 80% or higher. The final dataset contains 159 words, of which 36 are related to emotion.

As with other datasets, CAL500 full audio clips were not directly available at the authors' website. Some other problems have been reported[65], namely the fact that it contains 502 tracks and 503 song ids instead of the 500 described in the paper and that some of the audio clips are corrupt. While the original full dataset is unavailable, the features can be found in the Million Song additional datasets[66] and GitHub[67].

Yang Arousal and Valence (YangAV) Dataset – 2008

The YangAV dataset[68] is the dataset used by Yang et al. (2008) in their work, one of the first to explore MER as a dimensional model using a regression approach. The dataset contains 25-second clips of 195 popular songs from both Western, Chinese and Japanese albums.

The 195 songs were selected with two specific criteria: 1) they should be uniformly distributed in each quadrant of the AV emotion plane; 2) each sample should express a

[64] http://calab1.ucsd.edu/~datasets/cal500/

[65] https://highnoongmt.wordpress.com/2013/03/07/using_the_cal500_dataset/

[66] https://labrosa.ee.columbia.edu/millionsong/pages/additional-datasets

[67] https://github.com/yzhaobk/CAL500

[68] http://mac.citi.sinica.edu.tw/~yang/MER/taslp08/

certain dominant emotion. The clips were segmented manually, mainly with the chorus part, to be both representative of the song and containing a dominant emotion.

The annotation process was conducted with 253 subjects, mostly college students, asked to label "the emotion based on their feelings of what the music sample is trying to evoke" (Y.-H. Yang, Lin, Su, et al., 2008). Ten random samples were assigned to each subject, who annotated them using AV values from –1.0 to 1.0 in a total of 11 uniformly spaced ordinal levels. The subjects were also instructed to consider melody, lyrics, and singing (vocal) of the song. The quality of the collected ground-truth was assessed by calculating the standard deviation of annotations from different subjects to the same clips, where values between 0.2 and 0.4 were observed. In addition, a test-retest reliability study was conducted by asking 22 subjects to re-label their songs after two months. Since more than half of new annotations had an absolute difference to the original ones below 0.1, the authors concluded that the results were replicable and thus quality was acceptable (although absolute differences between 0.2 and 0.6 were observed). The final annotations were obtained by averaging individual annotations.

Although the dataset published online contains only metadata and features, the audio clips are available upon request. After analyzing the data, a few problems were observed. First, from the 195 songs, only 194 clips were provided and the online metadata is only available for 193 clips. By matching the AV annotations provided in the audio clips package with the annotations present in the provided feature data, we verified that the songs were in different order and five did not match, reducing our possible set to 189 clips and features.

Two other issues were found, related with the ground truth. First, although both Russell and Thayer studies place the emotions far from the center of the emotion space, the majority of the clips in this dataset are found close to the center. A possible explanation for this is the disparity between annotations of the same song, which when averaged end up close to the center. Secondly, despite the authors trying to have a balanced initial set, the final annotations created by the subjects originated a very unbalanced dataset. As an example, the second quadrant contains only 12% of the songs. For an extensive analysis on this see Section 5.1.1.

MoodSwings dataset (MTurk240) – 2008

The MoodSwings dataset[69] is another collaborative dataset originally created using an online game with a purpose (GWAP) (Y. E. Kim et al., 2008). In this study, the authors used 240 15-second clips of US pop music to collect arousal and valence values in a per-second frequency. To this end, the game paired two anonymous online players, who used the mouse to annotate the music segment over time in the AV space. The score of

[69] http://music.ece.drexel.edu/research/emotion/moodswingsturk

the players was defined by the overlap between their cursors, encouraging consensus in annotations.

Subsequently, the authors decided to redesign the experiment to address problems found after analyzing the obtained labels (Speck et al., 2011). Here, Amazon Mechanical Turk (MTurk) workers were payed to use a modification of the system where the players were not paired up, playing solo, and no points were awarded. The workers used the same 240 clips and, although they were extended to 30-seconds for listening, only 15 seconds were considered for annotation. In this case, two of the clips were used as tests to eliminate low quality workers.

Soundtrack dataset – 2011

The soundtrack dataset[70] was the byproduct of a study by Eerola et al. (2011) about the comparison of the two main approaches to the definition of emotion taxonomies: discrete and dimensional models. To overcome the possible bias of previous studies, which used mainly random selected excerpts from very well-known western music, the authors decided to use less known clips extracted from 60 movie soundtracks from 1958 to 2006.

A rigorous song selection and segmentation process was carefully planned and executed by 12 expert musicologists with at least 10 years of musical study. Each expert was given five soundtracks and asked to find five examples of six target emotions. Half of the subjects focused on discrete emotions (happiness, sadness, fear, anger, surprise and tenderness), while the remaining half focused on continuous emotions, specifically the six extremes of the 3-dimensional model defined by valence, energetic arousal and tense arousal (positive and negative valence, high and low tension arousal, as well as high and low energy arousal). To help the experts, three adjectives were associated to the extremes of the dimensional models (valence: pleasant–unpleasant, good–bad, positive–negative; energy arousal: awake–sleepy, wakeful–tired alert–drowsy; tense arousal: tense–relaxed, clutched up–calm, jittery–at rest). Additional criteria were defined to ensure a good selection of excerpts, i.e.,: each clip should have 10 to 30 seconds; it should not contain lyrics, dialogue or sound effects (e.g., car sounds, shootings); familiar scenes should be avoided; it should convey the target emotion in an optimal way. The musical features and devices that influenced its choice should also be annotated.

From the segmentation process, a total of 360 audio clips were created. To evaluate the segments, a pilot experiment was run where each expert was asked to rate 90 clips in terms of perceived emotion and familiarity. This process was carried out as a classroom experiment using high-quality equipment in laboratory conditions (Eerola & Vuoskoski, 2011). The authors presented a comprehensive analysis on the results. One

[70] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/past-projects/coe/materials/emotion/soundtracks/Index

of the most relevant is the consistency between raters, measured with the Cronbach's alpha (Cronbach, 1951), where the lowest score was 0.66 for surprise[71], while the remaining obtained values between 0.89 and 0.93. For this reason, the authors decided to remove surprise from the second experiment.

Based on the results of the pilot test, the authors selected a subset of 110 clips (50 discrete + 60 dimensional). These clips were then rated in a scale of 1 to 9 by 116 university students (later reduced to 110), aged 18 to 42 (68% females) with different musical knowledge. Before the annotation, several concepts about perceived versus induced emotion, the emotional models as well as related music examples and a practice session were given to the subjects. The experiment was conducted under very controlled conditions, where each subject rated at his or her own pace, using studio quality headphones in a sound proof room.

The soundtrack dataset is arguably the dataset with the most carefully planned and executed ground truth acquisition process. Unfortunately, such approach is extremely resource intensive and is the cause of its obvious weakness – its very limited size, with only 110 clips. Additionally, using movie soundtracks may also be considered a limitation, since such clips can be seen as a unique, very specific genre.

Million Song Dataset – 2011

At the opposite end of the spectrum to the soundtrack dataset, lies the Million Song Dataset. It consists of a massive, freely-available, collection of audio features and metadata for a million contemporary popular music tracks (Bertin-Mahieux et al., 2011). The main goal behind its creation was to encourage research on MIR solutions that scale to commercial sized databases. The core data is provided by The Echo Nest[72], a music intelligence and data platform for developers and media companies that is now owned by Spotify.

Although no audio clips are provided with the dataset, some scripts are available[73], which can be used to download the song preview from online services. The data provided consists of 54 fields per song, of which 25 include metadata such as title, release, duration, year, the ID of the song in several online services (7digital.com, The Echo Nest, musicbrainz.org, playme.com) and various details about the artist. The provided features complete the remaining fields, and include descriptors about tempo, bars, beats, energy, key, loudness, mode, timbre (such as MFCCs) and others. In addition to the

[71] There are different reports about the acceptable values of alpha, ranging from 0.70 to 0.9 (Tavakol & Dennick, 2011).

[72] http://the.echonest.com/

[73] https://github.com/tbertinmahieux/MSongsDB/tree/master/Tasks_Demos/Preview7digital

base data, there are also complementary sets for subsets of the million song dataset contributed by the community. These include lyrics from musiXmatch, song-level tags and similarities from Last.FM, and others. Some of the tags are related to emotions (e.g., love, happy, sad) and have been used in MER studies (Corona & O'Mahony, 2015).

The main issues with this dataset are related with the absence of audio samples and the nature of the metadata. First, no audio clips are available and, although this can be mitigated by downloading the songs' previews from online services, such services are restricted with an API key and might limit the accesses per day. In addition, there is no guarantee that the preview excerpts obtained can be used to replicate the provided features and the authors' even state that accessing the same Nest API used to create the dataset may even result in different data[74]. In addition, there is also no guarantee that the previews match the excerpts used by the listeners who generated the tags or match excerpts used in the future by other researchers. Moreover, the data that might be useful for MER research comes from Last.FM tags. Such tags are generated by users, which may tag a song with any random word under any circumstance. For illustration, the dataset contains 522,366 unique tags, where some of the most used are "rock", "pop", "favorites" and "00s"[75].

DEAP120 dataset – 2012

The DEAP120 dataset[76], an abbreviation for Dataset for Emotion Analysis using Physiological and Audiovisual Signals (DEAP), is a multi-modal dataset for the analysis of human affective states which includes video segments and the physiological signals of the subjects (Koelstra et al., 2012).

The dataset contains 120 one-minute segments of music videos, segmented with an affective highlighting algorithm proposed by the authors. From the 120 original videos, 60 were manually selected, while the remaining 60 were selected via Last.FM affective tags. The 120 segments were then rated in terms of arousal, valence and dominance (using a discrete scale of 1 to 9) by 14-16 volunteers each, using an online platform.

Using the results of the online self-assessment, a subset of the 40 most appropriate videos was selected, based on the videos with the "clearest responses". This subset was then used by 32 volunteers in an experiment where their electroencephalogram (EEG) and peripheral physiological signals were recorded as they watched them. The frontal face video of 22 participants was also recorded. In addition, they also rated the 40 videos

[74] https://labrosa.ee.columbia.edu/millionsong/pages/will-i-get-back-same-data-if-i-access-echo-nest-api

[75] https://labrosa.ee.columbia.edu/millionsong/lastfm

[76] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

regarding their arousal, valence and dominance, as well as their like/dislike and familiarity with the video.

The dataset is public and provides an extensive set of data: the individual online ratings from the 120 subjects; details about the videos used, including the YouTube link, the segment starting time, quadrant of each video according to the subjects and others; ratings given by the participants in the experiment; face videos of the participants; the original physiological data and the preprocessed version of the same data. To obtain the data, researchers must request access and sign an end-user license agreement.

The main limitation of the study is the dataset's very limited size, since 40 segments are very limitative and cannot be used to achieve robust results. Still, given the embryonic state of MER induced research and the variety of source data available, this dataset is surely relevant.

Multi-modal MIREX-like emotion dataset – 2013

The Multi-modal MIREX-like dataset[77] is a dataset of perceived emotions in music previously proposed by us (Panda, Malheiro, et al., 2013). It follows the same categorical taxonomy of the MIREX AMC task, with five clusters and 29 adjectives, where its main advantages are the combination of several sources of information – audio, lyrics and symbolic files (MIDI).

The proposed automatic creation method is divided in three main steps. First, the AllMusic API was queried with each of the 29 MIREX AMC adjectives, obtaining the respective metadata. Then, a script was developed to parse the metadata and download the 30-second audio samples available at the same service. The second step is the annotation. Since each song entry in AllMusic is tagged with several emotion tags, these were grouped according to the MIREX clusters and the final annotation was based on the most significant cluster. This resulted in a total of 903 MIREX-like song entries and audio clips, nearly balanced across clusters: 18.8% in cluster 1, 18.2% in cluster 2, 23.8% in cluster 3, 21.2% in cluster 4 and 18.1% in cluster 5. The third step was used to obtain lyrics and MIDIs for the 903 songs. This was done automatically using a set of tools developed by the authors. The lyrics were extracted from lyrics.com[78], ChartLyrics[79] and MaxiLyrics[80], while MIDI versions were obtained from freemidi.org[81], free-

[77] http://mir.dei.uc.pt/downloads.html

[78] https://www.lyrics.com/

[79] http://www.chartlyrics.com/

[80] Nowadays offline

[81] https://freemidi.org/

midi.org[82], midiworld.com[83] and cool-midi.com[84]. After filtering invalid files, the final dataset contained 903 song entries, for which 903 audio clips, 764 lyrics and 196 MIDIs are available.

This dataset is one of the few MER multi-modal datasets, providing audio, lyrics and MIDI for the same songs. Still, the number of MIDI files is very limited when compared to the remaining sources. Moreover, the selected taxonomy has been largely criticized due to the lack of scientific support, as above mentioned. Finally, the quality of the annotations and audio segments was not assessed since they were built from AllMusic without manual verification.

Computer Audition Lab 500 Expansion (CAL500exp) dataset – 2014

Following the CAL500 success, Wang et al. (2014) proposed CAL500exp[85], an enriched version of its predecessor. As opposed to the original CAL500, the expansion was created to study MER in a smaller temporal frame as approached in music emotion variation detection (MEVD) studies. In detail, the creation process can be described in three steps. First, the original clips were divided into 3,223 variable-length (i.e., 3 to 16 sec) segments based on their acoustic contents. These segments are clustered to select representative segments for annotations. Secondly, before the annotation process, the list of possible labels for each segment was reduced to the tags used by CAL500 for each specific track. Finally, 11 annotators with strong music background were asked to refine the labels of each segment, by insertion or deletion. The final dataset uses 67 of the original CAL500 labels.

As with the original CAL500, the audio clips in this study are not public. In CAL500exp they are available upon request to the authors. Moreover, by presenting to subjects the CAL500 annotations as a baseline for each segment, the authors might have influenced the subjects' annotations. After all, the original labels assigned in CAL500 could be representative of the dominant emotions (and other descriptors) of the entire track and may not account for fluctuations present in some of the homogeneous segment.

LiveJournal Two-million Post (LJ2M) dataset – 2014

The LJ2M dataset[86] was created to foster research on "user-centered music information

[82] Nowadays offline

[83] http://midiworld.com/

[84] http://cool-midi.com/

[85] http://slam.iis.sinica.edu.tw/demo/CAL500exp/

[86] http://mac.citi.sinica.edu.tw/LJ

retrieval" (J.-Y. Liu, Liu, & Yang, 2014). Similarly to previous datasets of a more "uncontrolled" nature, this dataset contains a great diversity of emotional content, stemming from online users' blog posts, with associated songs and emotions labels. As with Last.FM tags, the context where each blog post was created and what it meant is completely unknown.

The dataset contains 1,928,868 blog posts created by 649,712 online users. A total of 88,164 unique song titles of 12,201 artists are associated with the blog posts. Although no audio clips are available, the authors provide the EchoNest track ID of the songs, which can be used to query the EchoNest API for metadata and audio features. A set of co-occurrence tables describing the content of each blog post is also provided, as well as some linguistic (textual) features.

Although having massive scale, this dataset does not provide audio clips. Still, the authors state that it is possible to gather 30-second samples from the 7digital website. Given its size and the triplets of text, emotion label and song, it provides an interesting set of information to data mining and big data research, from which interesting relations can be extracted. Still, given the nature of the data, contributed by users "spontaneously in their daily lives, instead of being collected in a controlled environment", conclusions extracted from it regarding MER should be taken with caution.

Greek Audio and Music Datasets – 2014

The Greek Audio Dataset (GAD) is a dataset of Greek popular music inspired in the Million Song Dataset (Makris, Kermanidis, & Karydis, 2014). The dataset provides audio features, lyrics and metadata of 1000 popular Greek tracks from the 60s up to today, covering 8 genres, some of which unique to Greek music (*rembetiko*, *laiko*, *entexno*, modern *laiko*, rock, hip hop/R & B, pop, *enallaktiko*). The provided features were extracted using jAudio (Mcennis, Mckay, Fujinaga, & Depalle, 2005) and an audio feature extraction web service[87] provided by the Vienna University of Technology and are related with timbre, rhythm and pitch. The audio clips are not included due to intellectual property rights. Still, the authors provide the YouTube link for each song's video. The dataset was manually annotated, containing genre and emotion labels.

The Thayer model was adopted for the emotion annotations, using the AV axes. Five annotators were used in the process, being instructed to listen and read the lyrics and annotate both AV using discrete (integer) values from 1 to 4. No indication is given whether the values are from perceived or felt emotions. F1-Score was used to assess the inter-annotator agreement (Boisen, Crystal, Schwartz, Stone, & Weischedel, 2000), obtaining a value of "0.8 approximately". The authors also state that for "clusters of mood with smaller F1-Score, a discussion between the annotators was taken part in order to

[87] http://www.ifs.tuwien.ac.at/mir/webservice/

reduce controversy" (Makris et al., 2014).

The authors later released the Greek Music Dataset (GMD), an extension to the GAD increasing the number of songs to 1400 (Makris, Karydis, & Sioutas, 2015). In addition to the 400 new songs, this extension added symbolic data (MIDI) for 500 files as well as features extracted from lyrics and MIDI files.

Both GAD and GMD are interesting by providing data for a very specific and less studied type of music. Nevertheless, their major drawbacks are related with the annotations, since it is not clearly stated whether the annotators used perceived or felt emotion. Moreover, complete songs seem to have been used, which usually contain fluctuations in the emotional content, as well as lyrics, which may lead to the perception or induction of a different emotion when compared to the audio only.

Music Mood Rating Dataverse – 2014

The Music Mood Rating Dataverse[88] (MMRD) dataset is not a typical MER audio dataset as it does not include audio clips or even audio features extracted from the employed audio clips. MMRD is the result of an automatic technique proposed by Paasi et al. (2014) to represent the emotions in music tracks based on Last.FM social tags data.

The motivation behind MMRD was the infeasibility of applying a traditional ground truth construction method, made of laborious survey-based annotations, to the real-life dimensions of today online databases. In contrast to the rigorous laboratorial procedure applied previously by one of the authors' soundtrack dataset (Eerola & Vuoskoski, 2011), this method uses the massive amount of uncontrolled tags created by online users to achieve the same goal. To this end, the authors first collected 1083 words related to emotion from several studies and the "expert-generated source" list available in AllMusic (Saari & Eerola, 2014), which was later reduced to 568 by merging the inflected forms (e.g., "depressed", "depressing" and "depression"). Next, the Last.FM API was used to crawl tracks associated with these tags and their weights, based on the number of times the tag was assigned to the track. The resulting 1,338,463 tracks were then filtered to remove less represented tags and songs, resulting in a corpus of 259,593 tracks and 357 emotion terms. To produce a semantic space of the gathered tags, the authors then applied non-metric Multidimensional Scaling (MDS) (Kruskal, 1964), a "set of mathematical techniques for exploring dissimilarity data by representing objects geometrically in a space of a desired dimensionality" (Saari & Eerola, 2014). To represent a track in the obtained MDS space, the authors then applied a projection (using center-of-mass) based on the terms associated to the tracks.

The Hopkins' index (Hopkins & Skellam, 1954) was then used to estimate the de-

[88] https://dataverse.harvard.edu/dataverse/musicmoods

gree of "clusterability" of the obtained MDS space, with results suggesting that the optimal representation of the emotions is continuous rather than categorical. To this end, the authors proposed a novel technique called Affective Circumplex Transformation (ACT) influenced by Russell's affective circumplex model of emotions to conceptualize the dimensions. The first step consisted in gathering AV positions of 101 terms from Russell's (Russell, 1980, p. 1167) and Scherer's (Scherer, 1984, p. 54) studies, reducing them to 47 by matching against their MDS space terms. The second step transformed the 3D MDS space to optimal arousal, valence and tension (AVT) by "classical Procrustes analysis (Gower & Dijksterhuis, 2004), using sum of squared errors as goodness-of-fit". A similar method can be used to position new songs in the AVT space based on its tags.

The method was validated by using a test set of 600 15-second samples from Last.FM previews annotated by 59 subjects. The positioning of the test set in the AVT plane according to the subjects' annotations is then compared to the ACT predicted positioning, demonstrating the efficiency and robustness of the solution.

Although not providing audio clips, the authors made the test set data available, which includes track info with Last.FM and Spotify URLs, as well as mean and raw subjects' annotations. Even if other researchers can crawl the samples from the given URLs, there is no guarantee that the tags are representative of the segments used by subjects to generate the annotations, after all Last.FM tags are normally 30-second clips and here 15-second clips were used. Additionally, the 357 emotion terms used to build the MDS are also provided but not the data from the 259,593 tracks which could be used to replicate the study and generate a larger dataset automatically annotated with the proposed ACT method. Finally, it is important to note that while the subjects' annotations are specific to a 15-second clip, the social tags available in Last.FM can be related to any segment of the song, the full song or even a personal experience or memory from the user. As an example, a user might have tagged a song as "hate" to represent the perceived emotion or because he hates that specific song.

AMG1608 Dataset for Music Emotion Recognition – 2015

AMG1608[89] is a dataset for music emotion analysis of considerable size, comprising 1608 30-second music clips (Y.-A. Chen, Yang, Wang, & Chen, 2015). The songs are mostly western, and were selected from the AllMusic service, while the excerpts were gathered from 7digital, a digital music and radio service. The dataset can be divided into two distinct parts. The first part is called the "campus subset", consisting of a subset of 240 songs annotated by 22 subjects from the National Taiwan University and Academia Sinica. The second part, the "Amazon Mechanical Turk (MTurk) subset", consists of

[89] http://mpac.ee.ntu.edu.tw/dataset/AMG1608/

annotations for all 1608 clips collected through the MTurk platform. A total of 643 subjects collaborated in this, generating 15 annotations per audio clip. The subjects were asked to rate the perceived emotion in terms of arousal and valence, using Russell's dimensional model. The individual annotations of each of the 665 subjects is provided.

Some steps were taken to reduce the possible problems usually associated with crowdsourced annotations. Namely, the MTurk process was limited to residents of the United States which had completed at least 90% of their tasks on Amazon Mechanical Turk. Following, the inter-subject agreement on the annotations was assessed by using Krippendorff's alpha. The results were considered of "fair agreement", with 0.31 for valence and 0.46 for arousal (Y.-A. Chen et al., 2015).

Although having a considerable size and providing valuable data in addition to AV annotations, such as individual annotations and songs' metadata, the dataset contains a major limitation. Due to copyright restrictions, the data provided to researchers consists of a single MATLAB file, containing annotations, songs' metadata and 72 audio features extracted from the original clips. This severely limits its application in new studies, since no additional features can be extracted from the audio. Additionally, the annotations were created in an online environment, i.e., MTurk and the dataset is unbalanced in terms of quadrants, with most songs belonging to the first quadrant.

Emotify dataset – 2016

The Emotify dataset[90] is a MER dataset on induced emotion collected through the GWAP Emotify, developed by Utrecht University (Aljanaki, Wiering, & Veltkamp, 2016). The dataset consists of 400 song excerpts with 1 minute each divided in 4 genres – rock, classical, pop and electronic. The annotation process used the Geneva Emotional Music Scales (GEMS), an emotion model specifically devised to measure musically evoked emotions (Zentner et al., 2008). GEMS consists of 45 emotion labels (GEMS-45), which are organized into nine categories (GEMS-9) and these in three superfactors (see Section 2.2.1). In this study, each participant selected up to three of the GEMS-9 categories (two of which were renamed by the authors), based on the emotion felt while listening to the clips.

A total of 1778 people participated in the online annotation process. Since the users could skip songs and select the genre, the annotations are spread unevenly between songs. The provided dataset contains individual annotations of each user, consisting of the selected emotions from nine possible – amazement, solemnity, tenderness, nostalgia, calmness, power, joyful activation, tension and sadness. Information about the participant is also included, namely his/her age, genre, mother language, mood prior to playing the game and whether he/she liked or disliked the clip. The original audio clips are also

[90] https://www.projects.science.uu.nl/memotion/emotifydata/

provided.

Overall, the Emotify dataset provides audio clips and categorical annotations for 400 songs. Still, it is specific to the study of emotions induced in the listeners by music. Such topic has been less explored in MER, given the complexity associated with induced emotions, as described in Section 2.1 (e.g., it is relatively personal and context / memory dependent, when compared to perceived emotions). Moreover, the annotations were collected online, in an uncontrolled environment, which although being a "cheaper and more efficient approach", typically achieve this by compromising the quality of the obtained annotations (Burmania, Parthasarathy, & Busso, 2016).

Malheiro's (audio and lyrics) emotion dataset – 2016

Our team (Malheiro et al., 2016a) also proposed another dataset[91] combining different sources of information – lyrics and audio. To build the dataset, an initial set of 200 song lyrics and 30-second audio clips of various genres was selected. The initial set was uniformly distributed in terms of Russell's quadrants according to the authors' perceived emotion. Next, 39 subjects assigned values (between -4 and 4 with a granularity of one unit) to valence and arousal according to their perceived emotion. The final annotations were obtained by averaging the users' annotations, excluding songs with standard deviation higher than 1.2. A substantial agreement between annotators was observed according to Krippendorff's alpha.

After filtering the data, the final dataset contains 180 lyrics and 163 audio clips, with 133 songs containing both lyrics and audio clips. The limited size of the dataset is its main issue, a common characteristic shared with other datasets where a controlled annotation process was adopted. Still, the original paper is not very detailed on the annotation process. Also, the provided package[92] contains only the audio clips, while lyrics, due to copyright reasons, are represented with links where each one can be accessed.

MediaEval Database for Emotional Analysis in Music (DEAM) – 2016

The DEAM[93] dataset was created by Aljanaki et al. (2017) and is one of the largest public datasets available. It consists of 1802 royalty-free audio files (58 full-length songs and 1744 excerpts of 45 seconds) covering several genres (e.g., rock, pop, electronic, country, jazz). The dataset is an aggregation of the datasets from the "Emotion in Music" task at MediaEval benchmarking campaign 2013-2015 and respective labels. In detail, it is composed of the 2014 development set (744 songs), 2014 evaluation set (1000 songs) and

[91] http://mir.dei.uc.pt/downloads.html

[92] http://mir.dei.uc.pt/resources/MER_bimodal_dataset.zip

[93] http://cvml.unige.ch/databases/DEAM/

2015 evaluation set (58 songs). The 45 second clips have a sampling rate of 44100 Hz and are "extracted from random (uniformly distributed) starting point in a given song" (Soleymani, Aljanaki, & Yang, 2016).

The clips are annotated with valence and arousal values both continuously and over the whole song. The continuous annotations were collected using "a sampling rate which varied by browsers and computer capabilities" (Soleymani et al., 2016) and transformed into average annotations at 2 Hz sampling rate, ignoring the first 15 seconds due to instability. Yet, the raw individual annotations are still provided, as well as the standard deviation. This process was conducted partially in the researchers' lab and on the Amazon Mechanical Turk crowdsourcing platform by a minimum of 10 workers for the 2013 and 2014 songs, while 5 workers were used in the 2015 set, two of which were from the lab. To increase the annotations quality, these workers had to pass a filtering stage that excluded poor quality workers based on state-of-the-art crowdsourcing approaches (Soleymani & Larson, 2010), involving both multiple choice and free form questions.

The authors also assessed the annotations' quality by employing Cronbach's alpha, a coefficient of internal consistency (Cronbach, 1951). Based on it, the static annotations were considered far more consistent than the continuous ones, passing the threshold of 0.7, which was considered an acceptable agreement between annotators (Aljanaki, 2016). The continuous annotations achieved values lower than the threshold, showing worse consistency than "what could be achieved for static (per song) annotations" (Aljanaki, 2016).

Although this is a large dataset that provides audio samples, extracted features, dimensional annotations and other metadata, it has important problems. First, the criteria used to select audio clips, with a fixed length of 45 seconds randomly selected from the full song, might be problematic. Namely, segments with contrasting or unclear emotions, or even not representative of the song might have been selected. This was confirmed by inspecting the clips, where we found some poor quality samples containing noise (e.g., claps, speak, silences) or clear variation in emotions. Furthermore, other researchers have found additional issues such as low agreement between the annotations from the original subjects when converted to quadrants (Vale, 2017, p. 18). To this end, Vale transformed both workers' AV annotations and the final AV annotations of each song into quadrant annotations and calculated the agreement, obtaining only 47%. In addition, the same author verified that for 1133 clips (67.41%) the majority of the subjects did not even annotated the same AV quadrant (Vale, 2017, p. 18). Finally, the number of songs for each of the four quadrants defined by the AV axis is also very unbalanced, with 681 clips for quadrant 3 and only 200 for quadrant 2 (Vale, 2017, p. 21).

Summary of existent Audio Datasets for MER

Even though an extensive review was provided, some other datasets exist in addition to

the abovementioned ones. Most of these are less relevant and share similar characteristics and issues as the described. Some examples are:

1. Smaller datasets, as the EMusic dataset[94], containing 140 clips where most (100) are experimental music and the remaining (40) belong to 8 distinct musical genres, with AV annotations generated though crowdsourcing (J. Fan, Tatar, Thorogood, & Pasquier, 2017). Another example is Yang's MER60 dataset[95] (Y.-H. Yang, Su, Lin, & Chen, 2007).
2. Public datasets where the provided data is very limited, such as the Trohidis et al. (2008) dataset[96], which consists of 593 songs annotated using 6 emotions (multi-label) by 3 expert listeners. In this case, the only data provided is a matrix of 593 x 72 features and the annotation, without any information on the employed songs.
3. Larger datasets, such as the Magnatagatune[97] (Law, West, Mandel, Bay, & Downie, 2009), which, although providing very interesting data (audio clips, metadata, annotations, audio features, similarity data), due to its size (approximately 21,000 clips) compromised the quality of the annotations using a less controlled collection process. In this case, the taxonomy consists of 188 different tags and was annotated following a GWAP approach (TagATune), considered rather difficult to handle due to its size and skewed tag distribution (Seyerlehner, Widmer, Schedl, & Knees, 2010). Other example is the CAL10k dataset[98], which contains 10,870 songs tagged with 628 labels harvested from Pandora. Unlike its predecessors, this dataset does not contain tags directly related with emotion.
4. Private, as is the case of the dataset used to predict genre and emotion in Laurier's PhD Thesis (2011).

A brief summary of the described datasets is presented in Table 3.3.

[94] http://metacreation.net/project/emusic/

[95] http://mac.citi.sinica.edu.tw/~yang/MER/hcm07/index.html

[96] https://sites.google.com/site/hrsvmproject/emotions.rar?attredirects=0

[97] http://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset

[98] http://calab1.ucsd.edu/~datasets/

Dataset	Type of data / Availability	Size	Taxonomy	Emotion Type	MER problem	Annotations	Comment
MIREX AMC	Audio / no data is available.	600 clips (30-sec).	Categorical (data-derived) taxonomy of 5 clusters.	Perceived emotion.	Single-label / static MER.	Annotated by 2-3 subjects (experts).	Private, used in MIREX. Several issues have been found, taxonomy not psychologically validated.
CAL500	Audio / no original data available but features are available via Million Song dataset.	500 clips (full songs).	Categorical, 159 labels (36 emotion related).	Unknown.	Multi-label (binary and numerical [0, 1] / static MER.	Annotated by 66 students, at least 2 annotators per song and 80% agreement.	No songs are provided due to copyright, original data is offline nowadays.
Yang's AV	Audio / available upon request.	195 songs (25-sec clips).	Dimensional using Russell's model (AV).	Induced.	Regression / static MER.	253 annotators, at least 10 per clip. Test-retest with 22 subjects.	Clips were segmented manually. Issues have been reported with the annotations and data (see Section 5.1.1).
MTurk240	Audio / only features are available.	240 songs (15-sec clips).	Dimensional, using AV.	Perceived.	Regression / MEVD.	Initially from GWAP, later via MTurk.	Annotations remade with MTurk to mitigate initial issues.

Dataset	Type of data / Availability	Size	Taxonomy	Emotion Type	MER problem	Annotations	Comment
Soundtrack dataset	Audio / audio clips, documentation and user preferences directly available.	360 (pilot) / 110 (final) 10 to 30-sec clips.	Categorical using 6 labels. Dimensional using AVT.	Perceived.	Single-label and regression / static MER.	116 subjects (later reduced to 110), using a very controlled procedure.	Limited size, very rigorous ground truth collection.
Million Song dataset	Audio / audio features, metadata, additional data available for smaller subsets (e.g., lyrics). Audio samples can be downloaded indirectly.	Nearly 1 million song entries.	Categorical, +500,000 unique labels.	Unknown.	Single-label / static MER.	Extracted from Last.FM social data.	Very large size, ideal for real-life scenarios. Audio samples may not be the same as in the original study. Annotations are completely uncontrolled.
DEAP120	Video / excerpts used can be downloaded unequivocally from YouTube. Subjects ratings and biosignals available.	Set1: 120 music videos (1-minute). Set2: 40 (subset of the 120).	Dimensional using arousal valence and dominance (AVD).	Induced.	Regression / static MER.	Set1: online, 14-16 subjects per clip. Set2: controlled, 32 participants.	Limited number of videos, but massive amount of data. No segments available due to copyright. Interesting amount of information sources.

Dataset	Type of data / Availability	Size	Taxonomy	Emotion Type	MER problem	Annotations	Comment
Multi-modal MIREX-like	Audio, lyrics, MIDI / original data directly available.	903 songs (30-sec clips), 764 lyrics and 193 MIDI files.	Categorical, using MIREX AMC clusters (5) and labels (29).	Perceived	Single-label and multi-label / static MER.	Automatic extraction from AllMusic emotion tags.	Limited MIDI subset, unvalidated taxonomy, lacks validation of the audio segments and annotations.
CAL500exp	Audio / segments available upon request.	500 tracks (3,223 segments of 3 to 16 seconds).	Categorical, 67 labels from CAL 500.	Unknown.	Multi-label / static MER.	CAL500 annotations refined by 11 subjects.	Possible MEVD dataset, annotations method may produce bias.
LJ2M	Triplets (3-tuples) containing: blog post, emotion tag and music title. Metadata, EchoNest features and predicted emotion.	1,928,868 blog posts with 88,164 unique song titles of 12,201 artists.	Categorical, uses 132 emotion tags predefined by LiveJournal.	Unknown.	Single-label / static MER.	From 649,712 online users.	No clips available but 7digital IDs are provided. The blog data was previously generated by online users without any guidelines.
GAD	Audio and lyrics / the excerpts used can be	1000 songs.	Dimensional, using AV.	Unknown.	Regression / static MER.	Five annotators used.	Provides only audio features but YouTube links are available. Lacks

Dataset	Type of data / Availability	Size	Taxonomy	Emotion Type	MER problem	Annotations	Comment
	downloaded unequivocally.						detail on the annotations process.
GMD	Audio, lyrics, MIDI / similar to GAD.	1400 songs (extends GAD).	Dimensional, using AV.	Unknown.	Regression / static MER.	No information apart from "2421 annotations for 1400 songs" (1.7 per song).	Provides MIDI files and features only for 500 songs.
MMRD	Audio / only metadata, excerpts can be downloaded but with no guarantee that they match the ones used in the study.	600 songs (15-sec clips).	Dimensional, using AVT.	Perceived.	Regression / static MER.	The method was validated by 59 subjects.	Authors propose a method to automatically annotate songs using Last.FM data. Provides the data used to validate the method.
AMG1608	Audio / only features are available.	1608 songs (30-sec clips).	Dimensional, using AV.	Perceived.	Regression / static MER.	10 MTurk workers per song, subset of 240 was also annotated by 22 volunteers.	Unbalanced in terms of quadrants, no audio clips available.

Dataset	Type of data / Availability	Size	Taxonomy	Emotion Type	MER problem	Annotations	Comment
Emotify	Audio / original clips directly available.	400 songs (1-minute clips).	Categorical, using GEMS-9.	Induced.	Multi-label / static MER.	Obtained online via a GWAP from a total of 1778 players.	No metadata provided, uncontrolled annotations on induced emotions via online game.
Malheiro's (audio and lyrics) emotion dataset	Audio and lyrics / original clips directly available, lyrics used can be downloaded unequivocally.	200 songs (30-sec clips).	Dimensional, using AV. Categorical, using AV quadrants.	Perceived.	Single-label and regression / static MER.	Annotated by 39 subjects (audio and lyrics annotated separately).	Dataset: 180 lyrics, 163 audio clips, and 133 audio + lyrics after filtering. Limited size.
DEAM	Audio / original clips and features directly available.	1802 songs (48 full songs, 1744 45-sec excerpts).	Dimensional, using AV.	Perceived.	Regression / static MER and MEVD.	5-10 MTurk workers per song, continuous AV sampled at 2Hz.	Segmented randomly, audio clip issues have been reported.

Table 3.3: Summary of the most well-known audio datasets used in MER research.

3.2.3. Feature Extraction, Selection and Reduction

Having concluded the construction of the ground-truth, the second stage of MER studies is typically the feature extraction process. The goal is to extract musical descriptors that best represent and summarize the content of the selected audio clips. To this end, the audio signals (or other sources, such as MIDIs or lyrics text files) are processed by computational algorithms, extracting the selected audio features. The output of these audio extractors varies greatly, from points in time (e.g., when extracting onsets), to single values (e.g., tempo estimation) or series of values (e.g., spectral features extracted for each frame), to grids and other representations (e.g., similarity matrix). Before being ready for pattern recognition, these sets of data are usually cleaned and summarized, for example using statistical moments.

As described in Section 3.1, a significant number of features has been created over the years to capture information from audio signals. From these, several have been associated with different emotional states, either in musicological studies, as extensively addressed in Section 2.4, or experimentally by different MER studies. As an example, intensity and related measures of amplitude are usually correlated with arousal and have been used to predict arousal dimension (Zhang, Huang, Yang, Xu, & Sun, 2017). Moreover, timbre features such as MFCCs, spectral shape, spectral contrast and various spectral moments are frequently used to describe the sound quality (Fu, Lu, Ting, & Zhang, 2011). Other dimensions such as rhythm have been exploited, namely by extracting features such as tempo, rhythm strength or rhythm regularity (D. Liu & Lu, 2003).

Selecting the audio features to be extracted is a difficult task. While researchers can formulate hypotheses about limited sets of features that should be suitable, grounded on previous studies, acquired experience and knowledge, this strategy may end up limiting the results. Several factors contributing to this are the possible errors in extracted features (due to limitations in the employed algorithms, e.g., melody detection is still an open research problem), lack of knowledge and experience or personal bias, as well as contradictions or specificities of previous studies (e.g., applying only to a specific genre) that may not translate into newer ones.

Empirically, one might assume that extracting as much features as possible and using all of them is the solution, since it means more information describing the sets. However, this has been disproved, as using many features can, actually, lead to a decrease in classification results (Zhang et al., 2017). Understandably, having an excess of information, especially descriptors that are not relevant to the problem being tackled, will serve as noise and increase the complexity of the classifier. To address this problem of excessive feature dimension problem, feature selection and reduction techniques are employed.

The problem of dimensionality reduction is typically addressed in one of two ways: feature projection or feature selection. In feature projection, the original data is reduced by transforming it from a high-dimensional to a lower-dimension space. This transformation may be linear, as is the case of principal component analysis (PCA) (Hotelling, 1933; Pearson, 1901) or nonlinear, as with isometric feature mapping (ISOMAP) (Tenenbaum, Silva, & Langford, 2000) or locally linear embedding (LLE) (Roweis & Saul, 2000). In addition to PCA, the Karhunen-Loève (KL) expansion (Fukunaga & Koontz, 1970) has also been used effectively in MER to remove cross-correlative features (L. Lu et al., 2006). This algorithm, closely related to PCA, is applied to the extracted features, obtaining a new, reduced and decorrelated feature set by mapping them into an orthogonal space. The main issue with methods such as KL and PCA is that, although the dimensionality is reduced, the new set of features is transformed, making it difficult to understand the importance of each original descriptor.

A different approach to the previous methods, which relies on statistical analysis of the feature space, is to actually evaluate the performance of features by running actual tests with combinations of features. Since exhaustive tests with every possible combination of features is practically impossible, several approaches have been created. Depending on the ground-truth (i.e., labeled vs unlabeled), this process can be divided into supervised (e.g., (L. Song, Smola, Gretton, Borgwardt, & Bedo, 2007)), unsupervised (e.g., (Mitra, Murthy, & Pal, 2002)) or semi-supervised (e.g., (Zhao & Liu, 2007)) feature selection.

Supervised feature selection methods can be further divided into three groups: filter models, wrapper models and embedded models (Tang, Alelyani, & Liu, 2014). The first, filter model, does not relies on classifier learning to avoid bias between the machine learning and feature selection algorithm. Instead, the method uses statistics extracted from the training data (e.g., distance) and correlates it with the associated labels. Relief (Kira & Rendell, 1992) is one of the filter approaches found commonly on MER literature. On the other end of the spectrum, the wrapper model relies on learning algorithms, assessing the features' quality based on the prediction accuracy of such models. The application of such model is many times prohibitively expensive due to the dataset size and associated computational requirements. The embedded model is a compromise between the previous two. In this case, a first step uses statistical analysis, similarly to filter models, to select various subsets of candidate features. Following, the subsets are evaluated using learning models and the best performing one is selected. The embedded model has been shown to be a good compromise between the former two (e.g., (Cawley, Talbot, & Girolami, 2007)).

As no labels are available, unsupervised feature selection is normally based on clustering algorithms and associated quality metrics. Since such methods work with unlabeled data, one of the drawbacks is the difficulty to evaluate the relevance of features. For a comprehensive review on unsupervised feature selection refer to (Aggarwal &

Reddy, 2013).

A typical wrapper approach is forward feature selection (FFS), also called stepwise forward selection (Heinze, Wallisch, & Dunkler, 2018). This method, described in Algorithm 3.1, has been used extensively in MER (Hu & Yang, 2017), including in our works (Panda & Paiva, 2012b). In the opposite direction, backward feature selection (BFS), also called stepwise backward selection (SBS) (Heinze et al., 2018), starts with the full set of features, and is used to greedily remove the sequence of worst performing ones (i.e., backward elimination) until the stop criteria is met (e.g., no features are left or the accuracy starts to decrease).

Algorithm 3.1. Forward feature selection (also known as stepwise forward selection) algorithm.

1. Starting with an empty set of features, S, and a set of all the remaining features, R.
2. While the remaining set of features, R, is not empty:
   2.1. For each feature, R_i, in the remaining set, R:
      2.1.1 Add feature R_i and all the features in S to an empty set of features, T.
      2.1.2. Evaluate the prediction performance, P_i, of T using the selected machine learning strategy (e.g., 10-fold cross-validation using Support Vector Machines).
   2.2. Select the best performing feature, R_x, where $x = argmax(P)$, and move it from R to S.
3. The generated set, S, contains all the features ordered by relevance, where the first is the most relevant.

Although not exhaustive, such strategies are still computer-intensive, especially with a high number of features and large dataset. Hence, filter methods were proposed. As mentioned, one of these is Relief, a feature selection algorithm sensitive to feature interactions (Kira & Rendell, 1992) that has been used previously in MER (Malheiro et al., 2018; Y.-H. Yang, Lin, Su, et al., 2008). The original version of the algorithm supports only binary classification problems (two classes), computing the weight of each feature as follows.

Algorithm 3.2. Original Relief feature selection algorithm.

1. Given a dataset of n instances (e.g., songs), from which p

features were extracted, $F_{n,p}$, and annotated with one of two possible classes.

2. Scale each feature (normalize) to the same interval [0, 1].
3. Starting with an empty (zero-filled) weight vector W of length p, W_p, repeat for m times[99]:

   3.1. Select a random instance (song), i, and its feature vector, F_i.

   3.2. Select the two songs closer to F_i, one for each class, using the Euclidean distance between feature vectors.

   3.3. Compute the weight for instance (song) i, W_i, given by equation (3.1).

4. Compute the final weight vector, W, by dividing each element of the vector, W_i, by m.

The weight for a given feature, W_i, is computed as:

$$W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \quad (3.1)$$

Where $nearHit_i$ and $nearMiss_i$, represent the value of feature i of the closest same-class and different-class instances respectively.

Several variations have been introduced to the original Relief algorithm. Kononenko et al. proposed ReliefF (1997), which is more reliable when noisy features are present and supports multi-class problems. The first issue is solved by using the K-nearest hits and misses, averaging their contribution to the weights of each feature. For multi-class problems, ReliefF uses the K nearest misses from each different class and averages their contributions to W based on the probability of each class. Other approaches, such as RReliefF for regression problems, have also been proposed (Robnik-Šikonja & Kononenko, 1997). Relief algorithm and its variants output a value between -1 and 1 for each attribute, with more positive weights indicating more predictive attributes. Several variations exist, namely, ReliefFequalK, where the K nearest instances have equal weight, and ReliefFexpRank, where the K nearest instances have weights exponentially decreasing with increasing rank.

Finally, embedded approaches are a compromise of filter and wrapper models. An example of this consists in the application of Relief (a filter model), obtaining a feature

[99] The original paper defines m as the "sample size" while (Kononenko et al., 1997; Robnik-Šikonja & Kononenko, 1997) state that m is "a user-defined parameter" in both the original algorithm and their updated version.

rank and respective feature weights. Then, different subsets of the best features are selected and tested using machine learning algorithms. This was one of the solutions adopted by us, as described in Chapter 4.

In our experiments, the embedded strategy was used with Support Vector Machines (SVM) as the machine learning technique (Cortes & Vapnik, 1995). SVMs are supervised machine learning algorithms that work by building a hyperplane (or set of hyperplanes) that maximize the margin (i.e., separation) between the classes represented in the training data. Originally, SVMs work as a non-probabilistic binary linear classifier (separator). When data is non-linear SVMs use a kernel to project the data into a high-dimensional plane where it can be separated. This is a strength but also a weakness of SVMs, since its effectiveness depends on the selection of kernel and parameters and no single kernel is perfect for every dataset. A common choice is a radial basis function kernel (RBF), while a polynomial kernel performs better in a small subset of specific cases. In our preliminary tests RBF performed better and hence was the selected kernel (see Chapter 5). For multi-class classification the common strategy is to reduce the problem in multiple binary classification problems. Of the two typical approaches: "one-vs-all" and "one-vs-one", the selected SVM implementation – libSVM (Chang & Lin, 2011), uses the later ("one-vs-one"), combining the results with a max-votes system.

3.2.4. Classification and Evaluation

Following the creation of the dataset and feature extraction and selection, the third and final stage in a typical MER system is the creation of a computation model able to recognize patterns in the data, in order to classify new songs. To this end, the audio features representing the songs and the previous obtained annotations are fed to one of the many existent machine learning techniques. These algorithms, which vary greatly in their operating mode, aim to identify patterns in the input data – a set of statistics that correctly models the existent cases. Based on the identified rules, new songs can be classified according to their audio features.

A music emotion classification problem, or any other classification problem, can be organized in three types depending on the nature of the ground-truth and of the type of predictions: single label, multi-label or fuzzy.

Single label emotion classification consists in assuming that emotions in music are exclusive. As such, subjects annotate the songs using a single label and the model is trained to predict a single emotion to a given song. This method is by far the most frequently employed in MER also due to being simpler (e.g., (Feng et al., 2003; L. Lu et al., 2006; Panda & Paiva, 2012b)). It is also the classification method used in the MIREX AMC task. Since single-label classification regards emotions as something deterministic/discrete, it ignores the possible ambiguity of emotion terms and subjectivity across

human annotation.

Multi-label classification was introduced in MER to solve possible ambiguity and subjectivity problems. In this approach, human annotators can label audio segments with a variable number of discrete labels. Even though viewing emotion as multi-label can be considered closer to reality, fewer MER studies have used this approach (Li & Ogihara, 2003; Trohidis et al., 2008; Wieczorkowska, Synak, & Raś, 2006; B. Wu, Zhong, et al., 2014) and low results are typically reported (e.g., (Li & Ogihara, 2003; Wieczorkowska et al., 2006)). A possible justification is the increased complexity of the various steps (from ground-truth acquisition to classification).

Fuzzy classification is an extension of the multi-label approach, where each of the identified emotions may have a different weight or probability, as opposed to being binary (present or not present). As with multi-label classification, fewer researchers have tackled MER as a multi-label fuzzy problem (e.g., (Myint & Pwint, 2010; Y.-H. Yang, Liu, & Chen, 2006)).

Due to some difficulties observed in emotion classification, such as the known limitations in categorical taxonomies (e.g., lack of discrimination within categories, ambiguity and subjectivity – for further details refer to Chapter 2), as well as the difficulties to further improve the state-of-the-art, shown by MIREX AMC task stagnant results over the last decade[100], more researchers have focused on regression solutions. Tackling emotion recognition as a regression approach consists in numerical-value prediction and requires a ground-truth with similar-type annotations, following one of the existing dimensional models of emotion. To the best of our knowledge, Yang et al. (2007) was one of the first to approach MER as a regression problem in 2007, using arousal and valence as dimensions. Since then, many other researchers followed the idea and addressed MER as a regression problem (Hu & Yang, 2014; Malheiro, Panda, Gomes, & Paiva, 2016b; Panda & Paiva, 2011b; Schmidt, Turnbull, & Kim, 2010).

Over the last years, some authors have noted that, while representing emotion as a point in the emotional space removes the ambiguity caused by labels in categorical models, it fails to represent the subjectivity of the human annotation. This is caused by authors considering the emotion of a segment to be the mean value of all annotators. To mitigate this, Schmidt et al. (2010a) proposed a continuous probabilistic distribution representation of music emotion, combining the advantages of fuzzy classification and numerical-value prediction. Here, emotions in music are seen as areas in the dimensional space, representing the various human annotations.

The various MER approaches described above are illustrated in Figure 3.4.

Several model learning algorithms have been tested in MER over the decades to identify patterns between the extracted features and the ground-truth annotations.

[100] http://www.music-ir.org/mirex/

Among which, support vector machines (SVM) (e.g., (Panda, Rocha, & Paiva, 2015)), Gaussian mixture models (GMM) (e.g., (L. Lu et al., 2006)), neural networks (e.g., (Feng et al., 2003)), boosting (e.g., (Q. Lu, Chen, Yang, & Wang, 2010)), k-nearest neighbor (KNN) (e.g., (Y.-H. Yang et al., 2006)), Bayesian networks (e.g., (W. Wu & Xie, 2008)), decision trees (e.g., (Ma, Sethi, & Patel, 2009)) and others. Recently, deep learning techniques have been gaining ground thanks to the increased performance of recent computer processors (e.g., (S.-H. Chen et al., 2015; Delbouys, Hennequin, Piccoli, Royo-Letelier, & Moussallam, 2018)).

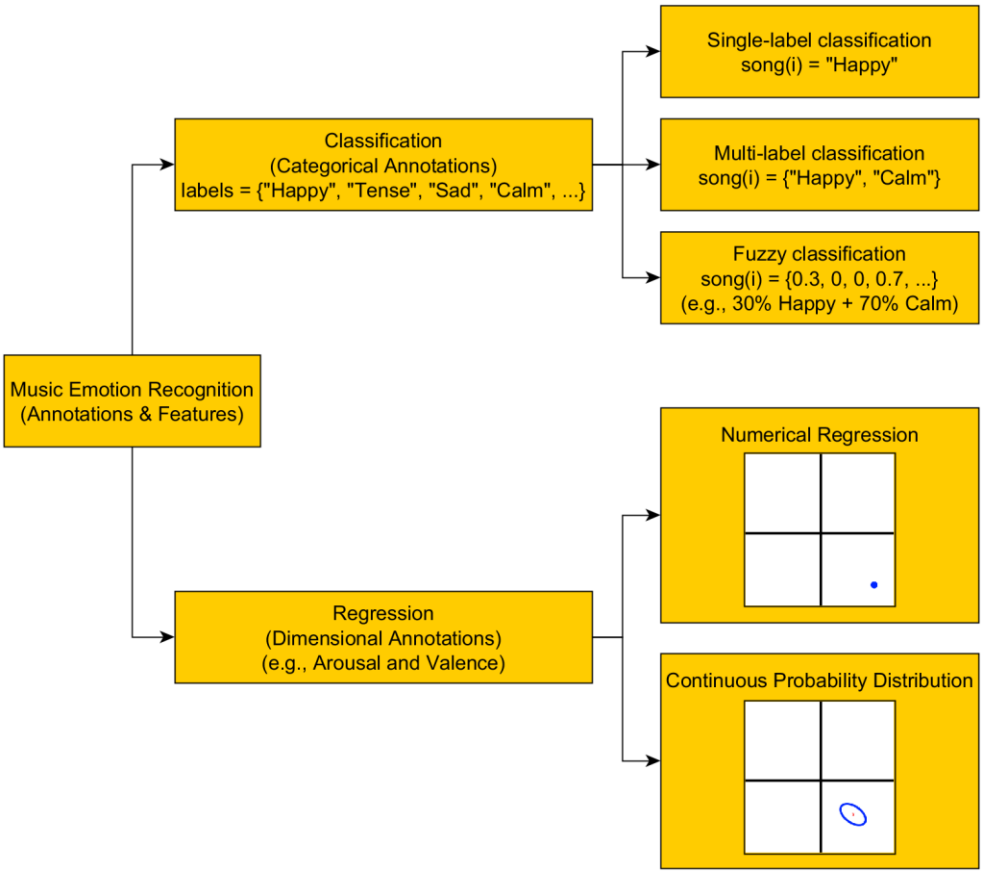


Figure 3.4: The different MER classification and regression approaches based on the type of ground-truth.

To evaluate the performance of a new approach, the dataset is usually split into training and the test sets. The main motivation behind this is to optimize two key objectives: select the best model and estimate its performance, since using the entire available data to select the model and estimate its error rate will usually result in overfitting and

overly optimistic error rates. This data splitting stage can be performed following different strategies, e.g., hold-out, train-validate-test, k-fold cross-validation. The hold-out method consists in splitting the dataset into the two abovementioned groups (i.e., train and test), using one to train the classifier and the other to estimate the error rate. Such simple method has some drawbacks, such as being highly dependent on the quality of the split, which can generate very different sets. More advanced solutions exist, such as cross-validation, which resample the two groups and repeat the train and test process. One example is *k*-fold cross-validation, where the dataset is split in *k*-subsets or folds and *k* experiments are run, using *k-1* folds for training and the remaining for testing (Refaeilzadeh, Tang, & Liu, 2009, p. 532).

Whatever the approach is, feature selection is performed using only the training set. The training set of songs and their annotations is used to train the classification or regression model. The test set of songs is used to evaluate the trained model's accuracy by predicting their annotations. The predicted values are then compared to the original labels, giving an indication on how accurate the prediction was. Several metrics exist to measure the classification and regression performance. For classification, typical measures include precision, recall, accuracy and F1-Score. For regression, typical measures are R^2 score, mean absolute error and mean squared error.

In classification problems, there are four terms used to compare the labels output by the classifier to the original labels generated by the annotators. Imagining a simple binary problem of classifying a song as "happy" or not, these are: 1) true positive (TP) – a correctly classified song, e.g., a happy song correctly classified as happy; 2) true negative (TN) – a correct rejection, as in a song that is not happy, correctly marked as such; 3) false positive (FP) – a false alarm, a song wrongly classified as "happy"; 4) false negative (FN) – a miss, in this example a happy song which was not classified as such. From these, precision, recall and accuracy metrics can be defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

F1-Score, also known as F-Score or F-Measure, is commonly adopted since it combines both precision and recall using a harmonic mean:

$$\text{F1-Score} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

Other variants exist, such as weighted, macro and micro versions of the F1-Score to be used in multi-class classification problems.

For regression problems, where dimensional models are used, a typical metric is the coefficient of determination, R^2. This coefficient indicates how well the data fits a statistic model. The best possible score is 1.0, indicating that the model perfectly fits the data. A value of 0 indicates that the model does not fit the data at all.

The coefficient of determination, R^2, is given by:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{3.6}$$

where y_i is the annotated value (from the ground-truth) of the song i in the dataset, $\hat{y}_i$ represents the predicted value to the same song and $\bar{y}$ represents the mean annotation of the observed data.

3.2.5. Review of Core MER problems: Brief Historical Contextualization

The previous sections described the typical MER system in a generic way, presenting each of the major steps and how researchers have addressed them over the years. This section gives an historic contextualization of the major breakthroughs in the area of MER, focusing on the problem of perceived emotion.

Particularly, based on the reviewed literature, MER research is focused on three core problems: emotion classification, emotion regression and emotion variation. Over the last three decades, authors have been proposing solutions to solve these cardinal problems, which are summarized below.

Over the decades, some researchers have published MER state-of-the-art reviews and books such as (Y. E. Kim et al., 2010; Wieczorkowska, 2004; X. Yang et al., 2017; Y.-H. Yang & Chen, 2011a, 2012). These shed light on the MER field and its evolution over time. Thus, they serve as a major reference to obtain a deeper understanding of the field.

1988: Classification using symbolic files

During almost two decades, emotion was represented in MER studies as a categorical problem. To the best of our knowledge, the first work in the area was published in 1988 by Katayose et al. (1988). There, audio music files were first manually transcribed to

notes and the author performed sentiment analysis from the resulting symbolic files. Features related to melody, chords, key and rhythm were used to estimate the emotion with heuristic rules. Given the complexity of dealing with raw audio signals, several other works also used symbolic representations (Livingstone & Brown, 2005; M. Wang, Zhang, & Zhu, 2004; Yeh, Tseng, Tsai, & Weng, 2006).

As an example, in 2004 Wang et al. (2004) proposed a music emotion recognition system using MIDI files, based on a categorical model of emotion derived from Russell's dimensional model of AV[101]. To this end, the authors define three terms for the valence axis (happy, neutral and anxious) and two terms for arousal (calm, energetic). The combination of these two dimensions generates the six adjectives employed in the study: joyous, robust, restless, lyrical, sober and gloomy. Next, 18 features were extracted from an unknown number of 20-sec MIDI excerpts, mostly of Western tonal music. The features were divided in statistical features and perceptual features. Statistical features include average and standard deviation of absolute pitch, interval, tempo and loudness, as well as note density, instrument number (timbre) and meter. The perceptual features used are tonality, key, mode, a stability score, average and standard deviation of perceptual pitch height and of the perceptual distance between two consecutive notes.

The extracted MIDI features were used with the emotion annotations created by 20 listeners to train a hierarchical SVM classification system. The first level was used to distinguish between calm and energetic (the arousal dimension), while the second further divided the songs in one of the three valence related classes. Although the authors presented very high results (between 62.97% for robust and 85.81% for restless), it is important to note that MIDI excerpts were used and few details are provided on their selection, dimension or genre.

2003: Single-label classification using raw audio files

In the 21st century the field started to gain more attention, linked to factors such as the massification of digital audio formats and Internet access. We believe the first paper on emotion detection in audio was published in 2003 by Feng et al. (2003). In it, the authors proposed a system to classify musical pieces into 4 emotion categories (happiness, sadness, anger and fear). Two musical attributes, tempo and articulation are said to be relevant to the problem. Based on this, three audio features were used – relative tempo, and mean and standard deviation of the average silence ratio, which estimates articulation by representing what percentage of sound in one frame is below the average level (see Section 3.1). The employed dataset contained 223 pieces of modern popular music (few details are given about the selection and duration of the clips). Of these, 200 were used to train a neural network, where only 23 pieces were used to test (less than 7% of

[101] As in other studies, the authors mistakenly mention the Thayer's two-dimensional model as the AV model of emotion.

the dataset). In terms of classification accuracy, three of the four categories (happiness, sadness and anger) obtained high results, between 75% and 86%, with fear reaching only 25%, to a total precision and recall of 66% and 67% respectively. The main limitations found in this study are the very small test corpus (7%), with only 3 songs tagged as fear, not providing evidence of generality, and the usage of only two musical attributes. Additionally, the paper lacks details on how the dataset was annotated.

In 2006, Lie Lu et al. (2006) studied the classification of audio clips into one of the 4 quadrants formed by the AV dimensional model. To this end, the authors compiled a dataset of 800 20-sec music clips (200 per quadrant), from 250 musical pieces (mostly classical). These clips were selected and annotated by three experts, discarding any clip where consensus were not found among the three. Next, three feature sets were extracted: intensity (sound level related), timbre (spectrum related) and rhythm (tempo related). These were used to train Gaussian mixture models in both hierarchical (as in previous works such as (M. Wang et al., 2004), to divide songs into high or low arousal classes) and non-hierarchical schemes. Even though the reported accuracy is very high (86.3%), it should be noted that only classical songs with a strong agreement between all the experts were considered.

Several other works have studied MER as a single-label classification problem. Wu et al. (2006) used 72 10-sec audio excerpts from 4 different (unspecified) genres annotated into the same four quadrants by 60 subjects. From these, 55 low and mid-level musical features were extracted using Marsyas and PsySound audio frameworks and classified using SVM models. The reported accuracy is very high, up to 98.67%, which considering the very low dataset size might indicate overfitting.

In 2008, Hu et al. (2008) published a revision of the submissions to the MIREX AMC 2007 edition (the first AMC edition). Being a single-label music emotion classification problem, served as a first benchmark of the various classification strategies proposed over the previous years. There, 600 30-sec audio clips spanning 27 different genres were used, annotated into one of five clusters by three experts, following the taxonomy proposed in (Hu & Downie, 2007). Several sets of features were extracted from the distinct submissions, mostly low-level descriptors (e.g., tonal, temporal, loudness related), but also some higher-level ones (e.g., danceability) and even symbolic ones such as pitch and note durations. Still, the authors noted that there were no significant differences among these sets and thus "the (high-level) features other than the basic spectral ones did not show any advantage in this evaluation", with an average accuracy of 53% and a maximum of 61%. Such results were much more modest, contrasting with the very high accuracy reported in previous works where very limited datasets might have been used.

Many other works have been published in MER using single-label classification, most following the same principles. These offer variations of the dataset size, annotation process or classification method (e.g., (Panda & Paiva, 2012b; Pao et al., 2008; Schmidt et al., 2010)), explore source separation from the audio signal (e.g., (Xu, Li, Hao, &

Yang, 2014)) or investigates cultural differences in MER (e.g., (Patra, Das, & Bandyopadhyay, 2013)).

2003: Multi-label Classification

In 2003, Li et al. (2003) proposed the first system on emotion detection as a multi-label classification system, thus acknowledging that a musical piece can be described with more than one emotion label. To this end, the ten emotions present in the Farnsworth model (Farnsworth, 1958), a refined and regrouped version of the Hevner's adjectives list, plus three extra emotions added according to a test subject who indexed the test songs, were used.

The musical database for this test was composed of 499 songs, 50% of which were used for training and the remaining 50% used for testing. These songs were selected from 128 music albums (at least four songs each) and the collection covered four major music types: ambient (120 files), classical (164 files), fusion (135 files) and jazz (100 files). From these, acoustic features consisting of timbral texture features, rhythm content features (beat and tempo detection) and pitch content features were extracted and classification was performed with Support Vector Machines (SVM). Due to the multi-label nature of the problem, a binary classifier was trained for each existing category. To determine labels of a test data, the binary classifiers were run independently, selecting every label where the classifier's output exceeds a given threshold. The final results were modest, with an F1-Score of 44.9% (micro average) and 40.6% (macro average). One of the major problems with the paper is related with the dataset, which was annotated by a single subject, selecting various adjectives for each song.

Building on his previous work, Li et al. proposed a new system for emotion detection and also similarity search (Li & Ogihara, 2004) – searching for music sound files similar to a given music sound file. Using the same multi-label classification strategy, a new dataset of 235 Jazz sound files was employed. This dataset was annotated independently by two subjects. Each track was labelled using a scale ranging from -4 to 4 on three bipolar adjective pairs: (Cheerful, Depressing), (Relaxing, Exciting), and (Comforting, Disturbing), where 0 was viewed as neutral. From these, 35 features were extracted, among which Mel-frequency cepstral coefficients (MFCC) and other timbral features such as spectral centroid, rolloff, flux and low energy. Additionally, Daubechies wavelet coefficient histograms (DWCH) were also extracted, calculating the average, variance, skewness and energy for each sub-band.

To solve the emotion detection problem, the multi-label classification strategy previously proposed by the same authors was used. Similarity was measured computing the Euclidean distances between features vectors from each song. Results identified emotion detection, ranging from 70% to 83% accuracy, as a harder problem than similarity meas-

ure, with 86% of accuracy. On the positive side, both emotion detection, using the binary classifiers, and the similarity measure are interesting solutions. However, the same issues, related with the low number of human annotators are still present in the study. Also, the dataset is composed solely of jazz music, with only vocal sound files being used for similarity. Additionally, there is a wide gap in the accuracy for the first two categories between the two subjects, something attributed to cultural differences by the authors.

Other researchers followed the multi-label classification idea, proposing new solutions and testing different classification algorithms. As an example, Wieczorkowska et al. (2006) built on Li et al. (2003) work, approaching it as a multi-label problem, concluding that, despite the low results observed (27.1% with 13 classes and 38.6% with 6 classes), these were "comparable to the results obtained in subjective tests for human listeners". Given this, the authors concluded that using a small set of emotions that can be graded on a scale (either continuous or discrete) might be generally better. Further studies proposed novel strategies, such as hierarchical Bayesian models (B. Wu, Zhong, et al., 2014), random k-labelsets (Trohidis et al., 2008), multi-label back-propagation (Trohidis et al., 2011) and others.

Building on the existent multi-label classification studies, some authors proposed fuzzy approaches to further address the subjectivity in emotion perception. One of the first was Yi-Hsuan Yang et al. (2006), proposing a system that outputs fuzzy vectors to classify song emotions according to the four quadrants of Russell's model[102], e.g., (0.1, 0.0, 0.8, 0.1) representing a song with class 3 as the strongest emotion. To this end, 243 25-second segments from popular Western, Chinese and Japanese songs were selected and annotated into one of the four classes by subjects. This set was then reduced to 195 segments by eliminating ambiguous ones. Two different classifiers were tested, fuzzy k-nearest neighbors (FKNN) and fuzzy nearest-mean (FNM), and validated with 10-fold cross-validation, with FKNN achieving the best results – 78.33% accuracy. Other classifiers such as fuzzy support vector machines (FSVM) have also been used in similar problems (Myint & Pwint, 2010).

2003: Music emotion variation detection

Still in 2003, Liu et al. (2003) published one of the first MEVD works while studying hierarchical versus non-hierarchical approaches to emotion detection. The study focused on classical music, one of the reasons being that emotion changes over a musical piece are empirically more frequent in such genre. Emotions were classified using Gaussian mixture models (GMM) into four classes based on the quadrants of Thayer's emotion model, where music intensity was linked to energy and both timbre and rhythm mapped to the stress component.

[102] The original paper cites the Thayer's model but we believe Russell's model is the correct name since they use arousal and valence instead of tension arousal and energy arousal.

The database used for evaluation consisted of 250 classical musical pieces, split into 20-second clips, 75% of which were used for training and 25% for testing. From these, features such as root mean square value in each sub-band (for intensity), spectral shape features like centroid, rolloff and spectral flux (for timbre) and a Canny estimator, used to detect beat (for rhythm) were extracted. The resulting data was then used with two distinct approaches, hierarchical and non-hierarchical. The non-hierarchical framework used a single GMM combining the four emotion clusters. It received results from all extracted features from the three sets (intensity, timbre and rhythm), returning the calculated results. On the other hand, the hierarchical framework used several GMMs. Each GMM was built using a set of features regarding each emotion cluster, organized in three layers. As an example, a song classified in GMM1 as having low intensity will be either contentment or depression, descending to layer two it will be classified by GMM2 (using timbre features) and GMM3 (using rhythm). The results are summed and one of the two emotions is chosen (contentment or depression). Even though more complex, the hierarchical framework gives better results and makes a better use of sparse training data, important when the available training data is limited. As for MEVD, the approach proposed by the authors tries to find potential emotion change boundaries in the entire segment, instead of using a sliding window and classifying each segment. This method consists of a two-step MEVD scheme. First, the goal is to find potential boundaries, resorting to intensity (using the intensity outline to detect possible boundaries), timbre and rhythm (to check for possible emotion changes in possible boundaries). Then, the musical clip is divided into several independent segments, each containing a constant emotion.

In 2006, the same authors proposed an improved version of their hierarchical and non-hierarchical systems (L. Lu et al., 2006), increasing the feature set with additional features such as MFCCs, rhythm strength, rhythm regularity, and tempo. The results showed an average precision on emotion detection of 86.3%, with average recall of 84.1%. In MEVD tests, the results showed that about 84.1% of the boundaries are recalled and the precision is about 81.5%. Although the results were high, it is important to note that only 800 classical music clips were used and thus the results might not generalize to bigger datasets containing different genres. One of the most interesting aspects of this study is the chosen feature set.

MEVD can also be regarded as an extension of music emotion classification, where the entire segment is further divided into per-second (or other unit) subsegments. Hence, other authors have also applied their solutions to it (e.g., (Panda & Paiva, 2011b; Y.-H. Yang et al., 2006)). As an example, Yang et al. (2006) tested (as a proof of feasibility) their fuzzy classification technique with MEVD in segments of 10 seconds with 1/3 overlap, then proposing equations to translate the fuzzy vectors into arousal and valence.

2004: Multi-Modal Music Emotion Recognition

A different approach to the emotion detection problem was proposed in 2004 by Dan Yang et al. (2004). The paper proposes a strategy for emotion rating to assist human annotators in the music emotion annotation process. In addition to acoustic data, used to extract emotion intensity information, the authors propose the use of song lyrics to distinguish between emotions, by assessing valence. In the acoustical experiments, 500 songs were used to extract several audio features, namely beats per minute (BPM), 12 low-level standard descriptors from the MPEG-7 standard, four timbral features (spectral centroid, rolloff, flux and kurtosis) and 12 features generated by Sony Extractor Discovery System's genetic algorithm (Pachet & Zils, 2003). The same songs were labelled by a human listener according to their intensity (between 0 and 9) and used to train an SVM regressor. The results, according to the authors, achieved almost 0.90 of correlation and BPM, sum of absolute values of normalized Fast Fourier Transform (FFT) and spectral kurtosis were the best features. Similar conclusions were previously attained by Dan Liu et al. (2003) where emotional intensity was highly correlated with rhythm and timbre features.

Following this, lyrics were analyzed in order to differentiate between emotions. To this end, 152 30-second audio clips of alternative rock were labelled into emotion categories (of the PANAS-X schedule (D. Watson & Clark, 1999), authors of the Tellegen-Watson-Clark model) by a single volunteer, of which only 145 had lyrics. Lyrics text files were transformed into 182 psychological features using an approach named General Inquirer[103] (GI) (Stone, Dunphy, Smith, & Ogilvie, 1966) and SVM classifiers were used. Fusing together both acoustical and text approaches resulted in a small increase in accuracy, from 80.7 to 82.8%. Although the results were interesting, the quality of the dataset raises some issues since a single volunteer annotated the 182 30-sec clips which are of a single genre (alternative rock). Some interesting results about the lyrics, such as the sadness being strongly related with positive words such as "love", "life" and "feel" came up. This higher correlation of sadness to positive affect is predicted in the emotion model in use. While the increase in accuracy was very low, the analysis of lyrics seems interesting to differentiate between negative emotions, where there is a higher ambiguity.

In the same direction, Laurier et al. (2008) used audio and lyrics to classify emotion into four non-exclusive binary classes: angry, happy, sad and relaxed (i.e., a song can be "happy" or "not happy" but also "angry" or "not angry"). To this end, a large set of songs containing related tags were pre-selected from Last.FM and validated by at least one listener. Several audio features were extracted from the 1000 30-sec audio clips. Namely, timbral (e.g., MFCCs and spectral centroid), rhythmic (e.g., tempo, onset rate), tonal (e.g., HPCP) as well as temporal descriptors. As for lyrics, several strategies were tested.

[103] "A unique set of procedures for identifying, in a useful and meaningful way, recurrent patterns within the rich variety of man's written and spoken communications" (Stone et al., 1966).

First, using the similarity between songs by reducing lyrics using bag-of-words (BOW), which can be briefly described as the set of words used in the lyrics as well as their frequency. These values are then compared using the term frequency–inverse document frequency (TF.IDF) statistics, which indicates how relevant a word is to a document (i.e., a song) in the dataset. The second method used latent semantic analysis (LSA) in combination with TF.IDF, which like other reduction algorithms such as PCA works by projecting the input data into a space of a given dimensionality. Finally, the authors explored language model differences, which consists in analyzing the language models corresponding to the different categories (e.g., the most frequent terms of "happy" songs, compared to the ones in "not happy").

The two sources of information were first tested independently using different classification algorithms (SVM, Random Forest and Logistic Regression), with the authors stating that SVMs performed better than the remaining. The results for audio ranged from 81.5% (happy binary classifier) to 98.1% (angry), while the best models using lyrics achieved 77.9% (angry) to 84.4% (sad). When combining both feature sources the results improved, ranging from 86.8% (happy) and 98.3% (angry), while reducing the observed standard deviation of the accuracies between folds. While the results show benefits of the multi-modal approach, it should also be noted that the performance using audio was already extremely high (especially with the "angry" class) and that the listener validation might be biased towards the audio part, since it was based on 30-sec clips which do not contain the full lyrics.

Three additional works on multi-modal MER were published by Hu et al. between 2009 and 2010 (Hu & Downie, 2010a, 2010b; Hu, Downie, & Ehmann, 2009). In these, the authors used a larger dataset of 5585 audio clips and lyric text files representing full songs. The dataset was annotated using a categorical model of 18 classes containing a total of 135 tags extracted from Last.FM tags and filtered with the WordNet-Affect[104] and "two human experts". The dataset was later reduced in the 2010 experiments to 5296 songs in order to balance the positive and negative emotion classes. From the audio files, the authors extract 65 spectral features (e.g., MFCCs, spectral centroid, spectral rolloff and spectral flux) using the Marsyas audio framework. Regarding the lyrical information, it was summarized using a myriad of descriptors related with BOW, part of speech (POS) (i.e., lexical items which have similar grammatical properties), GI and n-grams (sequence of n items). These features were combined and tested using multi-label SVM classifiers. Some of the most interesting results from the study are the fact that lyrics were able to significantly outperform audio for specific emotions (7 of the 18 categories): "aggressive", "angry", "anxious", "cheerful", "exciting", "hopeful" and "ro-

[104] WordNet-Affect is a linguistic resource for the lexical representation of affective knowledge (Strapparava & Valitutti, 2004).

mantic". Moreover, a more detailed analysis on the most significant lyrical features, provided in their last paper (Hu & Downie, 2010b), indicates a "strong and obvious semantic association between extracted terms and the categories".

Several other researchers have investigated the combination of multiple sources of information for MER (Malheiro, Panda, Gomes, & Paiva, 2013; Mcvicar & Freeman, 2011; B. Wu, Zhong, et al., 2014; Y.-H. Yang, Lin, Cheng, et al., 2008) or proposed new datasets with varied data sources, including symbolic versions of the songs (Panda, Malheiro, et al., 2013) or even biosignals (e.g., heart rate and skin conductance (Coutinho & Cangelosi, 2011); EEGs and facial expression videos (Koelstra et al., 2012)). Namely, we proposed a dataset combining audio, lyrics and symbolic (MIDI) versions of songs, which we used to demonstrate that combining descriptors from all sources lead to improvements in the classification accuracy (Panda, Malheiro, et al., 2013). The same problem was later approached using a smaller and more controlled dataset (annotated manually by multiple volunteers instead of relying on online sources) (Malheiro et al., 2016a), this time combining a myriad of audio features available from state-of-the-art audio frameworks (e.g., Marsyas, MIR Toolbox and PsySound3), with novel lyrical features proposed by Malheiro et al. (2018) and which proved to be emotionally-relevant.

2006: Dimensional approaches using regression

To the best of our knowledge, the first work to use dimensional models in the machine learning stage was published in 2006 by Korhonen et al. (2006) although using only 6 classical music segments. Before, dimensional emotion models such as the AV space had only been used in the annotation collection phase (T.-L. Wu & Jeng, 2006), after which the classification steps transformed the ground-truth into discrete categories (e.g., 4 classes).

In 2007, Yang et al. proposed one of the first works to view MER as a regression problem where emotions are represented in a dimensional taxonomy (Y.-H. Yang, Lin, et al., 2007). An improved version of it was released the year after (Y.-H. Yang, Lin, Su, et al., 2008). In this work, Russell's emotion model, consisting of continuous arousal-valence values, was used, thus enabling the differentiation of the music clips with similar or close emotional states, based on their proximity. A solution for MEVD by predicting the arousal and valence values of each music sample over intervals of time was also theorized.

As the ground-truth for this work, the authors used a dataset of 195 songs which was previously employed in their fuzzy classification system (Y.-H. Yang et al., 2006). As described, it consisted in 195 25-second clips, segmented by experts from Western, Japanese and Chinese albums (mainly pop and rock). According to the previous study, the dataset was balanced between quadrants. The ground truth was created with recourse to

253 volunteers with different backgrounds, in a subjective test where each clip was labelled by at least 10 different subjects. The volunteers were asked to annotate the evoked emotion in AV values, between [-1, 1] in 11 ordinal levels, i.e., {-1.0, -0.8, -0.6, ..., 1.0}, and to consider audio, lyrics and the singers' voice acoustics.

From the dataset, a total of 114 features were extracted, including spectral contrast (12 features), DWCH (28 features), many features available in PsySound2 (44 features), comprising loudness, level, pitch multiplicity and dissonance and also Marsyas features (30 features), including timbral texture, rhythmic content and pitch content. These were reduced to a subset with projection and selection algorithms such as PCA and RReliefF, and used with three distinct regression algorithms: Multiple Linear Regression (MLR), Support Vector Regression (SVR) and AdaBoost.RT (BoostR). There, SVR attained the best results.

Although the paper made a significant contribute to the MER field, namely by proposing regression as a solution for MER and MEVD, the obtained results were somewhat weak, especially for valence. The best solution, measured using R^2 statistics, reached 58.3% for arousal and 28.1% for valence and was obtained using the principal component space (resulting from applying PCA to the feature space) and RReliefF selection of features (18 features for arousal and 15 for valence). Also, according to the authors, the five features that performed best in arousal prediction were flux (standard deviation), tonality, multiplicity, flux (mean) and roll off (mean). As for valence, the five best were spectral dissonance, tonality, sum of the beat histograms, chord and sum of the pitch histogram. A deeper analysis of this paper uncovered a few problems with the ground truth, as detailed in Section 3.2.2 (under Yang's AV dataset) and in Section 5.1.1.

In the following years, several other authors proposed regression solutions to MER, e.g., (Schmidt et al., 2010). However, such approaches still fail to fully account for the ambiguity and subjectivity of human annotations. After all, generally the ground-truth is built by averaging the subjects' annotations for each song, condensing it in a single point of the continuous space. Looking back, previous categorical studies suffer from this when a single label is used.

To overcome the limitations of classical regression approaches, solutions based on probability distribution (provided by fuzzy classification) give more realistic representations of reality. In this direction, Schmidt and Kim (2010a) noted that, instead of a single AV point, emotion perceived by listeners is better represented as a distribution of ratings, more precisely, most annotations can be "well represented by a single two-dimensional Gaussian distribution" (Schmidt & Kim, 2010a). To this end, the authors proposed a system for music classification and MEVD that models the data as a two-dimensional Gaussian, predicting the distribution parameters from the acoustic content. Further studies built on this idea, experimenting with alternative classification schemes and datasets (e.g., (Y.-A. Chen, Wang, Yang, & Chen, 2014; J.-C. Wang, Yang, Jhuo, Lin, & Wang, 2012; J.-C. Wang, Yang, Wang, et al., 2012)).

A summary of the most relevant papers in the area, in chronological order, is presented in Table 3.4.

Paper	Dataset		Emotion model		MER problem		Annotations	Feature Extraction	Learning Model
	Type of data	Size	Taxonomy	Emotion Type	Classification Type	Static MER vs. MEVD			
(Katayose et al., 1988)	Symbolic.	3 piano songs presented in the results table, 1 used in the example figure.	Categorical (e.g., gloomy, urbane, pathetic, serious, hopeful and others).	Perceived.	Multi-label.	Static MER and MEVD.	None.	Related with melody, rhythm pattern, chord recognition, key recognition.	Heuristic rules defined by the authors (e.g., "Key F → Rural mood".
(Feng et al., 2003)	Audio.	223 musical pieces.	Categorical (4 classes: happy, sad, angry and fearful).	Perceived.	Single-label.	Static MER.	None, used rules derived by (Juslin, 2000).	Tempo and Articulation (mean and standard deviation of ASR).	Back propagation neural network.
(Li & Ogihara, 2003)	Audio.	499 clips from 128 albums (30-sec duration).	Categorical (13 adjective groups / 6 super-classes).	n/d.	Multi-label.	Static MER.	1 annotator.	30 features extracted using Marsyas. Related with timbre, rhythm and pitch.	SVM.
(Li & Ogihara, 2004)	Audio.	235 excerpts from 80 Jazz instrumental albums (30-sec).	Categorical (3 bipolar adjective pairs: cheerful, depressing; relaxing, exciting; and comforting, disturbing).	n/d.	Multi-label.	Static MER.	2 annotators labelled all clips (2 annotations per clip).	35 features, namely Daubechies wavelets coefficient histograms (DWCH) and timbral features, (e.g., MFCCs and Spectral moments, using Marsyas).	SVM.
(M. Wang et al., 2004)	MIDI.	n/d number of excerpts, mainly from western tonal music (20-sec).	Categorical (6 adjectives: joyous, robust, restless, lyrical, sober and gloomy).	n/d.	Single-label.	Static MER.	20 different annotators.	18 features (pitch, tempo, interval, loudness, note density, timbre, meter, tonality and other perceptual).	SVM (2 levels, first level divides songs in tranquil or energetic).
(D. Yang & Lee, 2004)	Audio and lyrics.	Set 1: 500 (20-sec) alternative rock clips. Set 2: 152 (30-sec) alternative rock clips (145 contained lyrics).	Dimensional and categorical (PANAS-X, which uses 2D linked to discrete emotions.	n/d.	Regression.	Static MER.	1 volunteer.	Several audio, e.g., beats per minute, 12 MPEG-7 low-level descriptors, timbral (spectral) descriptors and 12 features from Sony evolutionary extractor discovery system. 182 features using General Inquirer.	SVM regression (SVR).
(Leman, Vermeulen, de Voogdt, Moelants, & Lesaffre, 2005)	Audio.	60 excerpts (30-sec).	Dimensional (15 bipolar adjectives transformed to 3D space of AV and	Perceived.	Single-label.	Static MER.	8 subjects.	7 features (related with loudness, roughness, spectral centroid, onsets, bandwidth, articulation and pitch).	Linear regression models.

			Interest).						
(Tolos, Tato, & Kemp, 2005)	Audio.	34 segments of unknown duration, later reduced to 14. Tested using a different set of 616 pop songs.	Categorical (3 classes: happy, aggressive and melancholic + calm), from 2D AV model.	Perceived.	Single-label.	Static MER.	10 subjects.	27 spectral features (e.g., centroid, rolloff, flux, cepstral coefficients reduced using linear discriminant analysis).	Multivariate normal distribution estimates.
(Wieczorkowska, Synak, Lewis, & Raś, 2005)	Audio.	303 musical pieces.	Categorical (13 classes / 6 superclasses).	Perceived.	Single-label.	Static MER.	1 subject.	Mostly spectral metrics, e.g., tristimulus, even/odd-harm, frequency, brightness, irregularity.	KNN.
(Wieczorkowska et al., 2006)	Audio.	875 excerpts (30-sec).	Categorical (13 classes / 6 superclasses).	Perceived.	Multi-label.	Static MER.	1 subject ("database created by Dr. Rory A. Lewis").	29 timbral features (e.g., frequency, level, tristimulus, even-harm, oddharm, brightness, regularity).	KNN.
(L. Lu et al., 2006)	Audio.	800 excerpts (20-sec) from 250 musical pieces.	Categorical (4 classes from 2D AV model).	Perceived.	Single-label.	Static MER and MEVD.	3 experts (songs without consensus were ignored).	Several, related with intensity (sound level), timbre (spectrum), and rhythm (tempo). Reduced using KL transform.	GMM (hierarchical and non-hierarchical approaches).
(Y.-H. Yang et al., 2006)	Audio.	195 excerpts (30-sec).	Categorical (4 classes from 2D AV). Dimensional (from fuzzy vectors to AV).	Perceived.	Multi-label (fuzzy).	Static MER and MEVD.	n/d number of subjects (excerpts without consensus were ignored).	15 features from PsySound2 (spectral centroid, Loudness, sharpness, timbral width, volume, spectral and tonal dissonance, pure and complex tonal, multiplicity, tonality, chord). Reduced with SBS.	FKNN and FMN (fuzzy multi-label).
(T.-L. Wu & Jeng, 2006)	Audio.	75 excerpts (10-sec) of 4 different (unspecified) genres.	Categorical (4 classes from 2D AV).	Perceived.	Single-label.	Static MER.	60 subjects (using a 2D AV space, 66.2% agreement in 4-class).	55 low and middle-level musical features (using Marsyas and PsySound). Reduced using multivariate analysis of variance.	SVM.
(Korhonen et al., 2006)	Audio.	6 Western art musical pieces.	Dimensional (2D AV space).	Perceived.	Regression.	MEVD.	35 (21 male, 14 female).	18 features (using PsySound and Marsyas, custom tempo algorithm).	Autoregression with extra inputs (ARX) and space-state model.
(Skowronek, McKinney, & Par, 2007)	Audio.	1059 excerpts from 12 genres (n/d duration).	Categorical (12 classes, e.g., sad, peaceful, tender-soft, angry-furious).	Perceived.	Multi-label.	Static MER.	12 subjects (6 annotations per excerpt on a 4-point scale).	n/d total (4 types: "signal describing features", tempo and rhythm, Chroma and key, percussive sound events).	Quadratic classifier / quadratic discriminant analysis (QDA).

(MacDorman et al., 2007)	Audio.	100 excerpts (6-sec).	Dimensional (2D AV space, initially 3D but dominance was discarded).	Unclear.	Regression.	Static MER.	85 participants.	5 representations from MA Toolbox (MFCC, sonogram, fluctuation, spectrum and periodicity histograms).	Linear regression, "such as least-squares regression".
(Meyers, 2007)	Audio and lyrics.	372 songs.	Categorical (8 classes based on the updated Hevner's adjectives list by Schubert, placed in the 2D AV model by Russell.)	n/d.	Single-label.	Static MER.	None, since a decision tree is used built using information from (Gabrielsson & Lindström, 2001).	Audio: 5 features related with mode, harmony, tempo, rhythm and loudness. Lyrics: Features provided by ConceptNet's "guess_mood".	Decision tree and KNN.
(Y.-H. Yang, Su, et al., 2007)	Audio.	MER60: 60 excerpts (25-sec).	Dimensional (2D AV space).	Perceived.	Regression.	Static MER.	99 subjects, 40 annotations per song.	45 features (PsySound2 and Marsyas): spectral centroid, loudness, sharpness, timbral width, volume, spectral and tonal dissonance, multiplicity, tonality, and chord.	SVR (extension of SVM for regression).
(Y.-H. Yang, Lin, et al., 2007; Y.-H. Yang, Lin, Su, et al., 2008)	Audio.	YangAV: 195 excerpts (25-sec). MEVD tested using 6 songs from (Korhonen et al., 2006).	Dimensional (2D AV space).	Induced.	Regression.	Static MER and MEVD.	253 subjects, rated in 11 ordinal AV levels, 10+ ratings per song.	114 features (PsySound, Marsyas, DWCH and spectral contrast). PCA and ReliefF tested for reduction and selection.	SVR (best results), BoostR and MLR.
(Y.-H. Yang, Lin, Cheng, et al., 2008)	Audio and lyrics.	1240 Chinese pop song excerpts (30-sec) and full lyrics.	Categorical (4 classes from 2D AV).	n/d.	Single-label.	Static MER.	The only information provided is that "emotions are labeled through a subjective test".	Audio: 106 features from Marsyas (MFCC) and PsySound (spectral centroid, moments and roughness). Lyrics: 8100 features related with BOW (4000 uni-gram and 4000 bi-gram) and probabilistic LSA (100).	SVM.
(Bartoszewski et al., 2008)	MIDI.	104 musical pieces.	Unsupervised, dimensional model used to validate results (2D QA – quality and activation).	n/d.	n/d.	n/d.	3 subjects annotated 70 of the obtained segments to evaluate the results.	8 features: music scale, accuracy, sound intensity, basic sound, interval, direction, velocity, duration of notes.	Unsupervised, using agglomerative clustering; visualized with self-organizing map (SOM) neural network.
(Hu et al., 2008)	Audio.	600 excerpts (30-sec).	Categorical (MIREX AMC: 5	Perceived.	Single-label.	Static MER.	3 experts, 2 to 3 annotations per clip.	Various tested: spectral, tem-	SVM and KNN.

			clusters with 29 adjectives).					poral, tonal, high-level (danceability) and symbolic (e.g., note durations).	
(Laurier et al., 2008)	Audio and lyrics.	1000 songs (30-sec excerpts were used for manual validation by listeners).	Categorical (4 classes from 2D AV: happy, sad, angry and relaxed).	Perceived.	Multi-label.	Static MER.	From Last.FM tags, manual validation by at least one listener.	Audio: timbral, rhythmic, tonal and temporal descriptors. Lyrics: TF.IDF, LSA and language model differences.	SVM (best results), Logistic, RandForest.
(T.-L. Wu & Jeng, 2008)	Audio.	1200 (5-sec) segments extracted from 200 (30-sec) soundtrack clips.	Categorical (8 classes based on Hevner updated model).	n/d.	Single- and multi-label.	Static MER.	328 subjects, 28.2 annotations per song (online)	88 features (PsySound, Marsyas, DWCH and spectral contrast).	SVM.
(Trohidis et al., 2008)	Audio.	593 excerpts (30-sec).	Categorical (6 classes based on Tellegen-Watson-Clark model).	Perceived.	Multi-label.	Static MER.	3 expert annotators, only songs with full agreement.	72 features (8 rhythmic and 64 timbral using Marsyas, e.g., MFCCs, centroid, rolloff, flux and beat histogram statistics)	SVM based (binary relevance, label powerset, random k-labelsets) and KNN (multi-label).
(Pao et al., 2008)	Audio.	MER60: 60 excerpts (25-sec).	Categorical (4 classes from 2D-AV).	Perceived.	Single-label.	Static MER.	99 subjects, 40 annotations per song.	45 features (15 from PsySound and 30 from Marsyas similar to (Y.-H. Yang, Su, et al., 2007)).	Weighted-discrete KNN, KNN, SVM.
(Y.-H. Yang, Lin, & Chen, 2009)	Audio.	MER60: 60 excerpts (25-sec).	Dimensional (2D AV space).	Perceived.	Regression.	Static MER.	99 subjects, 40 annotations per song.	45 features (15 from PsySound and 30 from Marsyas similar to (Y.-H. Yang, Su, et al., 2007)).	SVR (extension of SVM for regression).
(Lin, Yang, Chen, Liao, & Ho, 2009)	Audio.	1535 songs from 300 albums, across 6 genres: blues, country, jazz, R&B, rap, and rock.	Categorical (12 classes by clustering AllMusic emotion tags).	n/d.	Multi-label.	Static MER.	Extracted from AllMusic and clustered into 12 classes.	Genre used as feature, plus 436 audio features (from Marsyas, 68 timbral textual features, 48 pitch content, 8 rhythmic, 120 LPCCs, 192 MPEG-7).	SVM scheme with two levels, 1st level predicts genre, 2nd for emotion.
(Han, Rho, Dannenberg, & Hwang, 2009)	Audio.	165 western pop songs.	Categorical (11 classes from Thayer's model).	n/d.	Single-label and regression (transformed into classes).	Static MER.	15 songs per class were selected from AllMusic.	7 features, related with scale (e.g., key, tonality), intensity, rhythm (e.g., tempo) and harmonic distribution.	SVM, SVR and GMM.
(Hu et al., 2009)	Audio and lyrics (text).	5585 full songs from several collections (e.g., US pop, Beatles, metal music).	Categorical (18 categories containing 135 tags, obtained from Last.FM).	n/d.	Multi-label.	Static MER.	From Last.FM tags, filtered with WordNet-Affect and "two human experts".	Lyrics: Many related with Bag-of-words (BOW), part-of-speech (POS), function words. Audio: 63 spectral features such as	SVM.

								MFCCs, spectral centroid, rolloff and flux (Marsyas).	
(Schmidt & Kim, 2009)	Audio.	MTurk240 subset of 120 songs (15-sec).	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	From MoodSwings GWAP, few details are given.	MFCC, spectral shape, contrast and chroma features.	Least-squares regression.
(Laurier et al., 2009)	Audio.	Soundtrack dataset: 110 excerpts from film soundtracks (mean duration of 15.3-sec).	Categorical (5 classes – fearful, angry, happy, sad, tender).	Perceived.	Regression.	Static MER.	116 listeners, rated on a 7-point scale.	200 features (such as timbral, rhythmic, tonal, dissonance, mode, loudness).	SVM.
(Eerola et al., 2009)	Audio.	Soundtrack dataset: 110 (mean duration of 15.3-sec).	Dimensional (3D: activity, valence, tension).	Perceived.	Regression.	Static MER.	116 listeners.	29 features related with timbre, harmony, register, rhythm, articulation and structure (from MIR Toolbox).	Multiple linear regression (MLR) and partial least-squares regression (PLS).
(Hu & Downie, 2010a, 2010b)	Audio and lyrics.	5296 songs based on (Hu et al., 2009).	Categorical (18 categories containing 135 tags, obtained from Last.FM).	n/d.	Multi-label.	Static MER.	From Last.FM tags, as previously described (Hu et al., 2009).	Audio: 65 spectral features using Marsyas (e.g., MFCC, spectral centroid, rolloff, flux). Lyrics: several, such as GI, BOW, n-grams.	SVM.
(Schmidt & Kim, 2010a)	Audio.	MTurk240: 240 clips (15-sec).	Dimensional (2D AV).	Perceived.	Regression (probability distribution / 2D Gaussian).	Static MER and MEVD.	From MoodSwings GWAP, few details are given.	Several related with MFCCs, Chroma, statistical spectrum descriptors (SSD) and spectral contrast.	Multi-variate parameter regression methods such as MLR, PLS and SVR.
(Schmidt & Kim, 2010b)	Audio.	MTurk240: 240 clips (15-sec).	Dimensional (2D AV).	Perceived.	Regression (probability distribution / 2D Gaussian).	MEVD.	From MoodSwings GWAP, few details are given.	Octave based spectral contrast (OBSC).	MLR, linear dynamical system (LDS) and Kalman/RTS smoothing.
(Schmidt et al., 2010)	Audio.	MTurk240: 240 clips (15-sec).	Categorical (4 quadrants from 2D AV) and dimensional (2D AV).	Perceived.	Single-label and regression.	Static MER and MEVD.	From MoodSwings GWAP, few details are given.	Several related with MFCCs, Chroma, SSD and spectral contrast.	SVM, LSR and SVR.
(Y.-H. Yang & Chen, 2011b)	Audio.	Set1: 60 English pop songs (30-sec). Set2: 1240 Chinese pop songs (30-sec).	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	Set1: 40 subjects per song. Set2: 4.3 per song.	157 features related with melody (10), timbre (142) and rhythm (5), extracted with Marsyas, MIR Toolbox and MA Toolbox.	SVM and ListNet (neural networks)-

(J. Kim et al., 2011)	Audio.	446 music clips (20-sec) of several genres (e.g., rock, hip-hop, jazz, metal, dance, country).	Categorical (8 classes from 2D-AV) and dimensional (AV).	n/d.	Single-label classification / clustering.	Static MER.	Emotion tags and AV values from 10 subjects.	None, based on clustering of AV values annotated by 10 subjects.	K-means clustering algorithm.
(Panda & Paiva, 2011b)	Audio.	Train set: Yang195 dataset, only 189 clips were used (25-sec). Test set: 29 full songs.	Categorical (4 classes from 2D AV).	Perceived.	Single-label and regression.	MEVD.	Train set: 253 volunteers, at least 10 annotators per song. Test set: 2 subjects.	172 audio features mostly spectral (63 from MIR Toolbox and 109 from Marsyas).	SVM (SVC and SVR).
(Panda & Paiva, 2011a)	Audio.	Yang195 dataset, only 189 clips were used (25-sec).	Dimensional (2D AV).	Induced (since the Yang195 dataset was used).	Regression.	Static MER.	253 volunteers, at least 10 annotators per song.	458 features (e.g., related with timbre, dynamics, rhythm and harmony – 44 PsySound2, 177 MIR Toolbox, 237 Marsyas).	SVM (SVR).
(Schmidt & Kim, 2011)	Audio.	240 clips (15-sec), subset of MTurk240.	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	From MoodSwings GWAP, few details are given.	Directly from magnitude spectra. MFCCs, spectral contrast, chroma, SSD and EchoNest timbral features for comparison.	Deep belief networks (DBN).
(Coutinho & Cangelosi, 2011)	Audio and biosignals.	9 musical pieces (classical), heart rate and skin conductance.	Dimensional (2D AV) and categorical (4 classes).	Induced.	Single-label and regression.	MEVD.	AV annotations and biosignals by 39 subjects (1 Hz).	Six audio features: loudness, tempo, pitch level (power spectrum centroid), melodic contour, timbre (sharpness), and texture (multiplicity). Two biosignals: heart rate and skin conductance.	Recurrent neural networks and linear discrimant analysis.
(J.-C. Wang, Yang, Wang, et al., 2012)	Audio.	MER60: 60 clips (30-sec). MTurk240: 240 clips (15-sec).	Dimensional (2D AV).	Perceived.	Regression.	Static MER and MEVD.	MER60: 40 subjects per song. MTurk240: 7 to 23 subjects per song using MTurk workers.	MER60: 70 features (dynamic, spectral, timbre and tonal using MIR Toolbox 1.3). MTurk240: no audio available, authors used the provided features (e.g., MFCCs, chroma, SSD and spectral contrast).	GMM.
(Panda & Paiva, 2012b)	Audio.	Multi-modal MIREX-like dataset: 903 clips (30-sec) belonging to 29 emotion tags from	Categorical (MIREX AMC taxonomy: 5 clusters, 29 emotion tags).	n/d (no information provided by AllMusic).	Single-label.	Static MER.	Annotations extracted directly from AllMusic emotion tags.	253 features (65 from Marsyas, 177 from MIR Toolbox, 11 from PsySound3).	SVM.

		AllMusic.							
(Y. Song, Dixon, & Pearce, 2012)	Audio.	2904 audio excerpts (either 30 or 60-sec) from 7digital.com.	Categorical (4 classes: happy, sad, angry and relaxed).	n/d.	Single-label.	Static MER.	From Last.FM social tags.	55 features from MIR Toolbox, summarized with mean and standard deviation (7 dynamics, 5 rhythm, 32 spectral, 10 harmony).	SVM.
(Soleymani et al., 2013)	Audio.	MediaEval: 1000 clips from the 2014 subset (45-sec, eight gen- res: blues, electronic, rock, classical, folk, jazz, country, and pop).	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	10+ annotators per song using MTurk workers.	Several: MFCCs (Rastamat toolbox), OBSC, SSD, chromagram. Additional EchoNest API features.	MLR.
(Patra et al., 2013)	Audio.	250 Hindi music clips (30-sec).	Categorical (5 clusters, i.e., excited, happy, calm, sad and angry).	n/d.	Single-label.	Static MER.	5 human annotators.	Several, related with rhythm (e.g., strength, regularity, tempo), timbre (e.g., MFCCs) and intensity (e.g., RMS energy).	Fuzzy C-means clustering algorithm (unsupervised).
(Imbrasaite, Baltrusaitis, & Robinson, 2013)	Audio.	MTurk240: 240 clips (15-sec).	Dimensional (2D AV).	Perceived.	Regression.	MEVD.	MTurk240: 7 to 23 subjects per song using MTurk workers.	MTurk240: no audio available, authors used the provided features (e.g., MFCCs, chroma, SSD and spectral contrast).	SVR and continuous conditional random fields (CCRF).
(Panda, Malheiro, et al., 2013)	Audio, lyrics and MIDI.	Multi-modal MIREX-like dataset: 903 audio clips (30-sec), 764 full song lyric files, 196 full song MIDI clips.	Categorical (MIREX AMC taxonomy: 5 clusters, 29 tags).	n/d (no information provided by AllMusic).	Single-label.	Static MER.	Annotations extracted directly from AllMusic emotion tags.	Audio: 275 (PsySound3, MIR Toolbox, Marsyas, and melodic features). Lyrics: 26 (jMIR/jLyrics). MIDI: 320 (jMusic, jSymbolic, MIDI Toolbox).	SVM, KNN, Naive Bayes, decision trees (C4.5).
(Xu et al., 2014)	Audio.	267 songs divided in 4542 clips (15-sec).	Categorical (4 classes – anger, calm, happy and sad from Last.FM tags).	n/d (since Last.FM tags were used).	Single-label.	Static MER.	Extracted directly from Last.FM social tags.	84 audio features (58 related with timbre, 4 intensity, 12 melody and 10 "other", using jAudio).	SVM.
(B. Wu, Zhong, et al., 2014)	Audio, lyrics.	1493 English pop songs.	Categorical (122 classes from AllMusic emotion labels).	n/d.	Multi-label.	Static and MEVD.	Used 122 AllMusic labels, average of 1.85 labels per song.	Audio: used the Million Song Database to obtain audio features. Lyrics: TF.IDF from synced lyrics files (LRC format) obtained	Hierarchical Bayesian models, several other methods also tested (e.g., binary relevance, label

								using Baidu search.	powerset, random k-labelsets).
(Hu & Yang, 2014)	Set1 and 2: audio. Set3: music video (only the audio was used).	Set1: 496 Chinese pop clips (30-sec). Set2: MER60 – 60 English pop clips (30-sec). Set3: DEAP120 – 120 Western video clips (1-minute).	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	Set1: annotated by 3 experts. Set2: 40 non-experts. Set3: 14 to 16 students.	Few details are given: "psychoacoustic features" which "have been used and reported as effective in previous studies".	SVR.
(Y.-A. Chen et al., 2015)	Audio.	AMG1608: 1608 Western music clips (30-sec).	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	643 annotators from MTurk and 22 students.	A total of 72 audio features (40 MFCC related, 17 tonal, 11 spectral and 4 temporal).	GMM and maximum a posteriori linear regression.
(S.-H. Chen et al., 2015)	Audio.	1080 music clips (n/d duration) collected from AllMusic and Last.FM (120 clips per class).	Categorical (9 classes: angry, sad, happy, bored, calm, relaxed, nervous, pleased, and peaceful).	n/d.	Single-label.	Static MER.	From AllMusic emotion tags and Last.FM social tags.	38 audio features (covering rhythm, dynamics, timbre, pitch and tonality extract with MIR Toolbox).	Deep Gaussian process (GP), also SVM and standard GP for comparison.
(Madsen, Jensen, & Larsen, 2015)	Audio.	20 excerpts (15-sec).	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	13 annotators.	56 audio features (12 chroma, 24 related with loudness, and 20 MFCCs).	Gaussian process model with multiple kernel learning.
(Ahsan, Kumar, & Jawahar, 2015)	Audio.	100 clips (30-sec) of various genres (e.g., classical, reggae, rock, pop) – subset of the (Trohidis et al., 2008) dataset.	Categorical (6 classes as used by (Trohidis et al., 2008)).	Perceived.	Multi-label.	Static MER.	3 expert annotators, only songs with full agreement.	72 features (8 rhythmic and 64 timbral using Marsyas, e.g., MFCCs, centroid, rolloff, flux and beat histogram statistics) as in (Trohidis et al., 2008).	SVM, KNN, multi-label KNN (ML-KNN) and max-margin multi-label classification (M3L).
(Malheiro et al., 2016a)	Audio and lyrics.	Bi-modal emotion dataset: 163 audio clips (30-sec), 180 lyrical files (full lyrics), 133 common songs (bi-modal tests).	Categorical (4 classes derived from 2D AV).	Perceived.	Single-label.	Static MER.	Annotated by a total of 39 subjects, 6 to 8 annotations per song (average).	1701 audio (e.g., related with timbre, rhythm, harmony, dynamics and others), 1232 lyrics (e.g., bag-of-words, general inquirer, structural analysis and semantic features).	SVM.
(Hu & Yang, 2017)	Audio.	Set1: MER60 – 60 English pop excerpts (30-sec). Set2: 818 Chinese	Dimensional (2D AV).	Perceived.	Regression.	Static MER.	Set1: 40 non-experts. Set2: 3 experts (3 per clip).	539 features divided in loudness, pitch, rhythm, timbre and harmony (extracted using Chroma Toolbox, MIR	SVR.

		pop excerpts (30-sec). Set3: AMG1608 – 1608 Western clips (30-sec).					Set3: 15 to 32 MTurk workers per clip, 665 total.	Toolbox, PsySound and Tempogram Toolbox).	
(Aljanaki et al., 2017)	Audio.	DEAM: 1802 – 58 full songs and 1744 excerpts (45-sec).	Dimensional (2D AV).	Perceived.	Regression.	Static MER and MEVD.	5 to 10 annotations per song using MTurk workers.	Many from several different teams, using different tools (e.g., Marsyas, MIR Toolbox for MATLAB, PsySound, openSMILE, Essentia, jAudio).	Many (e.g., SVR, Kalman, long-short term memory recurrent neural networks, CCRF, multi-level regression).
(Malik et al., 2017)	Audio.	DEAM (2015 subset): 431 excerpts (45-sec) + 58 full songs.	Dimensional (2D AV).	Perceived.	Regression.	MEVD.	5 to 10 annotations per song using MTurk workers.	260 features (by summarizing 65 low-level time-series descriptors) used as baseline, log Mel-band energies used as raw features.	Convolutional and recurrent neural networks.
(Thammasan, Fukui, & Numao, 2017)	MIDI and EEG.	40 MIDI songs (between 73 and 147-sec).	Dimensional (2D AV).	Induced.	Regression.	MEVD.	Each participant selected 16 songs, created continuous AV and recorded EEG signals while listening.	EEG: 17 features related with fractal dimension. Audio: 37 features related with rhythm (e.g., tempo, attack time and slope), dynamics (e.g., RMS), timbre (e.g., MFCC, ZCR) and harmony (e.g., HCDF) using MIR Toolbox.	SVM.
(Malheiro et al., 2018)	Lyrics.	Set1: Bi-modal emotion dataset: 180 lyrics (full songs). Set2: 771 lyrics extracted from AllMusic.	Categorical (4 classes from 2D AV).	Perceived.	Single-label.	Static MER.	Set1: 39 subjects, 8 annotations per song (average). Set2: from AllMusic emotion tags.	High number of lyrical features, among which: content related features (e.g., bag-of-words), stylistic based, song structure based and semantic based.	SVM.
(Panda et al., 2018)	Audio.	900 excerpts (30-sec).	Categorical (binary and 4 classes derived from 2D AV).	Perceived.	Single-label classification.	Static MER.	Derived from AllMusic tags, validated locally by two subjects.	1255 features related with the 8 musical dimensions (898 from Marsyas, MIR Toolbox and PsySound; 357 novel/manually extracted).	SVM.

Table 3.4: Summary of some the most relevant MER studies over the last three decades.

3.2.6. Explored Problems, Applications and Current Directions

Building on the research on core MER problems described earlier, a variety of different MER approaches and possible applications have also been studied. Some examples are: the automatic creation of playlists and music recommendation based on emotional content and metadata (Aucouturier & Pachet, 2002; Flexer et al., 2008; Panda & Paiva, 2011a); studies on the importance of personalization and context to MER applications (Y.-A. Chen, Wang, Yang, & Chen, 2017); understanding the possible cultural differences and exploring inter-cultural solutions (Hu & Yang, 2017); evaluate the relevance of common audio features to the field (Y. Song et al., 2012); and explore information from sources beyond audio or lyrics, such as video, EEGs and other physiological signals (in emotion induction studies) (Hsu, Zhen, Lin, & Chiu, 2017; Koelstra et al., 2012; Thammasan et al., 2017).

Most, if not all, of these problems are still to be addressed before MER can be usable in real life scenarios. To this end, we believe that there are several research paths that must be explored beforehand. These are briefly presented below, ordered by their importance (from the ones we consider the most urgent to address to the least ones).

Higher-level, emotionally-relevant audio features

A great number of audio features have been proposed over time, many of which were later employed in MER studies. Still, most of these have been developed for other MIR problems and may not directly relate with music emotion. To understand this, Song et al (2012) evaluated the influence of various audio features in emotion classification. First, 2904 clips tagged as "happy", "sad", "angry" or "relaxed" were obtained from Last.FM. Next, the MIR Toolbox was used to extract a total of 54 features, divided into dynamics (7), rhythm (5), spectral (32) and harmony (10). Although the quality of the dataset can be questioned, since it was based on Last.FM social tags and unverified clips from 7digital.com, the results showed that "no single dominant features have been found", and that many of the tested features can be removed without significant losses in accuracy. Furthermore, regarding features' groups, "the used spectral features outperformed those based on rhythm, dynamics, and, to a lesser extent, harmony". However, fusing all features barely increased accuracy (51.9% from spectral only to 53.6%).

Based on these results and on our literature review described earlier (Section 3.1), we believe that novel emotionally-relevant audio extractors are needed. Supporting this is the fact that some of the musical dimensions regarded as relevant to emotion (e.g., musical texture, musical form or expressive techniques) lack audio algorithms to capture this information. Moreover, many of the existent features may not be helpful for MER.

The proposal of novel emotionally-relevant audio features is the main contribution of this thesis, as is discussed in Chapter 4.

Public standard datasets combining categorical and dimensional models

The quality and availability of datasets to MER research has been a major issue for long. The problem is actually composed of two distinct parts: music files and audio annotations. Regarding the audio music files, the existence of copyright restrictions limits the usage and sharing, especially of longer segments, forcing researchers to use private or inadequate datasets.

Concerning the annotations, the difficulties are spread between the complexity of the process, subjectivity of human emotion perception and the resources required to conduct a robust procedure. Classification studies usually divide emotions into different classes without theoretical support. Most MER studies with acceptable results adopt few classes (e.g., 4 to 6), which although useful for the final user, may be of limited interest for massive information retrieval systems. However, using a large number of emotion categories leads to ambiguity problems, where users might have difficulties differentiating them. Regression approaches using dimensional models mitigate this problem. However, such models are far from how we as humans reason about and describe emotions, thus limiting also its application.

Initial steps have been given to address this issue, such as attempting to establish relations between emotional categories and emotional points (J. Kim et al., 2011) and combining ground-truth with both types of data (J.-C. Wang, Yang, Chang, Wang, & Jeng, 2012). Besides, the annotation process is complex, very resource-intensive, and prone to inaccurate data and errors, due to the subjectivity associated to emotion perception. Historically, this has led to one of two possible outcomes: 1) very robust /controlled ground-truth procedure for a small dataset; 2) uncontrolled ground-truth, extracted from online sources (e.g., Last.FM, MTurk or GWAP approaches), but of massive scale. Some proposals has been made to find a compromise and increase accuracy, such as having baseline annotations by experts to filter outliers (Speck et al., 2011), or using a ranking-based emotion AV annotation (ranking pairs of songs), instead of selecting a real value (Y.-H. Yang & Chen, 2011b).

Hence, another contribution of this study is the proposal of a methodology and dataset that we believe to the current stage of MER research, as is described in Section 4.1.

Source separation: audio, voice and lyrics

We know empirically from our personal experiences that not only music but also the voice acoustics and the lyrics being sang have influence in the emotional response. As

previously noted, researchers have already demonstrated that lyrical features (extracted from text) can significantly outperform audio features for specific emotions (Hu & Downie, 2010b). Regarding the voice signal and its acoustic information, researchers have also noted their relevance to emotion, especially to discriminate low arousal musical pieces. As an example, Xu el al. (2014) verified that using only the singing voice can be effective for distinguish between "calm" and "sad" emotions. However, when music signals containing both voices and accompaniments are used this effectiveness is lost. As a result, they concluded that "source separation can effectively improve the performance" of MER systems. Moreover, other authors have studied the emotion in speaking and singing voice (Scherer, Sundberg, Tamarit, & Salomão, 2015), as well as the related voice acoustic features (Eyben, Salomão, Sundberg, Scherer, & Schuller, 2015).

Thus, a possible path to improve MER are better multi-modal solutions which exploit the audio content of both music and voice, as well as the lyrical message. Such solution is still far away, still requiring advances in areas such as source separation and singing voice transcription. Still, some very recent works explored the extraction and alignment of lyrics in music from the audio signal (Gupta, Tong, Li, & Wang, 2018), which is a step in this direction.

This problem was partially approached in this thesis, where features based on the separated voice component (based on the work by (Z.-C. Fan et al., 2016)) were used, as will be described in Section 4.2.

New machine learning techniques

Some machine learning techniques have increased in popularity in the last years. The most notable example is the resurgence of neural network techniques, specifically deep learning, to a myriad of problems, fueled by the improvements in computer processing (e.g., using GPUs)). Some MER studies have already used techniques such as deep belief networks (DBN) (Schmidt & Kim, 2011), deep Gaussian process (deepGP) (S.-H. Chen et al., 2015) or convolutional (CNN) and recurrent neural networks (RNN) (Malik et al., 2017).

In (Schmidt & Kim, 2011), the authors use a DBN to learn features directly from the magnitude spectra. To this end, the authors use a subset of the MoodSwings dataset containing only 240 15-sec clips annotated using AV. Their magnitude spectra are fed to the network, which is said to learn emotion-based acoustic descriptors. To evaluate them, the DBN output is fed directly to a linear regression layer and used to predict AV values. The authors report good results when compared to predictions made by standard audio features sets such as MFCCs, spectral contrast and others. Still, the results are based on the average mean distance between predicted and real AV values using single feature sets (e.g., only MFCCs) vs. the DBN which was fed the entire magnitude spectra. It would be interesting to have results of state-of-the-art MER approaches to compare.

Moreover, the employed dataset is extremely limited in size for a deep learning based approach.

Sih-Huei Chen et al. (2015) proposed a deep learning system using deepGP to classify music into 9 emotion classes. For each class 120 audio clips from AllMusic or Last.FM were obtained (no details are given about their duration). In contrast to the typical deep learning approach, where the raw data (e.g., spectrogram or even the waveform signal) is fed to the network, the authors extracted 15 acoustical features using MIR Toolbox. These features are related with rhythm (e.g., tempo, event density), dynamics (e.g., RMS energy), timbre (e.g., zero-crossing rate, MFCC), and harmony (e.g., mode). Feeding the features as input to the deepGP network achieved higher results (71.3%) than a similar strategy using SVM models (63.0%) and standard Gaussian process models (67.4%). Although interesting, further research needs to be carried with a larger and better annotated dataset.

Finally, a more recent work by Malik et al. (2017) used both convolutional (CNN) and recurrent neural networks (RNN) in a stacked setup for music emotion recognition using the AV space. Here, the MediaEval2015 emotion dataset was used to test the hypothesis of the authors that neural networks could learn information such as first and second order derivatives and statistics from raw data on its own. To this end, the authors used the 65 features (summarized in four statistics to 260 features) of the best performing system to date in the MediaEval evaluation as the baseline, feeding it to the system. Alternatively, a similar system was trained with only raw features, consisting of the log Mel-band energy (16 feature vectors summarized in 64 statistics). The training process was conducted using 431 audio excerpts (30-sec), annotated every 500ms, while the evaluation was conducted on 58 full songs from the same set. While the system using the baseline features achieved the best results (RMSE of 0.202 for arousal and 0.268 for valence), both systems (baseline and raw features) outperformed the previously best results. Moreover, the raw features method used a "significant less amount of parameters" than the others effectively proving the authors hypothesis.

Given the large size of some new datasets (e.g., Million Song dataset), such methods might become valuable to uncover new and interesting relations in MER, since "lack of data tends to limit the outcomes of deep learning research" (Pons et al., 2018). In this direction, a recent paper by Jordi Pons et al. (2018) pointed out that end-to-end learning stacks (in this case for music auto-tagging) benefit from larger datasets and demonstrated that "waveform-based models outperform spectrogram-based ones in large-scale data scenarios" (i.e., 1+ million songs).

This approach was not studied in this thesis, mostly because of the dimension of the created dataset, which is not suited for deep learning.

Chapter 4

A Novel System for Music Emotion Recognition: New Dataset and Audio Features

Music emotion recognition is a very promising field in MIR. Considering the current market directions, with ever growing online multimedia databases and massive adoption of streaming services, there is clearly a need to take MER technology from academia to real life applications. To this end, a number of open questions still present in the MER need to be addressed.

As discussed in the previous chapters, the typical MER solutions have three distinct components: the ground-truth, providing data to explore the relations between music signals and emotional annotations; feature extraction, used to summarize musical signals in manageable sets of descriptors; and classification, where machine learning techniques are used to identify patterns between the extracted descriptors and emotional annotations. Based on our in-depth review of the MER literature, we believe an unbalanced amount of attention has been given to the last component, neglecting the former two. After all, being the last step, however good it may perform, its outcome is always dependent on the quality of the preceding ones.

We believe that significant improvements in the first two steps are key to advances in MER.

Section 4.1. Dataset Construction

Under this perspective, and given the restrictions in existent public datasets, we start by describing our dataset and the methodology used in its construction. We propose a compromise between the resource intensive fully manual annotation processes, which typically leads to limited sized datasets, and fully automatic massive datasets, built from online data at the expense of quality. To this end we fuse both ends of the spectrum, by starting with online annotations from experts and thus significantly relieving the manual annotation process.

Section 4.2. Novel Audio Features

Next, we propose a set of novel higher-level audio features aimed at improving emotion recognition. To this end, we intersect the knowledge acquired in Chapter 2 and Chapter 3, exploring musical dimensions that have been identified as relevant in music emotion but for which audio features are still lacking.

Section 4.3. Feature Extraction and Reduction

The dataset built by us is then used to extract standard audio features with three state-of-the-art frameworks: Marsyas, MIR Toolbox and PsySound3. Next, the novel proposed features are extracted from the same audio clips. Additionally, since some studies have hinted on the relevance of the voice signal, we use experimental source separation algorithms and extract the same features from the resulting voice-only audio signal. To conclude, we apply feature reduction techniques to the extracted feature set.

Section 4.4. Feature Selection and Emotion Classification

We then use the extracted features and gathered annotations in our classification experiments to assess the importance of the various feature sets. To this end, the features weight is computed using ReliefF and SVMs are used to classify songs based on quadrants (multi-class and binary), arousal (high and low) and valence (positive or negative).

Section 4.5. Classification Results and Discussion

After describing the feature selection classification strategies, we present the obtained results in each of the addressed problems.

Section 4.6. Feature Importance per MER Problem

Finally, based on the obtained feature rankings and classification results, we explore the uncovered relations between audio features and emotional categories.

4.1. Dataset Construction

Several MER datasets have been created over the years and used in a myriad of studies. However, most suffer from at least one of the following problems, as previously debated in Sections 3.2.1 and 3.2.2:

- Being private, used solely in their creators research;

- Missing data (e.g., audio samples), in part due to possible copyright restrictions;
- Very limited in size, with few tens or hundreds of song entries;
- Containing uncontrolled annotations (e.g., crawled from online social media);
- Limited to a specific genre, limiting research in a real-life scenario;
- Using unvalidated emotion taxonomies;
- Lacking details regarding its creation, especially in ground-truth collection;
- Containing low quality data (e.g., problems with samples and annotations)

As such, we propose a methodology for semi-automatic creation of a MER dataset that can be used to validate our work.

4.1.1. Dataset Requirements and Methodology

To avoid the pitfalls present in public available datasets, the following objectives were pursued in the process of creating our dataset:

1. To use a simple taxonomy, supported by psychological studies. Existent MER research is still unable to properly solve simpler problems with high accuracy. Thus, in our opinion, at this moment, there are few advantages to tackle problems with higher granularity;
2. To follow a semi-automatic construction process, reducing the resources needed to build a sizeable dataset;
3. To reach medium-high size, containing hundreds of songs;
4. To be prepared to a wide scope of research works, thus providing emotion quadrants as well as genre, artists or emotion tags for multi-label classification;
5. To be of varied and balanced nature, not being limited to a single genre or containing a very unbalanced number of instances per emotion or genre;
6. To have reduced ambiguity, containing songs with clear emotions during the manual validation, against excerpts with no agreement.

With this in mind, the methodology used to build our dataset consisted of the following steps, as described in Algorithm 4.1:

Algorithm 4.1. Algorithm used to build our dataset.

```
1. Gather songs and emotion data from AllMusic services. Ac-
```

cording to several authors, AllMusic data was curated by experts, relating music and emotions.

1.1. Retrieve the list of 289 emotion tags, E, using the AllMusic API.

1.2. For each emotion tag gathered, E_i, query the API for the top 10000 songs related with it, S_i.

2. Bridge the emotional data from AllMusic, based on an unvalidated emotional taxonomy, with Warriner's list, which associates adjectives to arousal, valence and dominance (AVD) values.

2.1. For each emotion tag, E_i, retrieve the associated AVD values from the Warriner's dictionary of English words. If the word is missing, remove it from the set of tags, E.

3. Data processing and filtering, in order to reduce the massive amount of gathered data to a more balanced but still sizeable set. This process contains several sub-steps:

3.1. Filter ambiguous songs, where a dominant emotional quadrant is not present. Compute the dominant emotional quadrant of a song:

3.1.1. Transform each emotion tag of a song to quadrant using the previously gathered AV values. Tags without AV values are considered Q0.

3.1.2. Set the dominant quadrant of a song as the most represented quadrant based on its emotion tags' list.

3.1.3. Compute the dominant quadrant weight of a song as the ratio of tags within this quadrant to the total number of tags.

3.1.4. Discard any song where the dominant quadrant is Q0 or the dominant quadrant weight is < 0.5.

3.2. Remove duplicated or very similar versions of the same songs by the same artists (e.g., different albums) by using approximate string matching against the combination of artist and title metadata.

3.3. Remove songs without genre information.

3.4. Remove songs associated with less than three emotion

tags.

4. Generate a subset dataset maximizing genre variability in each quadrant, as described in Algorithm 4.2 (page 176).
5. Manually validate the dataset in controlled conditions as described in Section 4.1.5. This follows a lighter process, where few resources are needed to blindly validate the audio clip and annotations.

The dataset creation process is described in detail in the next sections.

4.1.2. Data Collection from AllMusic

AllMusic[105] is a popular online music guide, cataloguing more than 3 million albums and 30 million tracks[106]. Originally launched as All Music Guide in 1991, its first release was as a 1,200-page reference book in CD-ROM format[107], predating the World Wide Web. Nowadays AllMusic is property of TiVo Corporation, previously known as Rov Corporation, which provides music metadata and recommendation platforms for several services and applications such as Spotify, iTunes, MTV or Billboard.

Regarding emotional information, AllMusic provides a list of 289 "mood tags"[108] that are associated with its albums and tracks. This data is said to be produced by experts (e.g., "annotations curated by experts" (Schmidt & Kim, 2010a), "experts at All Music Guide (AMG) have created a large mood taxonomy" (Pao et al., 2008), "provides professional reviews"(Hu & Downie, 2007), "mood labels that are applied to songs and albums by professional music editors" (Y.-H. Yang & Hu, 2012)). Additionally, AllMusic also supplies a list of 21 "genres"[109], each with a set of sub-genres and 2nd level sub-genres.

The AllMusic API[110] served as the source of musical information for our dataset, providing metadata as well as 30-second audio clips. To this end, we queried the API for the top songs for each of the 289 distinct emotion tags in it. The query response is very rich, containing metadata such as the results count (total number of songs that fit the

[105] https://www.allmusic.com/

[106] https://motherboard.vice.com/en_us/article/53djj8/the-story-of-allmusic-the-internets-largest-most-influential-music-database

[107] https://www.worldcat.org/title/all-music-guide-the-best-cds-albums-tapes-the-experts-guide-to-the-best-releases-from-thousands-of-artists-in-all-types-of-music/oclc/31186749/editions?referer=di&editionsView=true

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

[109] https://www.allmusic.com/genres

[110] http://developer.rovicorp.com/docs

query), list of artists, title, moods (representing the additional list of emotion tags that are associated with a specific song), genres, relevance of each entry to the query, styles (e.g., Singer/Song-writer, Country-rock), themes (e.g., "Romantic Evening"), audio sample, appearances (detailed information of the albums where the song appears), the text review and its reviewer and others. Although for this dataset only artist, title, mood and genre data was explored, the additional metadata is available to future studies.


```
searchResponse:
  totalResultCounts:      54567
  controlSet:             {…}
  results:
    0:
      relevance:          […]
      messages:           null
      type:               "song"
      id:                 "MT0003277857"
      song:
        styles:           […]
        sample:           "http://rovimusic.rovicor…qi0RS20pHqZddGAWQOA=&f=J"
        genres:           […]
        reviewUri:        "http://api.rovicorp.com/…8yx&trackId=MT0003277857"
        themesUri:        "http://api.rovicorp.com/…8yx&trackId=MT0003277857"
        title:            "If Not for You"
        appearances:      […]
        ids:              {…}
        review:
          text:           "[roviLink=\"MW\"]New Mor… release. ~ Thomas Ward"
          author:         "Thomas Ward"
        themes:           […]
        sampleUri:        "http://api.rovicorp.com/…8yx&trackId=MT0003277857"
        moods:
          0:
            id:           "XA0000000934"
            weight:       9
            name:         "Amiable/Good-Natured"
          1:              {…}
        moodsUri:         "http://api.rovicorp.com/…8yx&trackId=MT0003277857"
        appearancesUri:   "http://api.rovicorp.com/…8yx&trackId=MT0003277857"
        primaryArtists:
          0:
            id:           "MN0000066915"
            name:         "Bob Dylan"
        isPick:           true
```


Figure 4.1: Data sample extracted from AllMusic API response when queried by "bright" songs.

A condensed excerpt of the response obtained when querying for "Bright" mood is shown in Figure 4.1. Although the results count for each queried emotion tag is in some cases very high (for "Bright" there are 54567 entries, as the example shows), the access is limited to the first 10,000 entries. In addition, some of the emotion tags are rarely used. In particular, 77 of them are associated with 0 to 100 songs. The most used emotion tag is "Rousing", associated with a total of 135,437 songs.

According to the gathered data, a total of 5,558,883 songs in the AllMusic database are associated with the 289 emotion tags. Of these, we were able to crawl information on 1,731,141 songs (31.1%), an average of 5,990 songs for each emotion. These songs appear in a total of 18,850,162 albums, meaning that, on average, each song is included in 10.89 albums. Also, these entries were associated with 26,581,278 emotion tags (averaging 15.35 tags per song), 2,329,140 genre tags (averaging 1.35 genre tags per song), 7,897,847 theme tags (averaging 4.56 per song) and 5,761,312 styles tags (3.33 per song). Audio samples were available for a total of 1,596,082 song entries (92%). Since on average each song entry is associated with 15.35 emotion tags, it means that the same songs appear repeatedly in the results. After accounting for this, the number of unique songs drops to 370,611. This information is summarized in Table 4.1.

AllMusic DB	Song entries with emotion tags:	5,558,883
	Unique emotion tags:	289
Collected raw data	Crawled songs:	1,731,141 (31.1%)
	Emotion tags used:	26,581,278 (15.35 per song)
	Genre tags used:	2,329,140 (1.35 per song)
	Theme tags used:	7,897,847 (4.56 per song)
	Style tags used:	5,761,312 (3.33 per song)
	Audio samples (URLs):	1,596,082 (92%)
	Appearances (albums):	18,850,162 (10.89 per song)
Unique songs set	Song entries:	370,611 (21.4% of crawled)
	Distinct (unique) artists:	28,646
	Emotion tags used:	4,458,423 (12.02%)
	Songs' with sample URL:	330,700 (89.23%)
	Songs with genre tags	363,147 (97.99%)

Table 4.1: Summary of the initial set of data gathered using AllMusic API (in 2017).

Finally, the distribution of AllMusic emotion tags and genres per song for the raw collected data is shown in Figure 4.2. As illustrated, most songs contain a very high number of emotion tags, while one or two genre labels are frequently used.

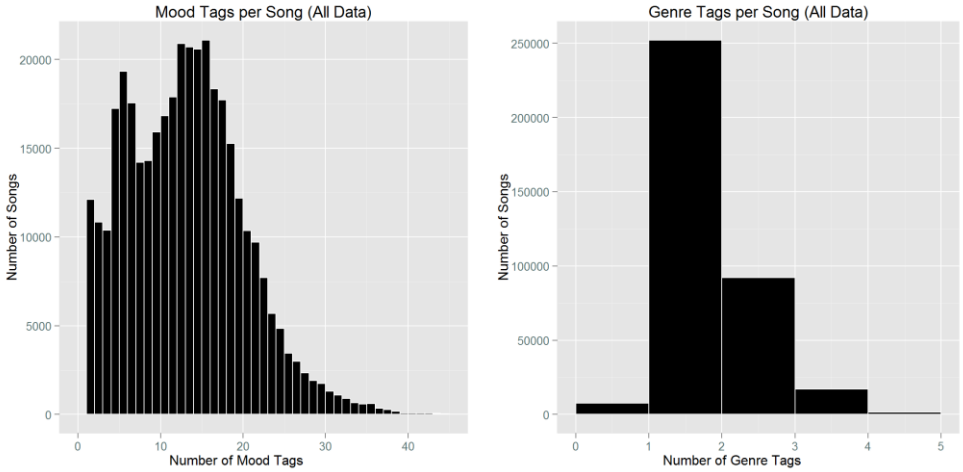


Figure 4.2: AllMusic mood tags and genre tags distribution in the raw collected data.

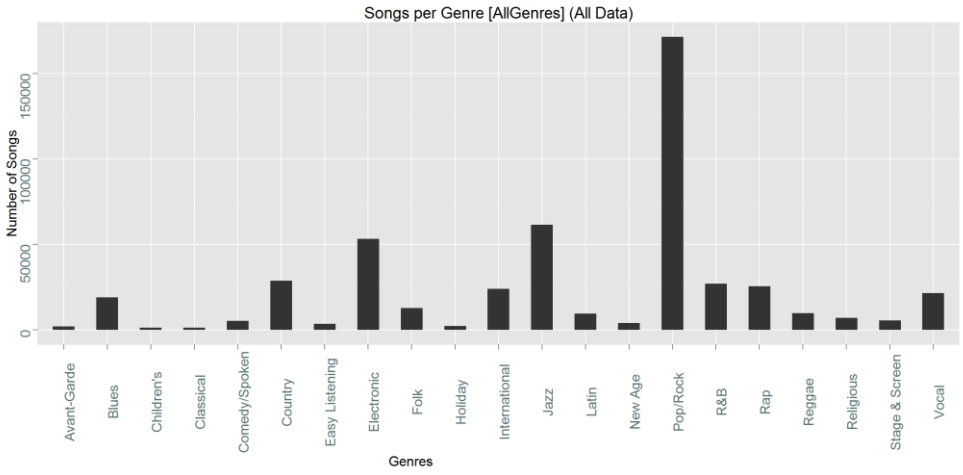


Figure 4.3: Number of songs per genre in the collected data.

The genre tags are highly unbalanced towards "Pop/Rock", as Figure 4.3 shows. Given the remaining AllMusic genres, "Pop/Rock" is actually an umbrella term covering distinct areas from heavy metal to pop, dance or punk rock and hence its massive usage.

More details are available in the AllMusic Pop/Rock sub-genres list[111]. Each of the 21 genres is further divided in sub-genres and second level sub-genres or styles.

Genre	*1st level sub-genres*	*2nd level sub-genres*
Avant-Garde	3	21
Blues	13	57
Children's	10	0
Classical	17	0
Comedy/Spoken	3	31
Country	7	35
Easy Listening	2	17
Electronic	7	60
Folk	2	21
Holiday	7	0
International	22	236
Jazz	11	67
Latin	5	63
New Age	24	0
Pop/Rock	15	215
R&B	3	34
Reggae	18	0
Religious	2	25
Stage & Screen	4	18
Vocal	17	0
Total	*192*	*900*

Table 4.2: AllMusic genres, sub-genres and styles.

As an example to better illustrate this imbalance, "Religious" genre has two sub-genres: "Contemporary gospel" and "Traditional gospel", which are then divided into 15 (e.g., "Christian rock" and "Latter-day Saints music" Latin gospel") and 10 (e.g., "Gospel choir" and "Hymns") styles respectively. In the same direction, the "Blues" genre

[111] https://www.allmusic.com/genre/pop-rock-ma0000002613

contains 13 sub-genres related with regions (e.g., "Chicago blues", "East coast blues", and "Louisiana blues") or instruments (e.g., "Harmonica blues", "Acoustic blues" and "Electric blues"). The same is true for other genres such as "Rap", "Latin", "Holiday", "R&B" or "Reggae", where the sub-genres are very specific divisions of the parent genre. However, the "Pop/Rock" genre is a much broader group, encompassing 15 sub-genres, among which "Alternative / Indie Rock", "Dance", "Folk / Country Rock", "Hard Rock", "Heavy Metal", "Punk / New Wave", and 215 styles (second-level sub-genres).

A list of the 21 AllMusic genres and number of sub-genres and styles is presented in Table 4.2. The full list of sub-genres and styles is omitted given its size but is available at the website[112]. As shown, "Pop/Rock" has a high number of sub-genres and styles, only matched by "International" genre. However, the latter is caused by the large number of very specific world styles (e.g., "Fado", "Flamenco" and "Carnatic"), for which less songs are available in the AllMusic database.

4.1.3. From AllMusic Emotion Tags to Russell's Quadrants

The reasons behind the selection of the 289 emotion tags in AllMusic are undocumented and, although assigned to songs by experts, lacks the scientific support that commonly used taxonomies such as Ekman's basic emotions or Russell's AV model offer.

For this reason, and to cope our first dataset requirement (Section 4.1.1), we derived a method to automatically transform the emotion tags of songs in Russell's quadrants based on AV values assigned to English words by previous Psychology studies. The two major studies in this subject were taken as source: the Affective Norms for English Words (ANEW) (Bradley & Lang, 1999) and the Warriner's adjectives list (Warriner, Kuperman, & Brysbaert, 2013). ANEW contains the mean and standard deviation of arousal, valence and dominance values for 1,034 English words, rated by Psychology students. Warriner's list is an improvement over ANEW and contains 13,915 English words with affective ratings in the same three dimensions. In addition to the more comprehensive list of words, it also contains a better documented annotation process.

AllMusic Emotion Tags vs. ANEW vs. Warriner's adjectives list

Two key aspects contributed to the selection of Warriner's adjectives list over ANEW. First, the fact that Warriner's list is an update to ANEW, covering its 1,034 words and adding 12,881 new words. The second reason was the coverage of AllMusic emotion tags by each of the two lists. To perform this analysis, the 12 AllMusic emotion tags composed by two words were split (e.g., "Anxious/Tense" to "Anxious" and "Tense"),

[112] https://www.allmusic.com/genres (at the bottom of each genre page).

leading to 301 tags. While ANEW contains only 50 of these 301 (16.6%), Warriner's list contains 200 (66.4%), maximizing the amount of collected data that can be used.

Nonetheless, the two lists were analyzed and compared to better support this decision. First, we compared the reported standard deviations in each word to understand if a significant difference between subjects exists. No relevant differences were found, namely, the mean value of the reported standard deviations was 1.679 (Warriner) vs 1.654 (ANEW) for valence and 2.300 vs. 2.367 for arousal (in a scale of 1 to 9). The fact that arousal annotations had higher standard deviation is interesting, showing that subjects have a harder time rating arousal for text, in contrast to what happens with music. This was confirmed by another study from our group (Malheiro et al., 2018).

Regarding the distribution of words across Russell's quadrants, Warriner's list is highly skewed towards low arousal quadrants (Q3 and Q4), while ANEW is much more balanced, in spite of a slight underrepresentation of Q3, as illustrated in Figure 4.4. There, Q0 represents words that have arousal and/or valence values equal to zero and thus cannot be placed unequivocally in one of the four quadrants. Still, it is important to stress the massive difference in size between both, where only 7.7% of words in Q1 for Warriner (1079 words, the less represented quadrant) is still more than the entire ANEW list (1034 words).

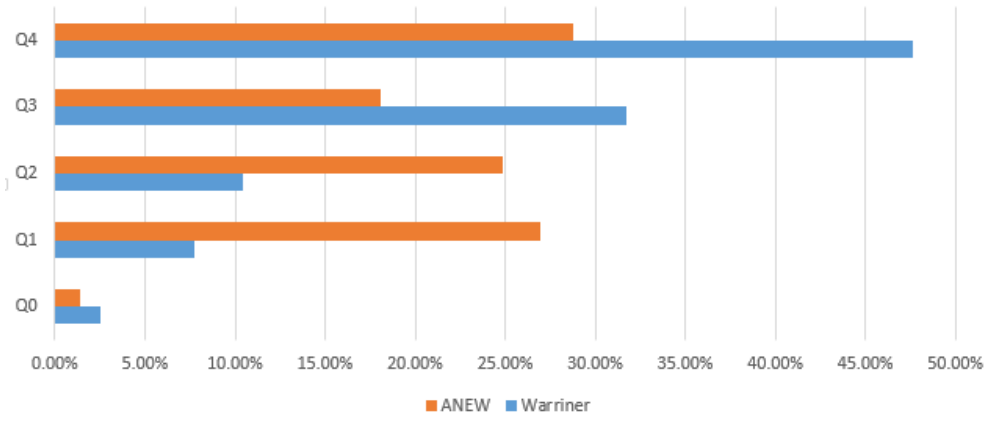


Figure 4.4: ANEW and Warriner's adjectives distribution (in %) across Russell's quadrants.

In addition, we compared the arousal and valence values for the 1034 common words across the two lists to understand if major differences exist. Regarding arousal values, the 1034 pairs of words are, on average, at a distance of 0.74 (± 0.53), with a correlation of 0.76. As expected, valence values are lower, with an average distance of 0.48 (± 0.39) and a correlation of 0.95, as illustrated in Figure 4.5. The histogram of

these distances is also provided in Figure 4.6.

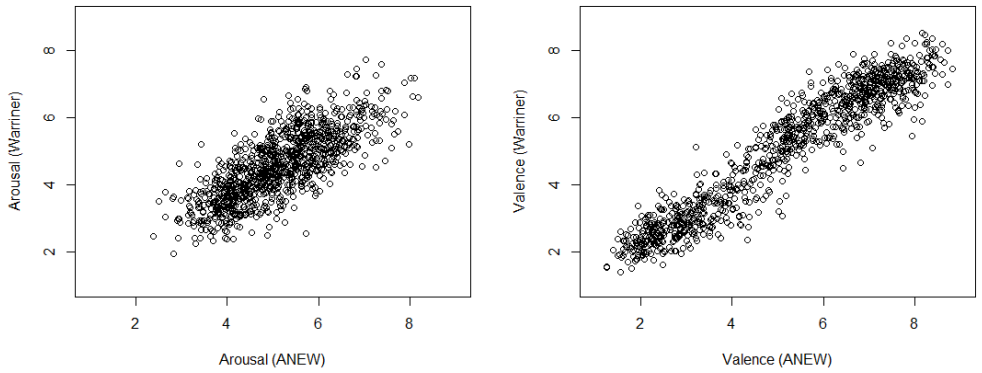


Figure 4.5: Arousal and valence correlation between pairs of common words in ANEW and Warriner's adjectives list.

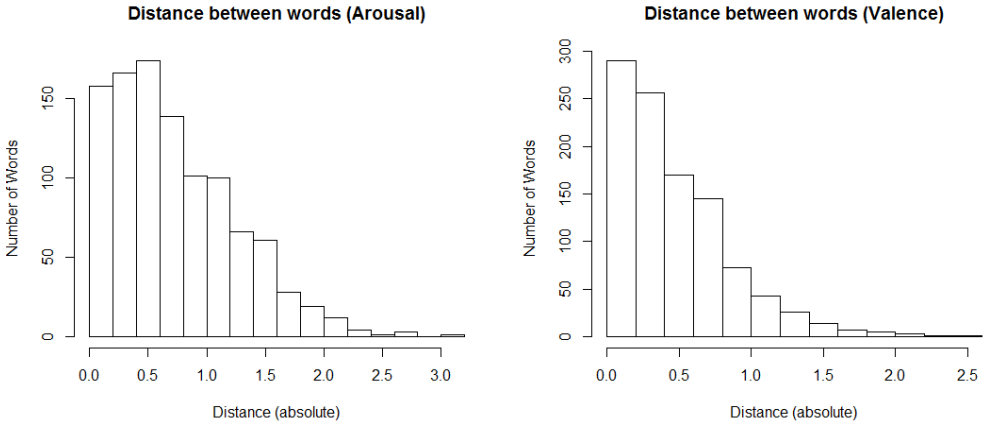


Figure 4.6: Distance between pairs of common words in ANEW and Warriner's adjectives list.

Finally, the differences between arousal, valence and dominance values for the entire lists (unpaired) and only for pairs of common words (paired) were statistically significant (significance level of 0.05). Despite this, and considering the high correlation found between the two lists and the two key aspects stated before, we believe Warriner's adjective list was the correct choice.

The final set of filtered tags and associated AV values is illustrated in Figure 4.7. As shown, a higher number of tags has positive valence (Q1: 49, Q2: 35, Q3: 33, Q4: 75). Of the 200, 8 are ignored in the next steps due to their null arousal or valence (Q0).

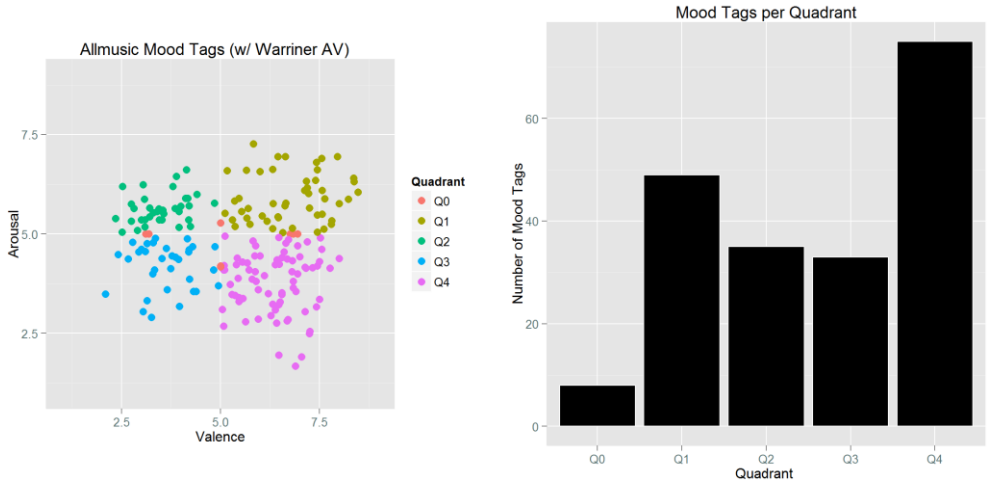


Figure 4.7: Filtered AllMusic emotion tags in the AV space and by quadrant.

4.1.4. Data Filtering and Mining

Having collected the set of raw data from AllMusic and a list of emotion tags associated with AV values, the next step is to transform this raw data into a balanced dataset and quadrant annotations of manageable size that can be manually validated.

The first step is to set the rules on how the quadrant of a song is computed and which emotion tags should be considered in this process. Regarding the later point, it is important to highlight two details of our data: 1) from the 289 emotion tags (expanded to 301), only 200 have AV values, thus the remaining tags cannot be assigned to a quadrant; 2) when querying the API for a specific emotion (e.g., "Happy"), the response contains a list of songs ordered by relevance (e.g., the songs considered the most "Happy"). However, a list of additional emotions for each song is also provided in the response but no information is given regarding its relevance and thus its usefulness must be reconsidered.

Several possible choices are possible to address these points. Regarding the unknown (no AV) emotion tags, we could either:

Option 1: Remove any song containing unknown tags;
Option 2: Keep the songs containing them but remove the unknown tags;
Option 3: Keep the songs and consider the tags, even if unknown, when computing the song's quadrant.

We chose option 3, i.e., to keep the songs where tags with no AV values are used and consider them in the quadrant calculation process as this was the option that would

better keep the original annotations given by AllMusic experts.

As for the second point, the two possible options are:

Option 1: Use all emotion tags available regardless of their relevance;
Option 2: Consider only the directly queried emotions, losing some less relevant emotional information (e.g., if a song is both "Happy" and "Bright" but did not appear in the list of the relevant songs for "Bright", only "Happy" will be considered).

In this case, we opted for the option 2, i.e., to consider only the directly queried emotions as they are known to be the most relevant for the song. Given the high number of tags used per song (15.35 on average) we believe that it is a good compromise to consider only the relevant emotions (e.g., the 5 relevant tags of a song) than to consider the full list as equally relevant (e.g., the full 15 tags, where the additional 10 are not considered as relevant). Although this choice eliminates some emotional information, it also generates a clearer vision of the prominent emotions in songs, useful to build a dataset with less ambiguity.

Even though these choices were used to create the final dataset, the following steps were also executed with all other possible options to guarantee that the best choice was made.

Considering the abovementioned decisions, to compute a single-label quadrant annotation of a song given its multi-label AllMusic emotion tags, we use Warriner's AV values to:

1. Convert each of the tags into quadrants;
2. Use the most represented quadrant as the song's quadrant. The weight of this single-label annotation is given by the ratio of the number of tags from the most represented quadrant against the total number of tags used, including the unknown ones.

Although only single-label quadrant annotations are used, this solution may be extended to accommodate further studies since we can generate:

- Single-label quadrant annotations;
- Single-label quadrant annotations and weight;
- Multi-label quadrant annotations;
- Multi-label quadrant annotations with weights (for fuzzy approaches);
- Multi-label emotion annotations (using tags) for higher granularity problems.

In addition, since emotion tags have AV values (mean and standard deviation), further efforts can be made to generate dimensional annotations from these, which can be either single points or fuzzy, reconciling all the available data – emotion tags, emotion AV values, quadrants weight and the additional relevance coefficient from the raw data.

Step 1: From AllMusic emotion tags to Russell's quadrants

The first data processing step was to transform song labels into Russell's quadrants, using Warriner's AV values. We found that tags associated with positive valence are more frequently used in the collected data (Figure 4.8). This can be partly explained by the unbalanced distribution of the 200 tags, as described earlier. As expected, there is a high number of tags that do not belong to any quadrant (NA, in the left plot), which account for all the emotion tags for which AV values are not available.

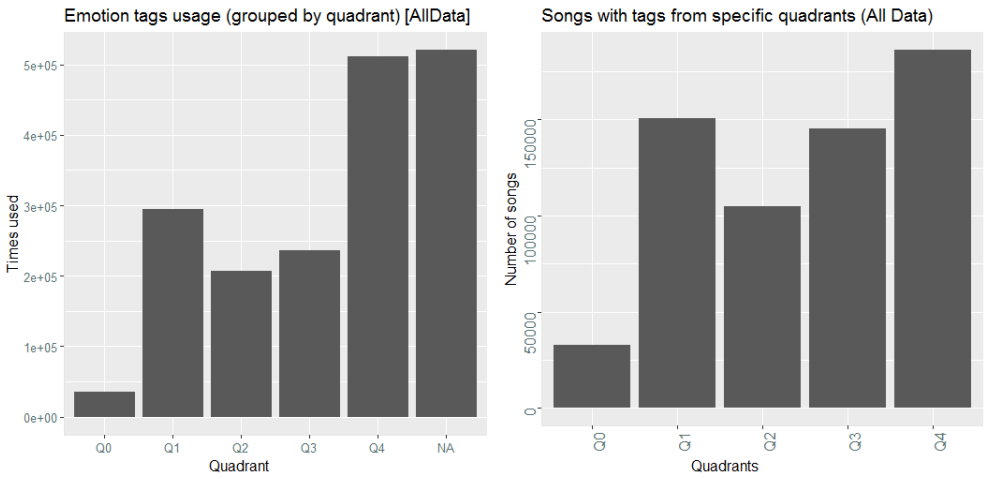


Figure 4.8: Emotion tags statistics in the raw data. Left: all the tags used according to the quadrants they belong to. Right: Number of songs containing at least one tag of a specific cluster.

After the transformation, the major quadrant for each song was assessed and its weight computed using the aforementioned method. This resulted in an average weight of 0.53 (± 0.32), with Q4 being the most represented (Figure 4.9). The major quadrant weight varied greatly, from 0.0 for songs containing only non-AV tags to 1.0, for songs where all the tags belong to the same cluster, many of them containing a single tag.

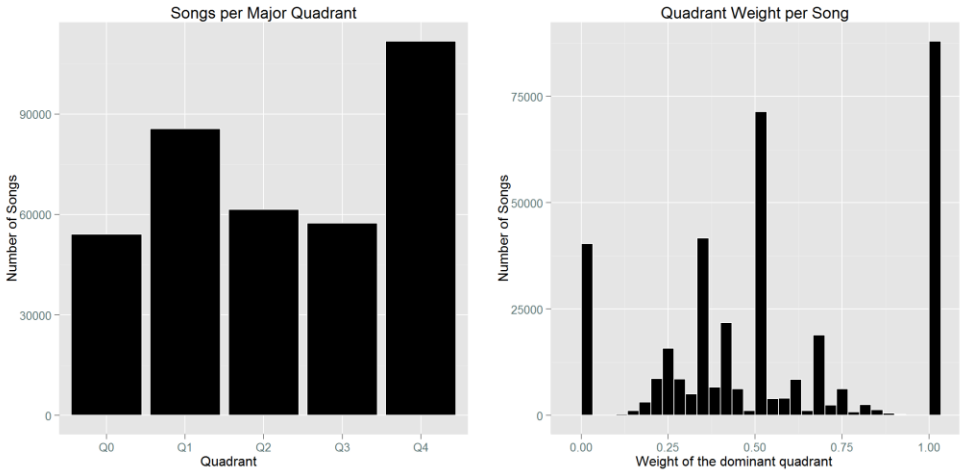


Figure 4.9: Songs per major quadrant (left) and major quadrant weights' distribution (right).

Step 2: Filter songs without a strong majority quadrant

To avoid songs where the emotional content is complex, unclear or ambiguous, we eliminate songs where the major quadrant weight is less than 0.5. In simple terms, we only consider songs with at least half of the emotion tags belonging to the same cluster.

This step reduced the number of songs to 120,733 (32.6%) of 17,676 distinct artists and further emphasized the unbalanced towards quadrant 4 (60,698 Q4 songs against 60,035 of the remaining three). This is somewhat explained by at least two reasons: given the higher number of Q4 emotion tags available in AllMusic it is natural to have a higher presence of Q4 tags in songs. For songs with a single tag, more of these will be Q4. The remaining, since more Q4 are available and used, have a higher chance of being Q4 and having a weight higher than 0.5.

One of the most interesting aspects of the dataset at this stage is the genre distribution across quadrants (Figure 4.10). Despite the expected skewness towards "Pop/Rock", we can see that different genres are more prevalent in specific quadrants (emotions). Namely, Q1 (happy songs) contains more Pop/Rock, Electronic, Jazz and R&B songs. Q2 (tense/anxious) is almost exclusively Pop/Rock (which contains sub-genres such as heavy-metal, punk or hard rock, as previously detailed in Table 4.2), Electronic and Rap. Q3 (sad) is mostly used in Blues, Jazz, R&B as well as Electronic and Pop/Rock. Finally, Q4 (calm/relaxed) is present in all genres but most of its songs are spread between Pop/Rock, Jazz, Country and Vocal. An additional interesting fact is that almost all songs of genre "Religious" belong to Q4.

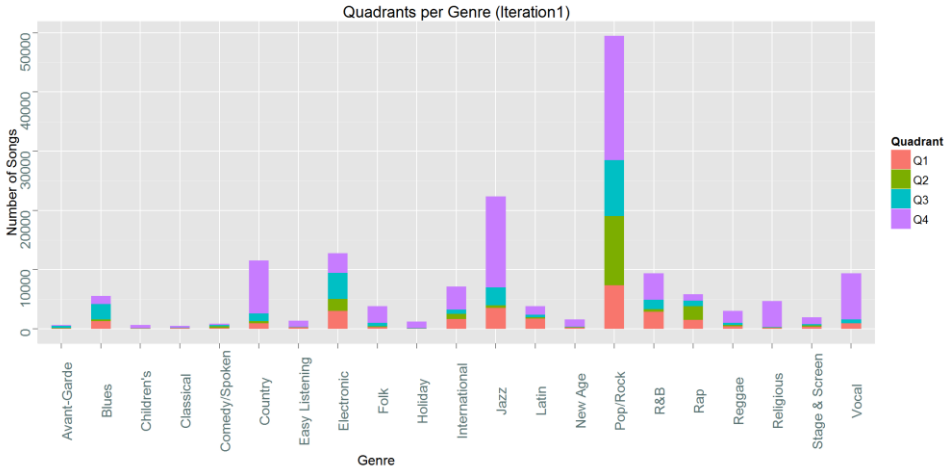


Figure 4.10: Songs distribution (by major quadrant) per each genre.

Step 3: Increase variability by removing similar songs

Very popular songs, especially from authors with a long career appear several times in the data. This is due to these song hits being released with slight variations over time, in different albums or featuring other artists. Even though such variations might contain slightly different emotions, caused mostly by variations in the artist's performance, we opted to keep only a single version of these.

The main goal is to increase the variability of the dataset and is accomplished by comparing similarities in metadata (mostly based on artist and title). For each duplicated song, we select the one to keep based on the one which contains more emotion and genre information. If the two duplicated songs are placed (major quadrant) in different quadrants we keep both, as this may indicate that the specific songs contain significant differences. In this process 11,477 duplicated songs were removed.

Step 4: Remove songs with no genre and few emotion tags (<3)

AllMusic experts use 289 emotion and 21 genre labels to review and annotate albums and songs, providing it to major commercial services. Hence, we consider that songs where no genre tag or only one or two emotion tags were selected (out of 289 possibilities) might indicate a not so accurate review process by the experts and thus choose to remove them.

At this point, only 3,917 songs did not contain genre information and were removed. Removing songs with less than three emotion tags greatly reduced the total number of candidate songs to 39,983, from which Q1 is the least represented (3,470) and

Q4 the most represented (24,982).

Step 5: Quadrant and genre balanced subset

At this point, the candidate dataset has already been reduced from 370,611 to 39,983 (10.8%) songs that better suit our requirements. However, this set is highly unbalanced in terms of quadrants and genres, and still too large for a manual validation by subjects to be attainable within our available resources. Thus, a final processing step was carried out to create a random subset of songs that is balanced in terms of quadrants and that maximizes genre diversity within each quadrant.

Since each song can have more than one genre tag and some genres are significantly more frequent, it is impossible to attain a perfect balance. Still, the following strategy was used to create the set:

Algorithm 4.2. Algorithm to maximize genre variability within a set.

1. For each quadrant, Q_i:
   - 1.1. Extract the list of distinct genres present in Q_i songs, G.
   - 1.2. Count the number of songs (belonging to Q_i) for each of these genres, S_g.
   - 1.3. Define the number of songs, N, to have in Q_i.
   - 1.4. Calculate the ideal number of songs, I, to have for each genre in Q_i, given by $I = N/length(G)$, where length(G) is the number of genres.
   - 1.5. Start with an empty set of songs for Q_i, S. For each genre in G:
     - 1.5.1. If the number of songs of that genre, S_g, is less than the ideal number, I, add all of them to the quadrant set, S.
     - 1.5.2. Else (if it has more), randomly select I songs (the ideal number) and add them to the quadrant set, S.
2. Eliminate duplicated songs from the quadrant set, S. (The generated set may have duplicated songs due to the multi-label nature of genre tags, e.g., picking a set of 10 songs from *genre1* and 10 songs from *genre2* might have common songs and thus result in less than 20 distinct songs).
3. Fill the remaining slots (due to song removal) with a random

sample from the songs still available in Q_i. (The generated set may be shorter than the desired size, caused by the duplicated removal and by genres with less songs than the ideal number).

The following figures illustrate the genre distribution before the algorithm is run (Figure 4.11) and the distribution of the created sample subset for Q2 (Figure 4.12).

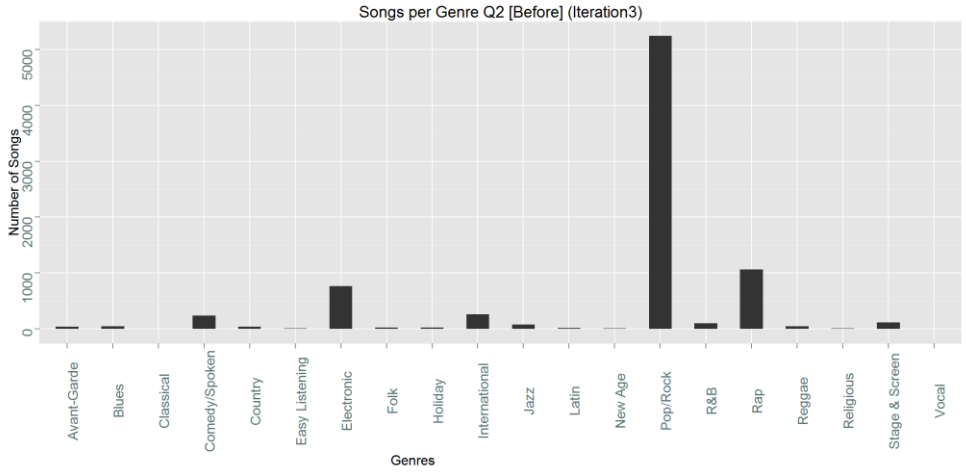


Figure 4.11: Number of songs of each genre in Q2.

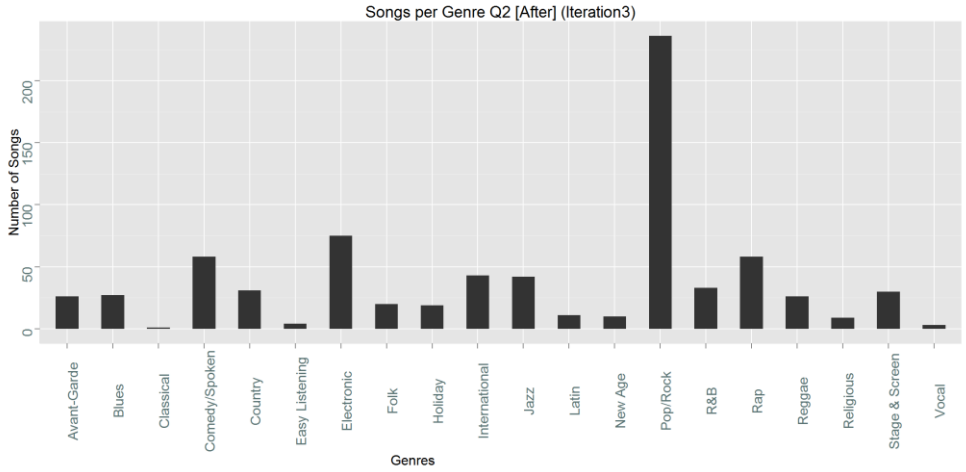


Figure 4.12: Number of songs of each genre in the generated sample subset of Q2.

Although it can be argued that the balancing process breaks the original genres proportions for the quadrant, it also increases the variability within the quadrant and, as both figures show, the most represented genres remain the same. Some genres are more prevalent in specific quadrants (e.g., "Pop/Rock in quadrant 2"), while others may be absent (e.g., there are no "New Age" songs in high arousal quadrants). This can be viewed in detail in Table 4.3, which presents the genre distribution in each quadrant. Since a song can have more than one genre tag the number of entries in the table exceeds the number of songs of our dataset.

Genre	*Q1*	*Q2*	*Q3*	*Q4*	*Total*
Avant-Garde	0	7	10	11	28
Blues	17	0	55	11	83
Children's	2	0	0	13	15
Classical	2	0	3	10	15
Comedy/Spoken	7	2	2	15	26
Country	21	4	33	27	85
Easy Listening	9	1	2	19	31
Electronic	37	54	18	32	141
Folk	5	4	27	14	50
Holiday	4	0	1	20	25
International	21	35	23	37	116
Jazz	56	8	52	51	167
Latin	35	4	13	12	64
New Age	0	0	7	37	44
Pop/Rock	86	151	72	59	368
R&B	38	8	26	22	94
Rap	8	41	10	9	68
Reggae	10	9	9	17	45
Religious	7	3	1	9	20
Stage & Screen	11	9	7	13	40
Vocal	23	0	35	30	88

Table 4.3: Number of songs by genre in each quadrant.

4.1.5. Dataset Validation

Not many details are known regarding the AllMusic emotion tagging process, apart from being made by experts (see Section 4.1.2). Still, several questions about this data have been raised by researchers over time. First, it is unclear whether the AllMusic experts evaluating music follow any guidelines regarding the analysis of the emotional content of songs. Namely, whether they are reporting evoked or perceived emotions and whether they consider only audio, only lyrics, a combination of both or if such decision is at the discretion of each. In addition, the segmentation process used by AllMusic to create the 30-second clips that represent each song is not documented. Given the commercial nature of the service, it would be expected that the most representative part of each song was selected. Yet, previous works demonstrated that some may be noisy (e.g., contain applauses, only speech, long silences or inadequate song segments such as the introduction), may contain clear changes in emotions or generate low agreement between annotators (e.g., (Vale, 2017); for details refer to Section 3.2.2).

To avoid the same pitfalls of previous datasets, which gathered subpar annotations in exchange for size, by using directly extracted data from internet sources and social media, a manual blind inspection of the candidate set was conducted. Two subjects from our lab were given sets of randomly distributed clips and asked to annotate them accordingly in terms of AV quadrants. The number of subjects was low given the fact that the validation process started with annotations generated by AllMusic experts. Still, a higher number of validators would be desirable if more resources were available. Beyond selecting a quadrant, the annotation framework allowed subjects to mark clips as unclear (if the emotion was unclear to the subject) or bad (if the clip contained noise, as defined above), as shown in Figure 4.13. In this way, additional ambiguous songs were discarded.

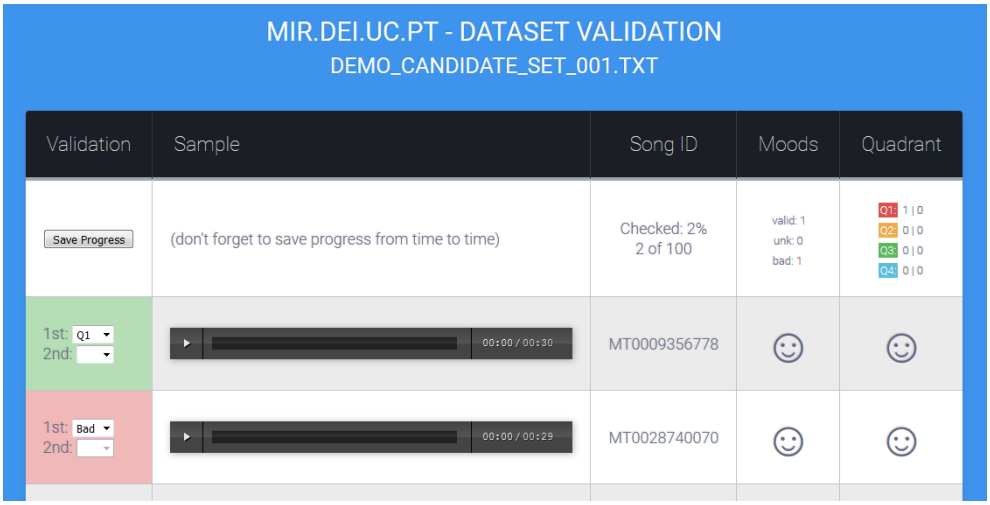


Figure 4.13: The annotation framework used to validate our dataset.

To construct the final dataset, song entries with clips considered bad or where subjects' and AllMusic's annotations did not match were excluded. With this procedure. we were able to reduce the number of subjects required (usually 10+) to annotate each song and still guarantee that both annotations and audio clips are verified, since both AllMusic experts' and our subjects' blind annotations matched.

Several interesting results were obtained from the validation process. First, our subjects demonstrated difficulties with low arousal songs, tagging many Q4 songs (according to AllMusic experts) as Q3, as shown in Figure 4.14. In addition, a high number of songs were marked as having bad samples, which highlights the importance of manually validating the audio clips gathered from online services.

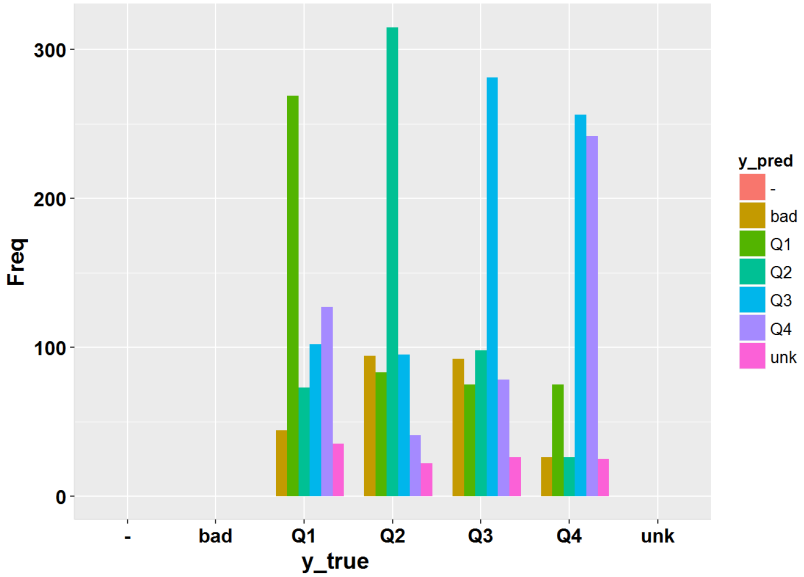


Figure 4.14: AllMusic experts' (y_true) versus our subjects' annotations (y_pred) after the first validation phase.

After the validation process, the amount of songs per quadrant was rebalanced to obtain a final set of 900 song entries, with exactly 225 for each quadrant. In our opinion, the dataset dimension is an acceptable compromise between having a larger dataset using MTurk workers as annotators or automatic but uncontrolled sources as annotations (e.g., Last.FM or AllMusic), and a very small and resource intensive dataset annotated exclusively by a high number of subjects in a controlled environment.

The final dataset also contains other types of annotations, possibly enabling its usage for different classification tasks (something that will be evaluated in the near future).

Each song entry is tagged in terms of quadrants, arousal and valence (positive or negative), and multi-label emotion tags. The distribution of these emotion tags is shown in Figure 4.15. In addition, most tags have an associated AV value, which might be used to place songs in the AV plane. Moreover, the remaining metadata (e.g., title, artist, album, year, genre and theme) can also be exploited in other MIR tasks. The dataset is nearly balanced in terms of emotion tags.

Figure 4.15: Dataset emotion tags. Words in red color are tags from Q1, Q2 in orange, Q3 in green and Q4 in blue. Black colored words represent the tags with no AV values.

Regarding genre, although the balance was increased, it was not possible to attain a fully balanced distribution. In fact, Pop/Rock is still predominant, mostly because it is overly generic and used together with other genres, as Figure 4.16 shows.

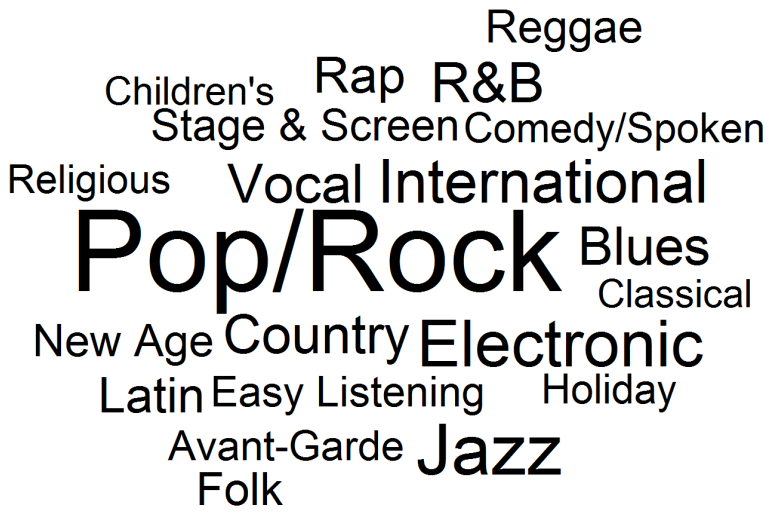


Figure 4.16: Dataset genre tags, with relatively balanced genres apart from Pop/Rock.

The exact number of songs labeled with each genre tag is available in column "total" of Table 4.3, presented previously (page 178). Given the high number of emotion tags available in our dataset (217), their distribution per quadrant and total is available in Table B.1 of Appendix B.

The final dataset is publicly available online[113] and can be used in other research studies. Moreover, it can be expanded in several ways, from the generation of new annotations as described in Section 4.1.2, to the increasing of its size by assigning resources to additional validation phases or by relaxing some of the processing steps that filtered out many of the candidate songs.

4.2. Novel Audio Features

Over the last decades, most MER studies have focused mainly on problems such as ground-truth collection and different machine learning approaches (e.g., regression, single- or multi-label classification). Still, regarding feature extraction, the research has been generally limited to experimenting different sets of standard audio features and feature reduction algorithms. Many of the standard audio features (see Chapter 3) employed in

[113] 4Q audio emotion dataset (2018), available at http://mir.dei.uc.pt/downloads.html

MER have been previously developed to solve different MIR problems. Moreover, several of these are low-level, extracted directly from the audio waveform or the spectrum.

In the previous chapters we identified several music elements that have been associated with music emotion. In addition, we reviewed the existent standard audio features, uncovering the musical dimensions that are less represented by these. Here, we build on the acquired knowledge and propose novel audio features that are more relevant to MER. As opposed to the information captured by low-level features, we naturally rely on clues like melodic lines, notes, intervals and scores to assess higher-level musical dimensions such as harmony, melody, articulation or texture. The explicit determination of musical notes, frequency and intensity contours are important mechanisms to capture such information and, therefore, we describe this preliminary step before presenting actual features, as follows.

4.2.1. From the Audio Signal to MIDI Notes

Going from audio waveform to music score is still an unsolved problem, and automatic music transcription algorithms are still imperfect (Benetos, Dixon, Giannoulis, Kirchhoff, & Klapuri, 2013). Still, we believe that estimating things such as predominant melody lines, even if imperfect, give us relevant information that is currently unused in MER.

To this end, we built on previous works by Salomon et al. (2012) and Dressler (2016) to estimate predominant fundamental frequencies (*f0*) and saliences. Typically, the process starts by identifying which frequencies are present in the signal at each point in time (sinusoid extraction). Here, 46.44 msec (1024 samples) frames with 5.8 msec (128 samples) hop-size (hereafter denoted *hop*) were selected.

Next, harmonic summation is used to estimate the pitches in these instants, as well as their respective salience. The result of this process is a representation of pitch salience over time (pitch salience function), as illustrated in the second panel (P2) of Figure 4.17, generated using the MELODIA[114] vamp plug-in under Sonic Visualizer (Salamon & Gómez, 2012). Given this, the series of consecutive pitches which are continuous in frequency are used to form pitch contours (panel 3 of the figure). These represent notes or phrases. Finally, a set of computations is used to select the *f0s* that are part of the predominant melody (Salamon & Gómez, 2012), shown in panel 4. The resulting pitch trajectories are then segmented into individual MIDI notes (P5) following the work by Paiva et al. (2006). The output of each intermediate step of this process is exemplified in Figure 4.17 using an excerpt of the song "S'posing" by Frank Sinatra.

[114] https://www.upf.edu/web/mtg/melodia

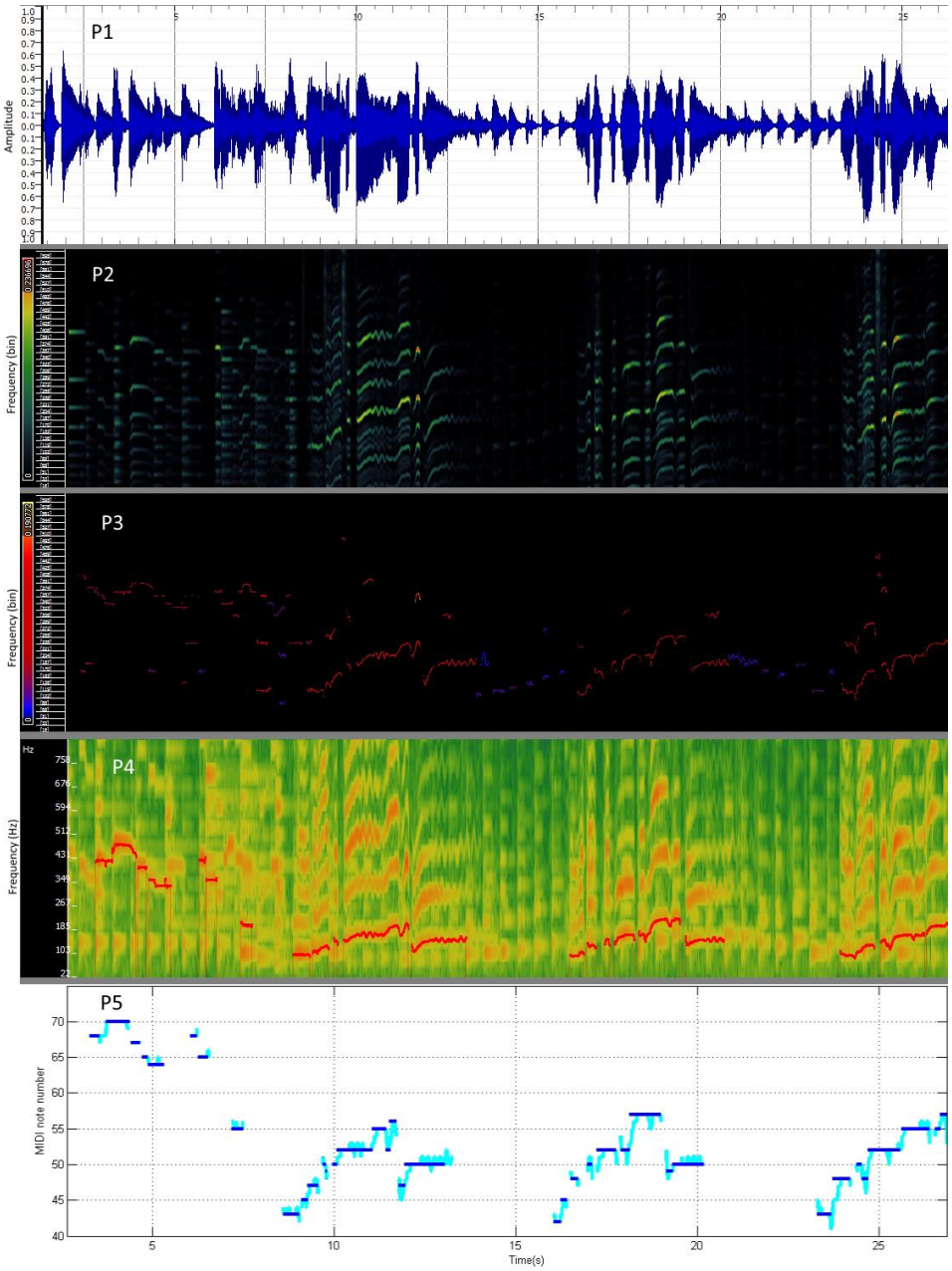


Figure 4.17: Excerpt from "S'posing" by Frank Sinatra, transformed from the audio signal to MIDI notes using the MELODIA plug-in (P1-P4) and Paiva et al. work (P5). P1: Audio waveform, P2: pitch salience function, P3: pitch contours, P4: extracted melody (in red) with the spectrogram as background, P5: MIDI notes.

Each of the N obtained notes, hereafter denoted as $note_i$, is characterized by: the respective sequence of $f0s$ (a total of L_i frames), $f0_{i,j}, j = 1, 2, ..., L_i$; the corresponding MIDI note numbers (for each $f0$), $midi_{j,i}$; the overall MIDI note value (for the entire note), $MIDI_i$; the sequence of pitch saliences, $sal_{j,i}$; the note duration, nd_i (sec); starting time, st_i (sec); and ending time, et_i (sec). This information is exploited to model higher level concepts such as vibrato, glissando, articulations and others, as follows.

In addition to the predominant melody, music may be composed of several musical lines produced by distinct sources. Although less reliable, there are works approaching multiple (also known as polyphonic) F0 contour estimation from these sources. We use Dressler's multi-F0 approach (2016) to obtain a frame-wise sequence of fundamental frequency estimates.

4.2.2. Melodic Features

Melody is a key dimension in music, defined as the horizontal succession of pitches. This set of features consists in metrics obtained from the notes of the melodic trajectory.

MIDI Note Number (MNN) statistics. Based on the MIDI note number of each note, $MIDI_i$ (see Section 4.4.1), we compute 6 statistics: $MIDImean$, i.e., the average MIDI note number of all notes, $MIDIstd$ (standard deviation of the MIDI note numbers), $MIDIskew$ (skewness), $MIDIkurt$ (kurtosis), $MIDImax$ (maximum) and $MIDImin$ (minimum).

Note Space Length (NSL)* and *Chroma NSL (CNSL). We also extract the total number of unique MIDI note values, NSL, used in the entire clip, based on $MIDI_i$. In addition, a similar metric, chroma NSL, $CNSL$, is computed, this time mapping all MIDI note numbers to a single octave (result 1 to 12).

Register Distribution. This class of features indicates how the notes of the predominant melody are distributed across different pitch ranges. Each instrument and voice type has different ranges, which in many cases overlap. In our implementation, 6 classes were selected, based on the vocal categories and ranges for non-classical singers (Peckham, Crossen, Gebhardt, & Shrewsbury, 2010). The resulting metrics are the percentage of MIDI note values in the melody, $MIDI_i$, that are in each of the following registers: Soprano (C4-C6), Mezzo-soprano (A3-A5), Contralto (F3-E5), Tenor (B2-A4), Baritone (G2-F4) and Bass (E2-E4). For instance, for soprano, it comes (4.1)[115]:

[115] Using the Iverson bracket notation, i.e., the bracket value is 1 if the condition inside holds true.

$$RDsoprano = \frac{\sum_{i=1}^{N} [72 \leq MIDI_i \leq 96]}{N} \tag{4.1}$$

Register Distribution per Second. In addition to the previous class of features, these are computed as the ratio of the sum of the duration of notes with a specific pitch range (e.g., soprano) to the total duration of all notes. The same 6 pitch range classes are used.

Ratios of Pitch Transitions. Music is usually composed of sequences of notes of different pitches. Each note is followed by either a higher, lower or equal pitch note. These changes are related with the concept of melody contour and movement. They are also important to understand if a melody is conjunct (smooth) or disjunct. To explore this, the extracted MIDI note values are used to build a sequence of transitions to higher, lower and equal notes.

The obtained sequence marking transitions to higher, equal or lower notes is summarized in several metrics, namely: Transitions to Higher Pitch Notes Ratio (*THPNR*), Transitions to Lower Pitch Notes Ratio (*TLPNR*) and Transitions to Equal Pitch Notes Ratio (*TEPNR*). There, the ratio of the number of specific transitions to the total number of transitions is computed. Illustrating for *THPNR*, follows (4.2):

$$THPNR = \frac{\sum_{i=1}^{N-1} [MIDI_i < MIDI_{i+1}]}{N-1} \tag{4.2}$$

Note Smoothness (NS) statistics. Also related to the characteristics of the melody contour, the note smoothness feature is an indicator of how close consecutive notes are, i.e., how smooth is the melody contour. To this end, the difference between consecutive notes (MIDI values) is computed. The usual 6 statistics are calculated, i.e., *NSmean* (mean value of *NS*, eq. (4.3)), *NSstd* (standard deviation), *NSskew* (skewness), *NSkurt* (kurtosis), *NSmax* (maximum), *NSmin* (minimum).

$$NSmean = \frac{\sum_{i=1}^{N-1} |MIDI_{i+1} - MIDI_i|}{N-1} \tag{4.3}$$

4.2.3. Dynamics Features

Exploring the pitch salience of each note and how it compares with neighboring notes in the score gives us information about their individual intensity, as well as intensity variation. To capture this, notes are classified as high (strong), medium and low (smooth) intensity based on the mean and standard deviation of all notes, as in (4.4):

$$SAL_i = \underset{1\le j\le L_i}{\text{median}}(sal_{j,i})$$

$$\mu_s = \underset{1\le i\le N}{\text{mean}}(SAL_i)$$

$$\sigma_s = \underset{1\le i\le N}{\text{std}}(SAL_i) \tag{4.4}$$

$$INT_i = \begin{cases} low, & SAL_i \le \mu_s - \frac{\sigma_s}{2} \\ medium, & \mu_s - \frac{\sigma_s}{2} < SAL_i < \mu_s + \frac{\sigma_s}{2} \\ high, & SAL_i \ge \mu_s + \frac{\sigma_s}{2} \end{cases}$$

There, SAL_i denotes the median intensity of $note_i$, for all its frames and INT_i stands for the qualitative intensity of the same note. Based on the calculations in (4.4), the following features are extracted.

Note Intensity (NI) statistics. Based on the median pitch salience of each note, we compute 6 statistics: *NImean*, i.e., the average pitch salience of all notes, *NIstd* (standard deviation of NI), *NIskew* (skewness), *NIkurt* (kurtosis), *NImax* (maximum) and *NImin* (minimum).

Note Intensity Distribution. This class of features indicates how the notes of the predominant melody are distributed across the three intensity ranges defined above. Here, we define three ratios: Low Intensity Notes Ratio (*LINR*), Medium Intensity Notes Ratio (*MINR*) and High Intensity Notes Ratio (*HINR*). These features indicate the ratio of number of notes with a specific intensity (e.g., low intensity notes, as defined above) to the total number of notes.

Note Intensity Distribution per Second. Low Intensity Notes Duration Ratio (LINDR), Medium Intensity Notes Duration Ratio (MINDR) and High Intensity Notes Duration

Ratio (HINDR) statistics. These features are computed as the ratio of the sum of the duration of notes with a specific intensity to the total duration of all notes. Furthermore, the usual 6 statistics are calculated, i.e., *LINDRmean* (mean value of *LINDR*), etc.

Ratios of Note Intensity Transitions. Transitions to Higher Intensity Notes Ratio (THINR), Transitions to Lower Intensity Notes Ratio (TLINR) and Transitions to Equal Intensity Notes Ratio (TEINR). In addition to the previous metrics, these features capture information about changes in note dynamics by measuring the intensity differences between consecutive notes (e.g., the ratio of transitions from low to high intensity notes) as illustrated in Figure 4.18.

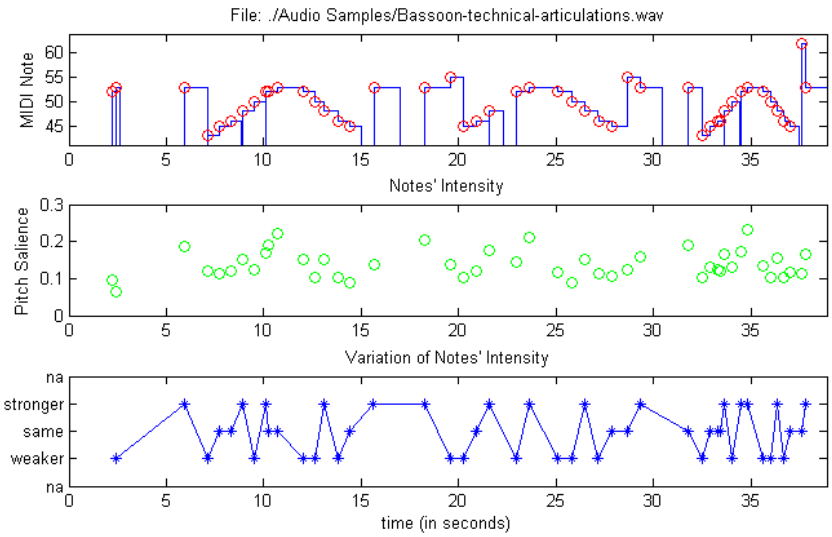


Figure 4.18: Assessing changes of intensity in consecutive notes.

Crescendo and Decrescendo (CD) statistics. Some instruments (e.g., flute) allow intensity variations in a single note. We identify notes as having crescendo or decrescendo (also known as diminuendo) based on the intensity difference between the first half and the second half of the note. A threshold of 20% variation between the median of the two parts was selected after experimental tests. From these, we compute the number of crescendo and decrescendo notes (per note and per sec). In addition, we compute sequences of notes with increasing or decreasing intensity, computing the number of sequences for both cases (per note and per sec) and length crescendo sequences in notes and in seconds, using the 6 previously mentioned statistics.

4.2.4. Rhythmic Features

Music is composed of sequences of notes and rests changing over time, each with a specific duration. Hence, statistics on note durations are obvious metrics to compute. Moreover, to capture the dynamics of these durations and their changes, three possible categories are considered: short, medium and long notes. As before, such ranges are defined according to the mean and standard deviation of the duration of all notes, as in (4.5). There, ND_i denotes the qualitative duration of $note_i$.

$$\mu_d = \underset{1 \le i \le N}{\text{mean}}(nd_i)$$

$$\sigma_d = \underset{1 \le i \le N}{\text{std}}(nd_i) \tag{4.5}$$

$$ND_i = \begin{cases} short, & nd_i \le \mu_d - \frac{\sigma_d}{2} \\ medium, & \mu_d - \frac{\sigma_d}{2} < nd_i < \mu_d + \frac{\sigma_d}{2} \\ long, & nd_i \ge \mu_d + \frac{\sigma_d}{2} \end{cases}$$

The following features are then defined.

Note Duration (ND) statistics. Based on the duration of each note, nd_i (see Section 4.2.1), we compute 6 statistics: *NDmean*, i.e., the mean of the duration of each note, *NDstd* (standard deviation of ND), *NDskew* (skewness), *NDkurt* (kurtosis), *NDmax* (maximum) and *NDmin* (minimum).

Note Duration Distribution. Short Notes Ratio (SNR), Medium Length Notes Ratio (MLNR), Long Notes Ratio (LNR). These features indicate the ratio of the number of notes in each category (e.g., short duration notes) to the total number of notes.

Note Duration Distribution per Second. Short Notes Duration Ratio (SNDR), Medium Length Notes Duration Ratio (MLNDR) and Long Notes Duration Ratio (LNDR) statistics. These features are calculated as the ratio of the sum of duration of the notes in each category to the sum of the duration of all notes. Next, the 6 statistics are calculated for notes in each of the existing categories, i.e., for short notes duration: *SNDRmean* (mean value of *SNDR*), etc.

Ratios of Note Duration Transitions. Ratios of Note Duration Transitions (RNDT). Transitions to Longer Notes Ratio (TLNR), Transitions to Shorter Notes Ratio (TSNR) and Transitions to Equal Length Notes Ratio (TELNR). Besides measuring the duration of notes, a second extractor captures how these durations change at each note transition. Here, we check if the current note increased or decreased in length when compared to the previous. For example, regarding the *TLNR* metric, a note is considered longer than the previous if there is a difference of more than 10% in length (with a minimum of 20 msec), as in (4.6). Similar calculations apply to the *TSNR* and *TELNR* features.

$$TLNR = \frac{\sum_{i=1}^{N-1} \left[\frac{nd_{i+1}}{nd_i} - 1 \right]}{N-1} \tag{4.6}$$

4.2.5. Musical Texture Features

To the best of our knowledge, musical texture is the musical dimension with less directly related audio features available (more precisely, zero feature, as discussed in Section 3.1). However, some studies have demonstrated that it can influence emotion in music either directly or by interacting with other features such as tempo and mode (Webster & Weir, 2005). We propose features related with the music layers of a song. Here, we use the sequence of multiple fundamental frequency estimates to measure the number of simultaneous layers in each frame of the entire audio signal, as described in Section 4.2.1.

Musical Layers (ML) statistics. As abovementioned, a number of multiple F0s are estimated from each frame of the song clip. Here, we define the number of layers in a frame as the number of obtained multiple F0s in that frame. Then, we compute the 6 usual statistics regarding the distribution of musical layers across frames, i.e., *MLmean*, *MLstd*, etc.

Musical Layers Distribution (MLD). Here, the number of *f0* estimates in a given frame is divided into four classes: i) no layers; ii) a single layer; iii) two simultaneous layers; iv) and three or more layers. The percentage of frames in each of these four classes is computed, measuring, as an example, the percentage of song frames identified as having a single layer (*MLD1*). Similarly, we compute *MLD0*, *MLD2* and *MLD3*.

Ratio of Musical Layers Transitions (RMLT). These features capture information about the changes from a specific musical layer sequence to another (e.g., ML1 to ML2). To this end, we use the number of different frequencies (*f0s*) in each frame, identifying

consecutive frames with distinct values as transitions and normalizing the total value by the length of the audio segment (in secs). Moreover, we also compute the length in seconds of the longest segment for each musical layer.

4.2.6. Expressivity Features

Few of the studied standard audio features are primarily related with expressive techniques in music. However, common characteristics such as vibrato, tremolo and articulation methods are commonly used in music. Their relation to emotions has been studied in some studies, e.g., (Dromey et al., 2015; Eerola, Friberg, & Bresin, 2013; Gomez & Danuser, 2007).

Articulation Features

Articulation is a technique affecting the transition or continuity between notes or sounds. To compute articulation features, we start by detecting legato (i.e., connected notes played "smoothly") and staccato (i.e., short and detached notes), as described in Algorithm 1. Using this, we classify all the transitions between notes in the song clip and, from them, extract several metrics such as ratio of staccato, legato and other transitions, longest sequence of each articulation type, and others.

Algorithm 4.3. Articulation detection.

1. For each pair of consecutive notes, $note_i$ and $note_{i+1}$:
   1.1. Compute the inter-onset interval (*IOI*, in sec), i.e., the interval between the onsets of the two notes, as follows: $IOI = st_{i+1} - st_i$.
   1.2. Compute the inter-note silence (*INS*, in sec), i.e., the duration of the silence segment between the two notes, as follows: $INS = st_{i+1} - et_i$.
   1.3. Calculate the ratio of INS to IOI (*INStoIOI*), which indicates how long the interval between notes is compared to the duration of $note_i$.
   1.4. Define the articulation between $note_i$ and $note_{i+1}$, art_i, as:
      1.4.1. *Legato*, if the distance between notes is less than 10 msec, i.e., $INS \leq 0.01 \Rightarrow art_i = 1$.

1.4.2. *Staccato*, if the duration of $note_i$ is short (i.e., less than $length_{max}$) and the silence between the two notes is relatively similar to this duration, i.e., $nd_i < length_{max} \wedge t_{low} \leq INStoIOI \leq t_{high} \Rightarrow art_i = 2$.

1.4.3. *Other Transitions*, if none of the abovementioned two conditions was met ($art_i = 0$).

Since no hard rules or values were found in scientific literature regarding articulation, the thresholds employed in Algorithm 4.3 were set experimentally (i.e., $length_{max}$ = *500 msec*, t_{low} = *0.25 and* t_{high} = *0.75*). The tests were run with a small set of different sound samples containing distinct times of articulation, as depicted in Figure 4.19, and thus need to be optimized with a more robust dataset.

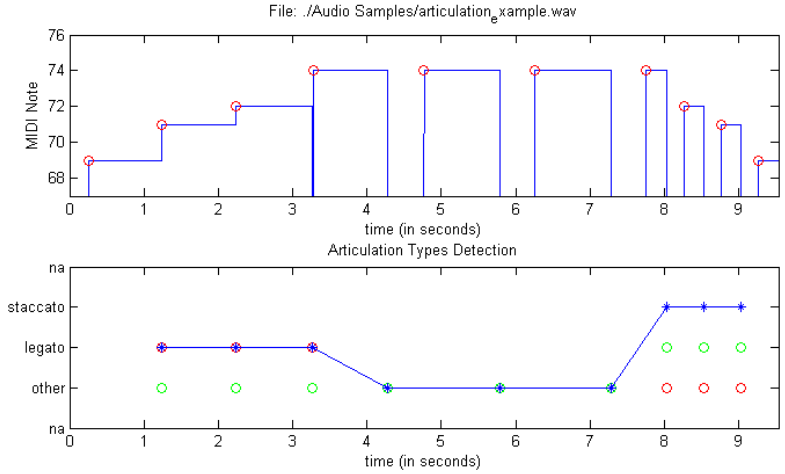


Figure 4.19: Testing articulation extraction with different note durations and intervals.

Based on the abovementioned algorithm, we define the following features:

Staccato Ratio (SR), Legato Ratio (LR)* and *Other Transitions Ratio (OTR). These features indicate the ratio of each articulation type (e.g., staccato) to the total number of transitions between notes.

Staccato Notes Duration Ratio (STNDR), Legato Notes Duration Ratio (LENDR) and Other Transition Notes Duration Ratio (OTNDR) statistics. Based on the notes duration for each articulation type, several statistics are extracted. The first is the ratio of the duration of notes with a specific articulation to the sum of the duration of all notes. Eq.

(4.7) illustrates this procedure for staccato (STNDR). Next, the usual 6 statistics are calculated, i.e., STNDRmean (mean value of STNDR, STNDRstd (standard deviation), STNDRskew (skewness), STNDRkurt (kurtosis), STNDRmax (maximum), STNDRmin (minimum).

$$STNDR = \frac{\sum_{i=1}^{N-1} \left[art_i = 1 \right] \cdot nd_i}{\sum_{i=1}^{N-1} nd_i} \tag{4.7}$$

Glissando Features

Glissando is another kind of expressive articulation, which consists in the glide from one note to another. It is used as an ornamentation, to add interest to a piece and thus may be related to specific emotions in music.

We extract several glissando features such as glissando presence, extent, length, direction or slope. In cases where two distinct consecutive notes are connected with a glissando, the segmentation method applied (mentioned in Section 4.2.1) keeps this transition part at the beginning of the second note (Paiva et al., 2006). The climb or descent, of at least 100 cents, might contain spikes and slight oscillations in frequency estimates, followed by a stable sequence. Given this, we apply the following algorithm:

Algorithm 4.4. Glissando detection.

1. For each note i:

   1.1. Get the list of unique MIDI note numbers, $u_{z,i}, z = 1, 2, \cdots, U_i$, from the corresponding sequence of MIDI note numbers (for each f0), $midi_{j,i}$, where z denotes a distinct MIDI note number (from a total of U_i unique MIDI note numbers).

   1.2. If there are at least two unique MIDI note numbers.

      1.2.1. Find the start of the steady-state region, i.e., the index, k, of the first note in the MIDI note numbers sequence, $midi_{j,i}$, with the same value as the overall MIDI note, $MIDI_i$, i.e., $k = \min_{1 \le j \le L_i,\, midi_{j,i} = MIDI_i} j$.

      1.2.2. Identify the end of the glissando segment as the first index, e, before the steady-state region, i.e., $e = k - 1$.

   1.3. Define

1.3.1. gd_i = glissando duration (sec) in note i, i.e., $gd_i = e \cdot hop$.

1.3.2. gp_i = glissando presence in note i, i.e., $gp_i =$ 1 if $gd_i > 0; 0,$ otherwise.

1.3.3. ge_i = glissando extent in note i, i.e., $ge_i = |f0_{1,i} - f0_{e,i}|$ in cents.

1.3.4. gc_i = glissando coverage of note i, i.e., $gc_i = gd_i/dur_i$.

1.3.5. $gdir_i$ = glissando direction of note i, i.e., $gdir_i = sign(f0_{e,i} - f0_{1,i})$.

1.3.6. gs_i = glissando slope of note i, i.e., $gs_i = gdir_i \cdot ge_i/gd_i$.

Then, we define the following features:

Glissando Presence (GP). A song clip contains glissando if any of its notes has glissando, as in (4.8).

$$GP\begin{cases}1, & if \exists i \in \{1,2,...,N\} : gp_i = 1 \\ 0, & otherwise\end{cases} \tag{4.8}$$

Glissando Extent (GE) statistics. Based on the glissando extent of each note, ge_i (see Algorithm 4.4), we compute the usual 6 statistics for notes containing glissando: GEmean, i.e., the mean of the glissando extent of each note with glissando, GEstd (standard deviation of GE), GEskew (skewness), GEkurt (kurtosis), GEmax (maximum) and GEmin (minimum).

Glissando Duration (GD) **and** ***Glissando Slope (GS) statistics.*** As with GE, we also compute the same 6 statistics for glissando duration, based on gd_i and slope, based on gs_i (see Algorithm 4.4).

Glissando Coverage (GC). For glissando coverage, we compute the global coverage, based on gc_i, using (4.9).

$$GC = \frac{\sum_{i=1}^{N} gc_i \cdot nd_i}{\sum_{i=1}^{N} nd_i} \tag{4.9}$$

Glissando Direction (GDIR). This feature indicates the global direction of the *glissandi* in a song, (4.10), as the ratio of upward glissando notes to total glissando notes. There, N_{gp} indicates the total number of notes with *glissando*:

$$GDIR = \frac{\sum_{i=1}^{N} gp_i}{N_{gp}} \text{ , when } gdir_i = 1 \tag{4.10}$$

Glissando to Non-Glissando Ratio (GNGR). This feature is defined as the ratio of the notes containing *glissando* to the total number of notes, as in (4.11):

$$GNGR = \frac{\sum_{i=1}^{N} gp_i}{N} \tag{4.11}$$

Vibrato and Tremolo Features

Vibrato is an expressive technique used in vocal and instrumental music that consists in a regular oscillation of pitch. Its main characteristics are the amount of pitch variation (extent) and the velocity (rate) of this pitch variation. It varies according to different music styles and emotional expression (Dromey et al., 2015).

Hence, we extract several vibrato features, such as vibrato presence, rate, coverage and extent. To this end, we apply a vibrato detection algorithm adapted from (Salamon, Rocha, & Gómez, 2012), as follows:

Algorithm 4.5. Vibrato detection.

1. For each note i:

   1.1. Compute the STFT, $|F0_{w,i}|, w = 1, 2, \cdots, W_i$, of the sequence $f0_i$, where w denotes an analysis window (from a total of W_i windows). Here, a 371.2 msec (128 samples) Blackman-Harris window was employed, with 185.6 msec (64 samples) hop-size.

1.2. Look for a prominent peak, $pp_{w,i}$, in each analysis window, in the expected range for vibrato. In this work, we employ the typical range for vibrato in the human voice, i.e., [5, 8] Hz (Salamon et al., 2012)[116]. If a peak is detected, the corresponding window contains vibrato.

1.3. Define:

1.3.1. vp_i = vibrato presence in note i, i.e., $vp_i = 1 \text{ if } \exists\, pp_{w,i}$; $vp_i = 0$, otherwise.

1.3.2. WV_i = number of windows containing vibrato in note i.

1.3.3. vc_i = vibrato coverage of note i, i.e., $vc_i = WV_i/W_i$ (ratio of windows with vibrato to the total number of windows).

1.3.4. vd_i = vibrato duration of note i (sec), i.e., $vd_i = vc_i \cdot d_i$.

1.3.5. $\text{freq}(pp_{w,i})$ = frequency of the prominent peak $pp_{w,i}$ (i.e., vibrato frequency, in Hz).

1.3.6. vr_i = vibrato rate of note i (in Hz), i.e., $vr_i = \sum_{w=1}^{WV_i} \text{freq}(pp_{w,i})/WV_i$ (average vibrato frequency).

1.3.7. $|pp_{w,i}|$ = magnitude of the prominent peak $pp_{w,i}$ (in cents).

1.3.8. ve_i = vibrato extent of note i, i.e., $ve_i = \sum_{w=1}^{WV_i} |pp_{w,i}|/WV_i$ (average amplitude of vibrato).

Then, we define the following features:

Vibrato Presence (VP). A song clip contains vibrato if any of its notes have vibrato, similarly to (4.8).

Vibrato Rate (VR) statistics. Based on the vibrato rate of each note, vr_i (see Algorithm 4.5), we compute 6 statistics: i.e., the weighted mean of the vibrato rate of each note, according to the respective note duration, nd_i, and vibrato coverage, vc_i, as in (4.12),

[116] The vibrato voice range is used since few works were found on vibrato ranges for sspecific instruments and those reported similar ranges (e.g., violin and erhu (L. Yang, Chew, & Rajab, 2013))

VRstd (standard deviation of VR), *VRskew* (skewness), *VRkurt* (kurtosis), *VRmax* (maximum) and *VRmin* (minimum).

$$VRmean = \frac{\sum_{i=1}^{N} vr_i \cdot vc_i \cdot nd_i}{\sum_{i=1}^{N} vc_i \cdot nd_i} \tag{4.12}$$

Vibrato Extent (VE)* and *Vibrato Duration (VD) statistics. As with VR, we also compute the same 6 statistics for vibrato extent, based on ve_i and vibrato duration, based on vd_i (see Algorithm 4.5).

Vibrato Coverage (VC). Here, we compute the global coverage, based on vc_i, in a similar way to (4.9).

High-Frequency Vibrato Coverage (HFVC). This feature measures vibrato coverage restricted to notes over note C4 (261.6 Hz). This is the lower limit of the soprano's vocal range (Peckham et al., 2010).

Vibrato to Non-Vibrato Ratio (VNVR). This feature is defined as the ratio of the notes containing vibrato to the total number of notes, similarly to (4.11).

Vibrato Notes Base Frequency (VNBF) statistics. As with the VR features, we compute the same 6 statistics for the "base frequency" (in cents) of all notes containing vibrato, that it the frequency of the notes themselves to understand if vibrato is more frequent in specific notes.

As for tremolo, this is a trembling effect, somewhat similar to vibrato, but regarding change of amplitude. Hence, a similar approach is used to calculate tremolo features. Here, the sequence of pitch saliences of each note is used instead of the *f0* sequence, since tremolo represents a variation in intensity or amplitude of the note. Given the lack of scientific supported data regarding tremolo, we used the same range employed in vibrato (i.e., 5-8Hz).

4.2.7. Other Features

Two existent approaches, previously used in other contexts were also tested: fractal dimension and a voice analysis toolkit.

Fractal dimension (FD) features. FD provides a metric of complexity by comparing (calculating the ratio) how a given pattern, normally a fractal, changes in detail with the change of scale. In this work we use Katz's FD implementation, which although slightly slower, is derived directly from the waveform (Esteller, Vachtsevanos, Echauz, & Litt, 2001). The FD of the signals was computed using a sliding window approach to promote stationarity, using windows of 50 ms with an overlap of 50%. The resulting FD signal was integrated using the 6 previously described statistical metrics.

Voice Analysis Toolbox (VAT) features. Another approach, previously used in other contexts was also tested: a voice analysis toolkit. We used the Voice Analysis Toolkit[117], a "set of MATLAB code for carrying out glottal source and voice quality analysis" to extract features directly from the audio signal. The selected features are related with voiced and unvoiced sections and the detection of creaky voice – "a phonation type involving a low frequency and often highly irregular vocal fold vibration, [which] has the potential [...] to indicate emotion" (Cullen, Kane, Drugman, & Harte, 2013).

Some researchers have studied emotion in speaking and singing voice (Scherer et al., 2015) and even studied the related acoustic features (Eyben et al., 2015). In fact, "using singing voices alone may be effective for separating the "calm" from the "sad" emotion, but this effectiveness is lost when the voices are mixed with accompanying music" and "source separation can effectively improve the performance" (X. Yang et al., 2017).

Hence, besides extracting features from the original audio signal, we also extracted all the previous audio features from the signal containing only the separated voice. To this end, we applied the singing voice separation approach proposed by Fan et al. (2016) (although separating the singing voice from accompaniment in an audio signal is still an open problem).

4.3. Feature Extraction and Reduction

To evaluate the relevance or our novel proposed features we used our dataset to extract standard and novel features. Given their high dimensionality, feature selection and reduction strategies were employed.

[117] https://github.com/jckane/Voice_Analysis_Toolkit

This section describes the methods and tools employed in this process.

4.3.1. Audio Feature Extraction

We use the newly created dataset do evaluate the performance of standard audio features, as well as to measure the influence of the novel features proposed in this work. As is customary in MER research, a preliminary step was used to convert the audio clips to a better suited format for the task – WAV PCM format, 22050 Hz sampling rate, 16 bits quantization and monaural. This is mainly done to reduce the load of the computer intensive extraction algorithms.

Based on the knowledge built on Section 3.1, we selected three audio frameworks – MIR Toolbox, Marsyas and PsySound3 to extract a total of 1603 audio features, distributed as shown in Figure 4.20. This high amount is in part caused by the summarization of features outputting time series data into six statistical measures – mean, standard deviation, skewness, kurtosis, maximum and minimum. Several factors contributed to the selection of the three abovementioned frameworks. In particular, because they are considered very relevant, being selected in many of the previous MER studies, as described in Section 3.2.5. Moreover, the three cover most of the standard audio features, as demonstrated in Section 3.1. Finally, our experience with these in our preliminary tests also contributed to this selection. In future experiments the Essentia audio framework should also be considered since it has been gaining importance in the field and contains a vast amount of standard features.

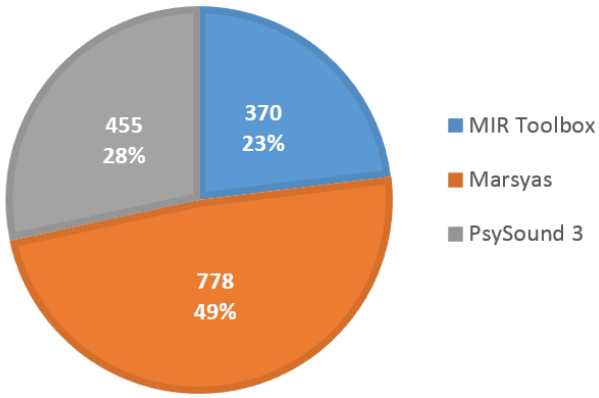


Figure 4.20: Standard audio features distribution per audio framework.

Next, the same audio clips were used to extract our novel proposed features, amounting to a total of 558 musical descriptors. As stated, some researchers have studied

emotion in speaking and singing voice and verified that using source separation to analyze the isolated voice signal may improve the performance. Therefore, in addition to the features extracted from the original audio signal, we also extracted a similar set of features from the signal containing only the separated voice by applying source separation algorithms, which although imperfect, may still bring relevant results to MER.

4.3.2. Reducing the Feature Dimensionality

A total of 2719 musical descriptors were obtained as the result of the feature extraction process. Such high feature dimensionality leads to very complex and resource intensive machine learning models, compromising its usage in reasonable time. Moreover, many features might be capturing similar information, especially within the three standard audio frameworks since they all offer some common features (e.g., many tone color features such as MFCCs or spectral moments are implemented in all). While such cases could have been solved from the start, by extracting only one version of the features, we decided not to follow this idea and extract repeated features (e.g., MFCCs from Marsyas and MIR Toolbox) as this allowed us to compare both implementations.

To reduce the high dimensionality and possible duplication of information across the feature set, we derived a two-step method based on standard deviation of the observed values, detection of outliers and correlation of pairs of features. The first step consists in eliminating features where the standard deviation of the observed data is zero, possibly caused by a bugged implementation or irrelevant extractor. The second step uses the correlation of pairs of features to eliminate equal or very similar features capturing the same information. This process, described in detail in Algorithm 4.6, was first carried on the standard feature set comprising 1603 features.

Algorithm 4.6. Feature dimensionality reduction.

1. Remove features where no variation was found in all the extracted values:
   1.1. For each feature, F_i, exclude the n lowest and highest values. Here, $n = 6$ was set experimentally.
   1.2. Compute the standard deviation of F_i, $std(F_i)$.
   1.3. If $std(F_i) == 0$, remove feature i.
2. Remove highly correlated pairs of features:
   2.1. Order features based on their importance using ReliefF feature selection algorithm.

2.2. For each feature, F_i, starting from the last feature (the one with the lowest weight), until $i = 2$:

2.2.1. Compute outliers of F_i, $OutF_i$, by using the box-and-whisker method and selecting as outliers the values lieing beyond the extremes of the whiskers. These extremes were defined as r times the interquartile range from the box, where r was set to 15 experimentally.

2.2.2. For each feature F_j, starting from $j = i - 1$ until $j = 1$:

2.2.2.1. Compute the outliers of F_j, $OutF_j$, similarly to 2.2.1.

2.2.2.2. Compute the correlation between F_i and F_j, $corr(F_i, F_j)$, excluding values which are outliers to F_i or F_j, e.g., $outF_i \cup outF_j$

2.2.2.3. If $corr(F_i, F_j) \geq maxCorrelation$, remove feature F_i. Here, $maxCorrelation$ was set experimentally to 0.9.

This strategy led to the reduction of the standard features set to 898 descriptors, removing several features ranging from features with no variations in the output, the same features extracted by different frameworks, different extractors capturing the same information, as illustrated in Figure 4.21. For each duplicated pair of features, the one with the highest weight according to the ReliefF algorithm is kept. This guarantees that the best features for our MER problem are not ignored.

Next, a similar strategy was used to remove similar features within the novel proposed features and also between the novel features and the standard features. This was performed by adapting Algorithm 4.6 to include the standard features before the ordered list of novel features. This strategy ensures that if a pair of features contains a novel and standard feature, the novel one is always removed. This decision guarantees that only novel features which are different than what already exists are used.

Figure 4.22 presents the number of standard and novel audio features extracted, organized by musical dimension. As previously discussed, most are timbral features, for the reasons pointed out. The novel features derived from the Voice Audio Toolbox and fractal dimension are considered Tone Color. The number of novel features used during emotion classification tests is twice what is present in the image, as the novel features are extracted two times – from the original audio and from the voice-only separated signal.

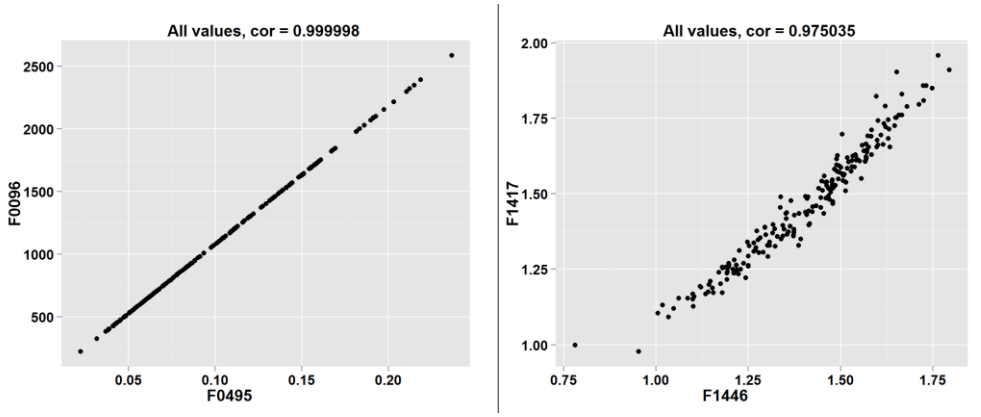


Figure 4.21: Correlation between pairs of features. Left: Zero Crossing Rate extracted with Marsyas (feature code F0495) and MIR Toolbox (feature code F0096). Right: Sharpness using two different loudness algorithms implemented in PsySound3 – Dynamic loudness (C & F) by Chalupper and Fastl (F1446), and Loudness (MG & B PsySound2) by Moore, Glasberg and Baer (F1447).

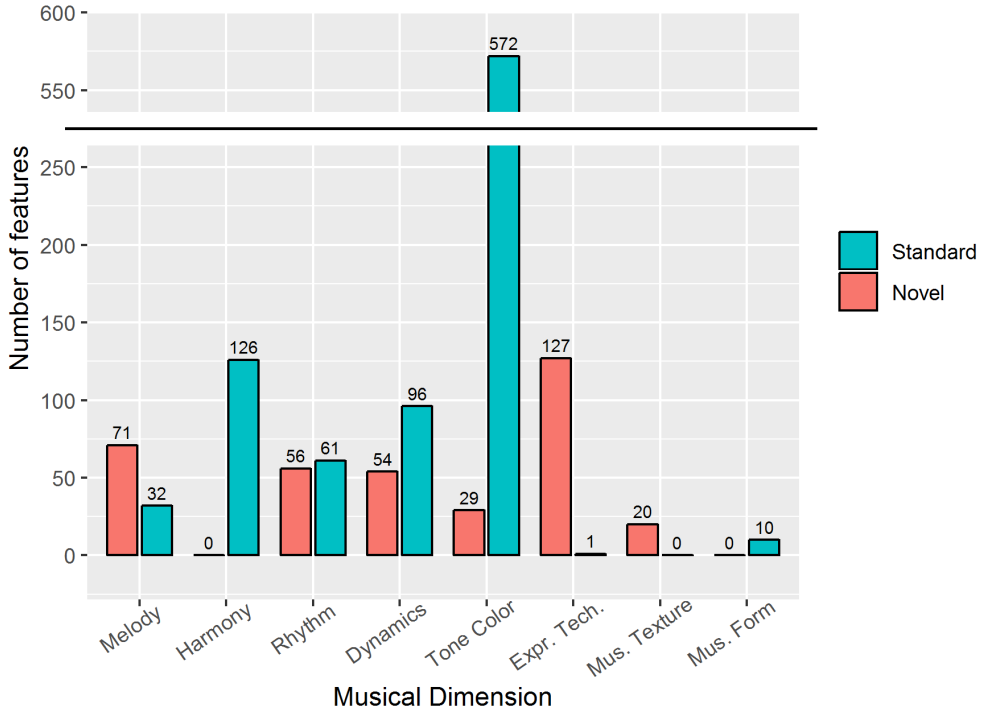


Figure 4.22: Feature distribution across musical dimensions.

4.4. Feature Selection and Emotion Classification

After the extraction of standard and novel features, followed by the removal of the unneeded (e.g., features that capture very similar information), the next steps consisted in feature selection and emotion classification. Feature selection algorithms were used to remove audio features providing information that is irrelevant to our MER problem, as well as to understand how relevant each feature and associated musical dimension are according to our dataset. Following, classification algorithms were used, exploring patterns in our data in order to classify music into different categories (e.g., quadrants, arousal or valence).

This section describes the strategies employed in these two steps.

4.4.1. Feature Selection

The number of audio features was greatly reduced by the feature reduction process described in the previous section. However, this number is still very high, making the classification experiments complex and resource-intensive. Moreover, the features removed pertained mostly to duplicated or invalid information, caused by redundant or bugged extractors. Although the current set of features contains no invalid or duplicated information, much of its content may be irrelevant to the problem we are addressing – MER.

The family of ReliefF feature selection algorithms (Robnik-Šikonja & Kononenko, 2003) was employed to select the better suited features for each classification problem. Several pertinent facts weighted in the choice of ReliefF. First, given the high number of features, an exhaustive feature selection algorithm was discarded. From the statistical selection methods available and previously discussed in Section 3.2.3, some transformed the feature space (e.g., PCA) and were also discarded, since we wanted to understand possible relations between (novel) features and emotional content. From the remaining, ReliefF has been previously used in MER studies and regarded as valuable (e.g., (Malheiro et al., 2018; Y.-H. Yang, Lin, Su, et al., 2008)). A description of Relief and its variants is available in Section 3.2.3.

The ReliefF algorithm uses the distance between K instances of the same and different classes to compute the weight of features. For robustness, two algorithm variants were used for averaging the weights: ReliefFequalK, where K nearest instances have equal weight, and ReliefFexpRank, where K nearest instances have weight exponentially decreasing with increasing rank.

Several feature rankings (ReliefF) were computed, measuring the influence of each feature in the entire set to different problems. Namely, the weight of each feature for:

- Quadrants classification (to differentiate between Q1, Q2, Q3 or Q4)

- Arousal classification (positive versus negative arousal)
- Valence classification (positive valence or negative valence)
- Identification of a specific quadrant as a binary problem (e.g., weight of each feature to classify a song as Q1 or not Q1)

The obtained feature ranking and weights were then used in several classification experiments, as detailed in the following sections.

4.4.2. Emotion Classification

As for classification, although new classification techniques have been gaining attention recently (e.g., deep learning (Delbouys et al., 2018; Pons et al., 2018)), in our experiments we used Support Vector Machines (SVM) (Chang & Lin, 2011). The focus of these experiments is not on classification algorithms but rather on the impact of novel features, and based on our experiments and in previous MER studies, this technique is robust and "is usually more efficient than the other classifiers" (X. Yang et al., 2017).

In addition to the default SVM model to predict quadrants, we evaluated two other approaches to better understand how the tested audio features relate to emotions in music.

First, a hierarchical approach was followed, dividing the problem into two levels. The first model is used to predict between hemispheres (high or low arousal) or meridians (positive or negative valence). In a second level, two models are trained to predict between each of the two classes, and the one selected depends on the result of the first level. As an example, the first level model trained to classify either high or low arousal, predicts a song to be low arousal. Next, a second model trained specifically to distinguish positive or negative valence in low arousal songs is used. The final result is then transformed into one of the four quadrants.

A second approach, binary quadrants, uses the training set to create four binary SVM models. Each model is trained to identify whether a song belongs to a specific quadrant or not (e.g., Q1 or not Q1), outputting a probability estimate for the test cases. The estimates of the four models are compared and the winning quadrant is the one with the highest probability estimate.

All experiments were validated with repeated stratified 10-fold cross validation (Duda, Hart, & Stork, 2000) (using 20 repetitions) since, according to the literature, "there are more performance estimates, and the training set size is closer to the full data size, thus increasing the possibility that any conclusion made about the learning algorithm(s) under test will generalize to the case where all the data is used to train the learning model" (Refaeilzadeh et al., 2009, p. 536). The average obtained performance

is reported. The result of feature selection is the complete list of features ordered by weight (relevance to a given problem).

4.5. Classification Results and Discussion

The classification tests were divided into four distinct problems to better understand the importance of features, all based in the Russell's AV taxonomy:

1. Classification by quadrant (multi-class);
2. Classification by arousal hemispheres (binary);
3. Classification by valence meridians (binary);
4. Classification by specific quadrant (one-vs-all).

The following sections detail the results of each approach.

4.5.1. Classification by Russell's Quadrants

A summary of the classification results of the different classification approaches is presented in Table 4.4. The model trained with standard features attained 67.5% F1-Score (macro weighted) using SVM and 70 standard features. The same solution achieved a maximum of 71.7% with a very high number of features (800). Adding the novel features (i.e., standard + novel features) increased the maximum result of the classifier to 76.4% (0.04 standard deviation), while using a considerably lower number of features (100 instead of 800). This improvement is statistically significant (at $p < 0.01$, paired T-test).

Classifier	*Feature set*	*# Features*	*F1-Score*
SVM	**standard**	**70**	**67.5% ± 0.05**
SVM	standard	100	67.4% ± 0.05
SVM	standard	800	71.7% ± 0.05
SVM	standard + novel	70	74.7% ± 0.05
SVM	**standard + novel**	**100**	**76.4% ± 0.04**
SVM	standard + novel	800	74.8% ± 0.04
SVM (HA av)	standard + novel	70	75.3% ± 0.04
SVM (HA va)	standard + novel	400	74.1% ± 0.04
SVM (BA)	standard + novel	150	75.6% ± 0.04

Table 4.4: Results of the classification by quadrants.

The remaining SVM classification strategies achieved close results. The hierarchical approach, using arousal in the first level, followed by valence (HA av) performed slightly better overall than the opposite (HA va), with the major difference being the number of features needed to obtain the maximum result. This is somewhat expected since, usually, predicting valence is a harder problem than predicting arousal (A. P. Oliveira & Cardoso, 2010; Y.-H. Yang, Lin, Su, et al., 2008). Thus, using valence in the second level, having two distinct models (one to classify Q1 vs Q2 and another to Q3 vs. Q4) might simplify the problem. As for the SVM binary approach (BA) using probability estimates, it proved to be slightly worse than the default SVM strategy, which uses "one-vs-one" for multi-class problems (according to the libSVM implementation). The results of all approaches are compared in Figure 4.23.

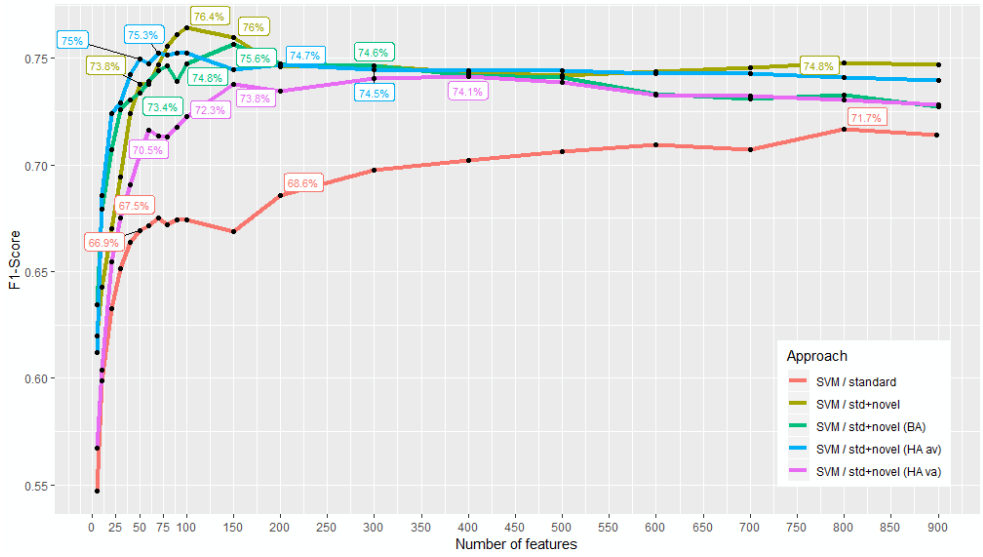


Figure 4.23: Results of the classification by quadrants.

Besides showing the overall classification results, we also analyze the results obtained in each individual quadrant (Table 4.5), which allows us to understand which emotions are more difficult to classify and what is the influence of the novel features in this process. In all our tests, a significantly higher number of songs from Q1 and Q2 were correctly classified when compared to Q3 and Q4. This seems to indicate that emotions with higher arousal are easier to differentiate with the selected features. Out of the two, Q2 obtained the highest F1-Score. This comes as no surprise to us, since several of the excerpts from Q2 belong to the heavy-metal and similar subgenres of pop/rock, which

are characterized by very distinctive, noise-like, acoustic features.

	standard			*standard + novel*		
Quads	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score
Q1	62.6%	73.4%	67.6%	74.6%	81.7%	78.0%
Q2	82.3%	79.6%	80.9%	88.6%	84.7%	86.6%
Q3	61.3%	57.5%	59.3%	71.9%	69.9%	70.9%
Q4	62.8%	57.9%	60.2%	69.6%	68.1%	68.8%

Table 4.5: Results per quadrant using 100 features.

The lower results in Q3 and Q4 (on average 12% below the results from Q1 and Q3) can be a consequence of several factors. First, a higher number of songs in these quadrants seem to be more ambiguous, containing unclear or contrasting emotions. During the manual validation process, we observed low agreement (45.3%) between the subject's ratings and the original AllMusic annotations. Moreover, subjects reported having greater difficulty distinguishing valence for songs with low arousal. In addition, some songs from these quadrants appear to share musical characteristics, which are related to contrasting emotional elements (e.g., a happy accompaniment or melody and a sad voice or lyrical message, despite the recommendations to ignore the lyrical part in the annotation validation process).

When using a similar number of features (100), the experiment with the addition of novel features shows an improvement of 9% in F1-Score when compared to the one using only standard features (baseline). This increment is noticeable in all four quadrants, ranging from 5.7% in quadrant 2, where the baseline classifier performance was already high, to a maximum increment of 11.6% in quadrant 3, which was the least performing using only standard features. Overall, the novel features improved the classification in general, with a greater influence in songs from Q3.

Regarding the misclassified songs, analyzing the confusion matrix (see Table 4.6, averaged for the 20 repetitions of 10-fold cross validation) shows that the classifier is slightly biased towards positive valence, predicting more frequently songs from quadrants 1 and 4 (total of 466.3, especially Q1 with 246.35) than from 2 and 3 (433.7). Moreover, a significant number of songs were wrongly classified between quadrants 3 and 4, which may be related with the ambiguity described previously. Based on this, further MER research needs to tackle valence in low arousal songs, either by using new features to capture musical dimensions currently ignored or by combining other sources of information such as lyrics. In fact, lyrics seem to have higher predictive value for valence, as concluded in (Malheiro et al., 2018).

		predicted				
		Q1	Q2	Q3	Q4	*Total*
actual	Q1	**183.85**	14.40	8.60	18.15	*225*
	Q2	23.95	**190.55**	7.00	3.50	*225*
	Q3	14.20	8.40	**157.25**	45.15	*225*
	Q4	24.35	1.65	45.85	**153.15**	*225*
	Total	*246.35*	*215.00*	*218.70*	*219.95*	*900*

Table 4.6: Confusion matrix using the best performing model.

4.5.2. Classification by Arousal and Valence

The experimental results of classification by arousal hemispheres and valence meridians (positive or negative) are presented in Table 4.7 and Table 4.8. In addition, similar results split by the opposite axis are also presented, e.g., arousal classification using only the positive valence songs is identified as Arousal (V+).

Classifier	*Feature set*	*# Features*	*F1-Score*
Arousal	standard	10	84.6%
Arousal	standard	300	88.9%
Arousal (V+)	standard	10	82.8%
Arousal (V+)	standard	300	87.2%
Arousal (V-)	standard	10	90.6%
Arousal (V-)	standard	200	93.3%
Arousal	standard + novel	10	86.5%
Arousal	standard + novel	100	90.7%
Arousal (V+)	standard + novel	10	84.0%
Arousal (V+)	standard + novel	400	89.2%
Arousal (V-)	standard + novel	10	92.0%
Arousal (V-)	standard + novel	400	93.8%

Table 4.7: Classification by arousal hemispheres (positive or negative).

Classifier	*Feature set*	*# Features*	*F1-Score*
Valence	standard	10	70.1%
Valence	standard	300	79.1%
Valence (A+)	standard	10	81.9%
Valence (A+)	standard	600	87.1%
Valence (A-)	standard	10	67.6%
Valence (A-)	standard	300	73.2%
Valence	standard + novel	10	72.7%
Valence	standard + novel	500	81.2%
Valence (A+)	standard + novel	10	84.9%
Valence (A+)	standard + novel	40	89.0%
Valence (A-)	standard + novel	10	72.2%
Valence (A-)	standard + novel	50	76.8%

Table 4.8: Classification by valence meridians (positive or negative).

Based on the experiments, it becomes evident that arousal classification (in positive or negative) is an easier problem than valence classification, with an F1-Score of 86.5% with only 10 features and a maximum of 90.7% with 100 features. Classification of valence meridians resulted in 72.7% with 10 features and 81.2% with 500 features. Summing up, the best 500 features used in valence meridians classification are still unable to obtain a result close to the best 10 features for arousal hemispheres. This indicates us that either: i) it is harder to discern valence (positive or negative) in music audio signals than arousal (high or low) and thus a higher number of audio features needs to be employed in the former problem; ii) or that the existent audio features (standard and ours) are still missing some important musical clues that we, as listeners, naturally associate with valence.

Moreover, our results also show that, in our dataset, differentiating between positive and negative arousal is harder for songs with positive valence: 84.0% and 89.2% versus 92.0% and 93.8% using 10 and 400 features respectively. This may be in part influenced by the dataset and difficulties felt by human subjects to place songs with close emotions such as glad, pleased, content and satisfied in either Q1 or Q4. On the other hand, songs from Q2 and Q3 have drastic acoustic differences, which make them generally easier to distinguish.

Regarding valence, and even though our novel features significantly improve its prediction, we observe a considerable difference in its classification under high versus low

arousal song sets. For songs with low arousal, valence classification obtained substantially worse results, especially with fewer features: 72.2% F1-Score, against 84.9% for high arousal songs using only 10 features. Moreover, the best results for both cases were obtained with 40 to 50 features, which may indicate that fewer of the tested features are as relevant to valence, when compared to the arousal problem. This indicates that to further improve MER algorithms, future work needs to focus on valence, specifically for low arousal songs. This coincides with our previous observation that distinguishing Q3 from Q4 is often ambiguous, for the abovementioned reasons.

4.5.3. Binary Classification

As a complement, creating one binary classification model for each specific quadrant against the remaining (e.g., Q1 vs non-Q1) helps us to understand how different are these quadrants and which features better separate them. Results from this experiment are shown in Table 4.9.

Classifier	*Feature set*	*# Features*	*F1-Score*
Quadrant 1	standard	20	82.8%
Quadrant 1	standard	600	85.9%
Quadrant 1	standard + novel	30	85.9%
Quadrant 1	standard + novel	100	89.1%
Quadrant 2	standard	10	88.3%
Quadrant 2	standard	898	92.1%
Quadrant 2	standard + novel	10	88.6%
Quadrant 2	standard + novel	800	92.6%
Quadrant 3	standard	20	77.9%
Quadrant 3	standard	600	81.7%
Quadrant 3	standard + novel	20	80.6%
Quadrant 3	standard + novel	150	84.6%
Quadrant 4	standard	5	78.1%
Quadrant 4	standard	500	82.1%
Quadrant 4	standard + novel	5	80.7%
Quadrant 4	standard + novel	200	83.5%

Table 4.9: Binary classification – each quadrant vs. the remaining (e.g., Q1 vs. non-Q1).

For each quadrant, we present two results: the maximum result obtained (e.g., 76.6% for Q1 using 100 features) and the first result higher than 95% of this best value (e.g., for Q1 this value is 72.0%, obtained with 50 features).

All the four models were able to discriminate songs with an F1-Score (macro) higher than 80%. Again, it is also noticeable that the high arousal quadrants (Q1 and Q2) are more distinctive. This is especially true for Q2, where 10 features are sufficient to separate Q2 from non-Q2 songs with an 88.6% F1-Score, while the maximum is obtained with 800 features. On the other hand, low arousal quadrants obtained a slightly lower result and in the case of Q4, going from 5 to 200 features resulted in less than 3% increase. This seems to indicate that songs in these two quadrants share more musical characteristics, containing less distinctive features.

4.6. Feature Importance per MER Problem

The importance of each audio feature to the MER problems studied in this work was measured using ReliefF. Some of the novel features proposed here appear consistently in the top 10 features for each problem and many others are in the first 100, demonstrating their relevance to MER. There are also features that, while alone may have a lower weight, are important to specific problems when combined with others. In this section we identify which audio features were the most relevant to the description and discrimination of specific classes.

4.6.1. Best Features for Quadrant Classification

The best result (76.4%, Table 4.4) was obtained with 29 novel and 71 standard features, which demonstrates the relevance of adding novel features to MER. In the paragraphs below, we conduct a more comprehensive feature analysis.

The best 10 features to the quadrants classification problem are presented in Table 4.10. As shown, three of these features are novel. The first two are related with musical texture, a dimension that was identified in Section 3.1 as lacking audio features primarily associated with it, while the third captures tremolo information. The standard features in the top 10 are related with tone color, capturing information about the spectrum (such as the energy distribution and symmetry of this distribution). Using only these first ten features results in 64.2% F1-Score. Given the complexity of the problem, a total of 100 features are needed to obtain 76.4%. Of these 100 features, 29 are novel, related to the following dimensions: expressive techniques (20), musical texture (8) and tone color using the Voice Analysis Toolbox (1). The 20 expressive techniques' features

are divided into: vibrato (11), glissando (4), tremolo (4) and articulation (1). The remaining 71 standard features are related with: tone color (51), dynamics (10), rhythm (4), harmony (4) and melody (2). This information is summarized in Figure 4.24, at the end of this section.

Feature	*Type*	*Dimension*	*Weight*
MFCC1 (mean)	standard	Tone Color	0.1781
FFT Spectrum - Average Power Spectrum (median)	standard	Tone Color	0.1729
Musical Layers (mean)	novel	Texture	0.1666
FFT Spectrum - Spectral 2nd Moment (median)	standard	Tone Color	0.1624
Spectral Skewness (std)	standard	Tone Color	0.1581
Spectral Skewness (max)	standard	Tone Color	0.1458
Rolloff (mean)	standard	Tone Color	0.1408
Musical Layers (std)	novel	Texture	0.1371
Rolloff (MeanA/StdM)	standard	Tone Color	0.1371
Tremolo Notes in Cents (mean)	novel	Tremolo	0.1366

Table 4.10: Top 10 features for quadrants classification.

From the novel features, the most relevant ones are related with musical texture and high-level expressive techniques (tremolo, glissando and vibrato) which are dimensions that were lacking in existent standard features. After all, if standard features were already extracting similar information from the audio signal, these novel features would have been discarded during our feature reduction phase (for details, see Section 4.4.1). The majority of the existent standard features are low-level tone color descriptors, which in part explains the higher number of these in the top 100. Nonetheless, dimensions such as harmony, melody and rhythm are underrepresented but have been identified as relevant to MER (as reviewed in Section 2.4). This may indicate that the existent features in these categories are not good enough and still need to be improved.

4.6.2. Best Features to Classify Arousal and Valence

Despite its coarse granularity, binary classification of arousal or valence (positive or negative) gives us additional information on the importance of features, which is hidden when using quadrants. Table 4.11 and Table 4.12 present the five best features for each

case.

Feature	*Type*	*Dimension*	*Weight*
FFT Spectrum - Average Power Spectrum (median)	standard	Tone Color	0.1969
MFCC1 (mean)	standard	Tone Color	0.1953
Spectral Skewness (std)	standard	Tone Color	0.1927
FFT Spectrum - Spectral 2nd Moment (median)	standard	Tone Color	0.1910
Musical Layers (mean)	novel	Texture	0.1790

Table 4.11: Top 5 features for arousal (high vs low).

For arousal, the top five features are able to achieve 84.9% F1-Score. In the top 10, two different standard features appear as relevant to classify arousal, when compared to the quadrants top. These features – Average Power Spectrum and Events Density are related with the speed and energy and thus expected to be related to arousal. The top result (90.7%) is obtained with 100 features: 24 novel and 76 standard. The novel are related with: expressive techniques (11), of which 6 are vibrato, 4 tremolo and 1 articulation; musical texture (8); dynamics (4); and tone color/VAT (1). The 76 standard audio features are divided in: tone color (55); dynamics (12); rhythm (4); harmony (2); melody (2); and musical form (1).

If the problem is further divided (e.g., arousal classification only for songs with positive valence), the best features differ. In terms of musical dimensions, the major differences between positive valence songs (PVS) and negative valence songs (NVS), when classifying arousal are (top 100): PVS rely less on novel features (20% against 27%); use less expressive techniques features (4 vs. 14), not relying on any vibrato features; and use more rhythm and melody related features (13 vs 6). NVS classification differs by using vibrato (9 features), but less features related with rhythm, melody and dynamics.

Feature	*Type*	*Dimension*	*Weight*
Vibrato Extent (std)	novel	Vibrato	0.1210
Vibrato Base Freq (kurtosis)	novel	Vibrato	0.1168
Vibrato Length (kurtosis)	novel	Vibrato	0.1127
Vibrato Rate (kurtosis)	novel	Vibrato	0.1109
Vibrato Length (skewness)	novel	Vibrato	0.1076

Table 4.12: Top 5 features for valence (high vs low).

For valence classification, the top five features selected were all novel features related with vibrato. This indicates that, at least for the tested dataset: i) vibrato features are highly correlated with valence. The definite reason for this is still unclear but previous studies have found that variations in vibrato rate and extent result in distinct emotions (Dromey et al., 2015; Konishi, Imaizumi, & Niimi, 2000); and ii) no standard features were able to capture the same information. Moreover, from the best 100 features, 43 are novel divided into: expressive techniques (33), of which 17 capture vibrato, 13 glissando, 2 articulation and 1 tremolo; musical texture (6); melody (2) and rhythm (2). The remaining standard features are divided between tone color (38); harmony (9); dynamics (4); rhythm (3); and musical form (3). From the results obtained in Section 4.5.2, we know that valence is better classified for songs containing high (positive) arousal. Analyzing the gathered feature rankings for valence in these cases, in order to understand their differences (with arousal either high/positive or low/negative), shows that the vibrato (19) and glissando (16) features are highly relevant for valence when arousal is positive. In this case, 48 new features are present in the top 100, of which 39 capture cues related with expressive techniques. Interestingly, this value drops to 42 in 100 for negative arousal songs, with only two vibrato features in the first 100. Of the 42, 20 are expressive techniques features, mostly glissando related (17). Moreover, and contrasting to the previous rankings, 16 novel features directly extracted from the separated voice-only signal are present in these 42. These are related with expressive techniques (glissando, 10), melody (2), tone color (VAT, related with voice sections, 2), musical texture (1) and fractal dimension (1). These results are resumed in Figure 4.24, at the end of the section.

Summarizing the obtained knowledge: several novel features are relevant for valence classification. Expressive techniques such as vibrato and glissando are related with valence, with the former having higher importance to more energetic songs. When the energy/stress of a song is lower, besides glissando, other features such as musical texture, melody and voice characteristics and signal/waveform complexity (fractal dimension) are also relevant. Furthermore, to these songs, for the first time, several features obtained from the separated audio signal containing only the voice were selected. Such result supports the hypothesis that, in this situation (predicting valence in low arousal songs), the voice carries more emotional information than in the previous situations and analyzing it separately is beneficial. It also raises several questions:

1. Given the results, is the non-voice (accompaniment) signal in this hemisphere (low arousal) less important and in some cases even hardening MER classification? This idea seems to be supported by Xu et al. (2014).
2. Could it be that we, as listeners, when faced with this type of songs, tend to unconsciously ignore the accompaniment in them and focus on voice/singer when looking for emotional cues?

3. Moreover, assuming that the vocal acoustic information is more relevant to discriminate songs in this hemisphere than the others, what about the message in that voice signal? Can the lyrics improve it further to a level similar to the arousal classification?

4.6.3. Best Features to Discriminate each Quadrant

In this section we discuss the best features to discriminate each specific quadrant from the others, according to specific feature rankings (e.g., ranking of features to separate Q1 songs from non-Q1 songs). The top 5 features to discriminate each quadrant are presented in Table 4.13 (page 216).

Except for quadrant 1, the top5 features for each quadrant contain a majority of tone color features, which are overrepresented in the complete feature set in comparison to the remaining. It is also relevant to highlight the higher weight given by ReliefF to the top5 features of both Q2 and Q4. This weight is confirmed by results in Table 4.9, where few features are needed to obtain 95% of the maximum score for both quadrants, when compared to Q1 and Q3.

Musical texture information, namely the number of musical layers and the transitions between different texture types (two extracted from voice-only signals) were also very relevant for quadrant 1, together with several rhythmic features. However, the ReliefF weight of these features to Q1 is lower when compared with the top features of other quadrants. Happy songs are usually energetic, associated with a "catchy" rhythm and high energy. The higher number of rhythmic features used, together with texture and tone color (mostly energy metrics) support this idea. Interestingly, creaky voice detection extracted directly from voice is also highlighted (it ranked 15th), which has previously been associated with emotion (Cullen et al., 2013).

The best features to discriminate Q2 are related with tone color, such as: roughness – capturing the dissonance in the song; rolloff and MFCC – measuring the amount of high frequency and total energy in the signal; and spectral flatness measure – indicating how noise-like the sound is. Other important features are tonal dissonance (dynamics) and expressive techniques such as vibrato. Empirically, it makes sense that characteristics like sensory dissonance, high energy, and complexity are correlated to tense, aggressive music. Moreover, research supports the association of vibrato and negative energetic emotions such as anger (Scherer et al., 2015).

In addition to the tone color features related with the spectrum, the best 20 features for quadrant 3 also include the number of musical layers (texture), spectral dissonance, inharmonicity (harmony), metrics of signal complexity (fractal dimension) and expressive techniques such as tremolo. Moreover, nine features used to obtain the maximum score are extracted directly from the voice-only signal. Of these, four are related with

intensity and loudness variations (crescendos, decrescendos); two with melody (vocal ranges used); and three with expressive techniques such as vibratos and tremolo. Empirically, the characteristics of the singing voice seem to be a key aspect influencing emotion in songs from quadrants 3 and 4, where negative emotions (e.g., sad, depressed, and miserable) usually have not so smooth voices, with variations in loudness (dynamics), tremolos, vibratos and other techniques that confer a degree of sadness (Scherer et al., 2015) and unpleasantness.

Q	*Feature*	*Type*	*Dimension*	*Weight*
Q1	FFT Spectrum - Spectral 2nd Moment (median)	standard	Tone Color	0.1467
	Transitions ML1 -> ML0 (Per Sec)	novel	Texture	0.1423
	MFCC1 (mean)	standard	Tone Color	0.1368
	Transitions ML0 -> ML1 (Per Sec)	novel (voice)	Texture	0.1344
	Fluctuation (std)	standard	Rhythm	0.1320
Q2	FFT Spectrum - Spectral 2nd Moment (median)	standard	Tone Color	0.2528
	Roughness (std)	standard	Tone Color	0.2219
	Rolloff (mean)	standard	Tone Color	0.2119
	MFCC1 (mean)	standard	Tone Color	0.2115
	FFT Spectrum - Average Power Spectrum (median)	standard	Tone Color	0.2059
Q3	Spectral Skewness (std)	standard	Tone Color	0.1775
	FFT Spectrum - Skewness (median)	standard	Tone Color	0.1573
	Tremolo Notes in Cents (mean)	novel	Tremolo	0.1526
	Linear Spectral Pairs 5 (std)	standard	Tone Color	0.1517
	MFCC1 (std)	standard	Tone Color	0.1513
Q4	FFT Spectrum - Skewness (median)	standard	Tone Color	0.1918
	Spectral Skewness (std)	standard	Tone Color	0.1893
	Musical Layers (mean)	novel	Texture	0.1697
	Spectral Entropy (std)	standard	Tone Color	0.1645
	Spectral Skewness (max)	standard	Tone Color	0.1637

Table 4.13: Top 5 features for each quadrant discrimination.

The majority of the employed features were related with tone color, where features

capturing vibrato, texture and dynamics and harmony were also relevant, namely spectral metrics, the number of musical layers (and its variations) and measures of the spectral flatness (noise-like). More features are needed to better discriminate Q3 from Q4, which musically share some common characteristics such as lower tempo, less musical layers and energy, use of glissandos and other expressive techniques.

A visual representation of the best features in each classification problem, grouped by categories and separated in standard or novel (extracted from original audio or voice-only audio), is represented in Figure 4.24 and Figure 4.25. As previously discussed, all quadrants use tone color features. On the other hand, some categories of features are more relevant to specific quadrants, such as rhythm and glissando for Q1, or fractal dimension and voice characteristics (VAT) to Q2.

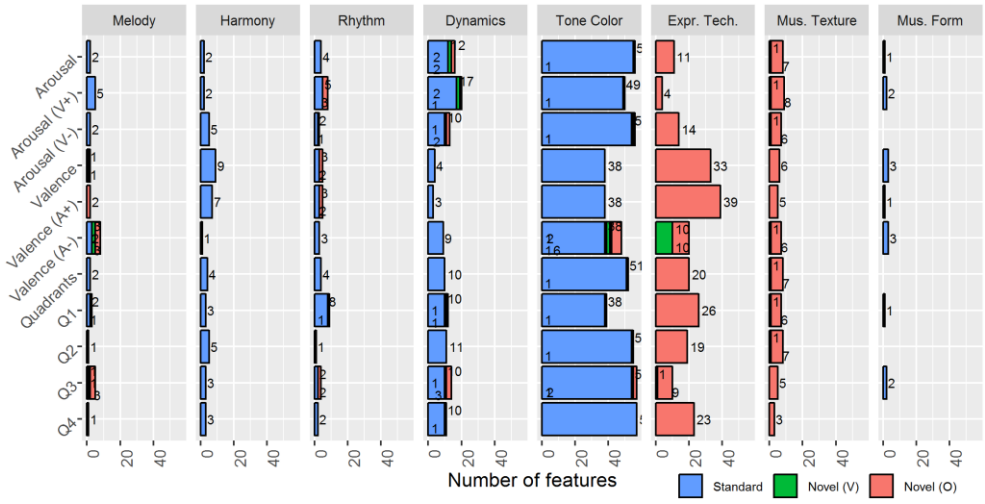


Figure 4.24: Best 100 features in each classification problem studied, organized by musical dimension. Novel (O) are extracted from the original audio signal, while Novel (V) are extracted from the voice-separated signal.

To conclude, the weight of the entire feature set to each problem is shown in Figure 4.26. Although complex, this illustration allows for the comparison of the weights for each problem. As can be seen, for some problems, especially valence prediction with low arousal songs – Valence (A-), the relevance of the available features is much lower than the others. This can be caused by lack of relevant features, higher ambiguity in the dataset and annotations or both. Additional visualizations of feature weights and ranks are available in Appendix C.

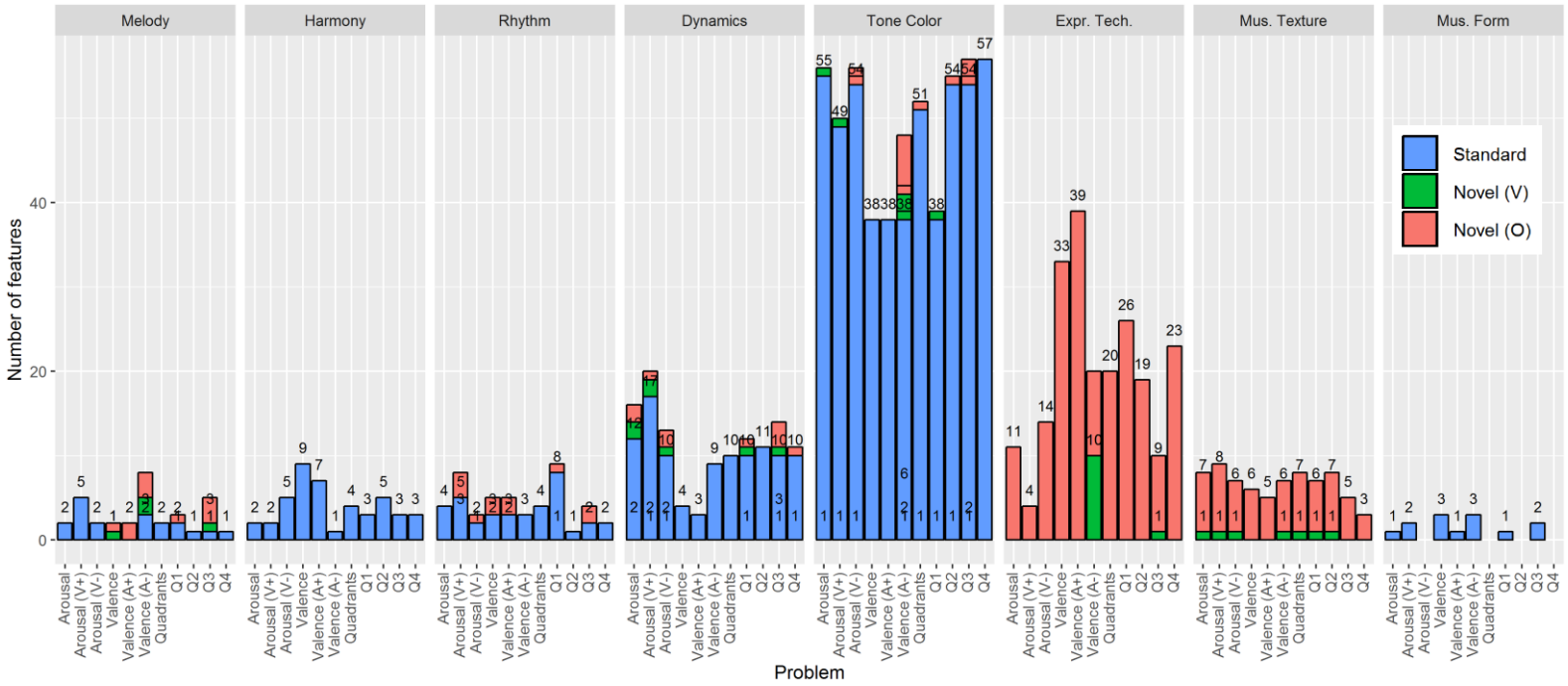


Figure 4.25: The top 100 features selected by ReliefF for each emotion classification problem studied, organized by musical dimension.

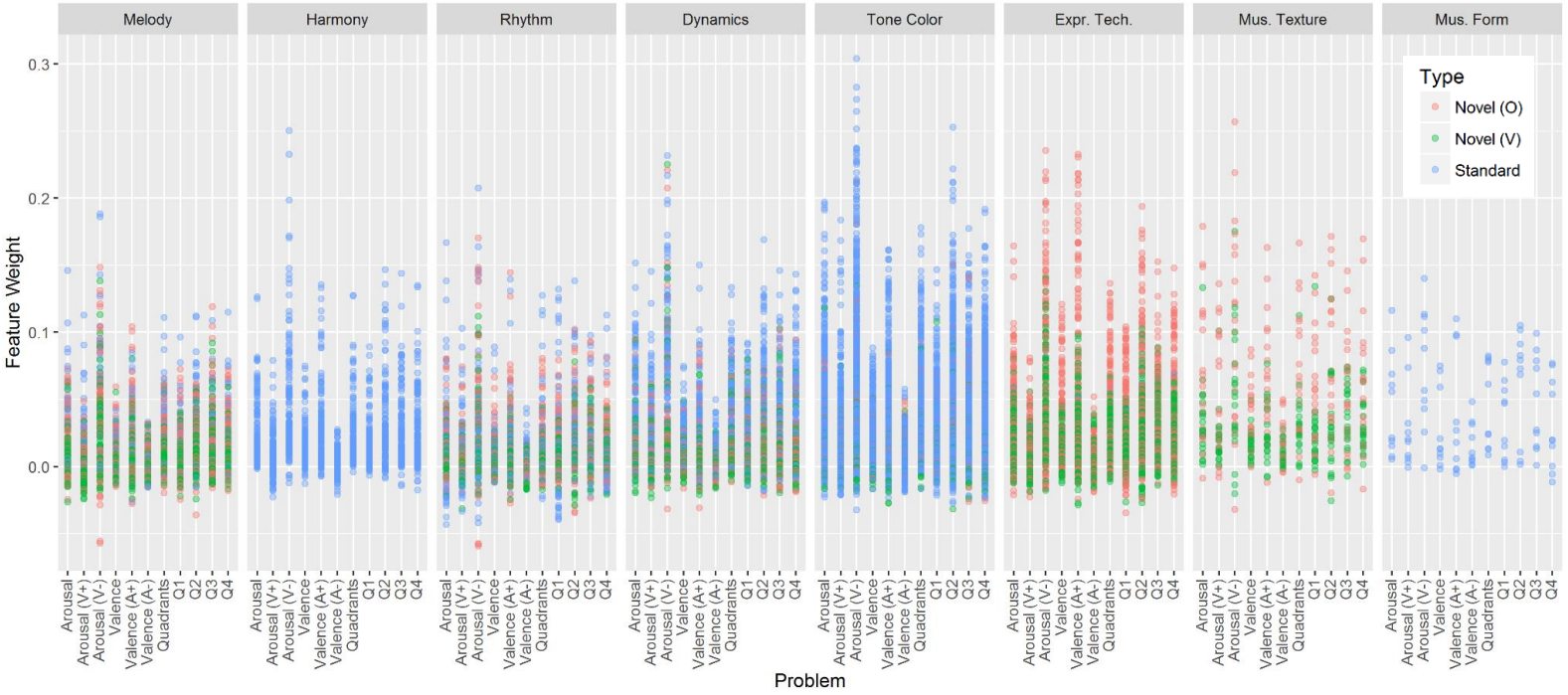


Figure 4.26: Feature weight of the entire feature set, grouped by musical dimension and divided by problem (each point is a different feature).

Chapter 5

Other Experiments

In addition to the main contributions described in previous chapters, a number of other experiments were also conducted at an earlier stage of the our research work. Although these are not major contributions, each of these studies had a specific purpose and, together, laid the foundations that enabled the results described in the previous chapters. Some of these consisted in the construction of other datasets, bi-modal approaches exploring audio and lyrics or the analysis of existent works and datasets. In these chapters, we describe some of these additional experiments as well as the results attained.

Section 5.1. Evaluation of MER Strategies and Datasets

Under this perspective, our initial work was focused on replicating state-of-the-art works, evaluating existing datasets and exploring both dimensional and categorical MER approaches. These experiments gave us a valuable idea of the main limitations in the field.

Section 5.2. Emotion-based Playlist Generation

Building on our music emotion recognition experiment using dimensional models, we explored the automatic generation of music playlists based on emotion. As a result, a computational prototype to demonstrate this functionality was built and made available online.

Section 5.3. MER Multi-modal Approaches

Additionally, a third experiment has been conducted to assess the influence of lyric content in music emotion and how combining different modalities might improve the prediction accuracy.

5.1. Evaluation of MER Strategies and Datasets

In the beginning of this work, one of the strategies to gain a deeper understanding of the MER field was to study and replicate the published state-of-the-art. To this end, we first selected the AV-based regression approach proposed by Yang et al. (2008), later studying categorical approaches such as the MIREX taxonomy (Hu & Downie, 2007).

5.1.1. Yang's Dimensional Approach

As discussed in detail in Section 3.2.5, the authors approached MER as a regression problem, using Thayer's model[118] to classify emotions in terms of arousal and valence (AV). To this end, the authors employed a dataset containing 195 song clips, collected in their previous study (Y.-H. Yang et al., 2006) and annotated by 253 subjects, with at least 10 subjects annotating each clip. A deeper analysis on this dataset was presented in Section 3.2.2. From the audio clips, a total of 114 features were extracted, mostly using PsySound2 and Marsyas and reduced with PCA and RReliefF, obtaining an R^2 of 58.3% for arousal and 28.1% for valence.

We built on Yang's work, using the same dataset to test an extended feature set and gain a better knowledge of the current limitations regarding dimensional approaches.

Feature Extraction

In our initial series of tests (Panda & Paiva, 2011a), we selected PsySound, the MIR Toolbox (MIRT) and Marsyas (MAR), described in Section 3.1.1, to measure the relevance of each one in MER. In their work (Y.-H. Yang, Lin, Su, et al., 2008), Yang et al. used PsySound2, an older version that is available only for the Mac PowerPC architecture. Since then, the program was rewritten in MATLAB, resulting in PsySound3. This version contains inconsistencies and lacks documentation, making it very hard to replicate the exact Yang's feature set and thus compare the results between PsySound2 and 3. For this reason, we employed the same PsySound2 features extracted and kindly provided by Yang[119]. A set of 15 features from PsySound2 were previously identified by the authors as particularly relevant to emotion analysis, hereafter denoted as Psy15, while the full set is denoted Psy44.

A brief summary of the extracted features and their respective framework is given in Table 5.1. Regarding Marsyas and the MIR Toolbox, the analysis window size used for

[118] Although the original article cites the Thayer's model, the authors used an AV model which is closer to the circumplex model of emotion proposed by Russell.

[119] http://mac.citi.sinica.edu.tw/~yang/MER/taslp08/

frame-level features is 23 ms, later transformed to song-level features using mean and variance, i.e., MeanVar model, (Meng, Ahrendt, Larsen, & Hansen, 2007). All extracted features were normalized to the [0, 1] interval. A total of 12 features extracted with Marsyas returned the same (zero) value for all songs and, thus, were not used in this experiment.

In a later experiment (Panda, Rocha, & Paiva, 2013), we added an additional set of 51 melodic features (MF), consisting of statistics extracted from the pitch contour of the melody using a transcription step, as described in (Salamon et al., 2012). Such features are related with the typology of the pitch contour, as described in (Adams, 1976), as well as vibrato, pitch and duration information of these contours.

Framework (# of features)	*Features*
Marsyas (237)	Spectral centroid, rolloff, flux, zero-crossing rate, linear spectral pair, linear prediction cepstral coefficients (LPCCs), spectral flatness measure (SFM), spectral crest factor (SCF), stereo panning spectrum features, Mel frequency cepstral coefficients (MFCCs), chroma, beat histograms and tempo.
MIR Toolbox (177)	Among others: root mean square (RMS) energy, rhythmic fluctuation, tempo, attack time and slope, zero-crossing rate, rolloff, flux, high-frequency energy, MFCCs, roughness, spectral peak variability (irregularity), inharmonicity, pitch, mode, harmonic change and key.
PsySound2 (44)	Loudness, sharpness, volume, spectral centroid, timbral width, pitch multiplicity, dissonance, tonality and chord, based on psychoacoustic models.
Melodic (51)	Pitch (mean pitch height, pitch deviation, pitch range and interval), duration (in seconds), vibrato (rate, extent, coverage), pitch contour typology statistics.

Table 5.1: Frameworks used and respective features.

Feature Selection and Emotion Regression

A wide range of supervised learning methods are available and have been used in regression problems before. The idea behind regression is to predict a real value, based on a previous set of training examples. Since we are using a continuous representation of emotion, two distinct regression models need to be trained – one for arousal and another for valence. Three different supervised machine techniques were evaluated: Simple Linear Regression (SLR) (Lane, 2009), K-Nearest Neighbors (KNN), and Support Vector

Regression (SVR). These algorithms were run using both Weka[120], a suite of machine learning algorithms developed in Java, and the libSVM library[121].

In order to assess the importance of each feature and to improve results, while reducing the feature set size at the same time, feature selection and ranking was also performed. To this end, the RReliefF algorithm and Forward Feature Selection, both described in Section 3.2.3, were used. In RReliefF, the resulting feature ranking was then tested to determine the number of features providing the best results. This was performed following an embedded approach, by adding one feature at a time to the set and evaluating the corresponding results. The best top-ranked features were then selected.

All experiments were validated using 10-fold cross validation with 20 repetitions, reporting the average obtained results. Moreover, parameter optimization was performed, e.g., grid parameter search in the case of SVR. In order to compare our results with the original study, the performance of the regression models were measured using R^2 statistics. As previously mentioned, this metric represents the coefficient of determination, "which is the standard way for measuring the goodness of fit for regression models" (Y.-H. Yang, Lin, Su, et al., 2008).

Results and Discussion

Our first experiments did not include the melodic features and were based only on SVMs for regression (Panda & Paiva, 2011a). There, the best results were 63% for arousal and 35.6% for valence, using a total of 53 and 80 features respectively. This subset of features was obtained from the combination of the features from the three frameworks, reduced with the FFS algorithm. Although the results obtained with the RReliefF feature selection method were lower, they were mostly obtained resorting to less features, helping us to identify the most important features for both problems (AV). For instance, using only the first ten features selected with RRF resulted in 31.5% for arousal and 15.2% for valence. On the other hand, the same number of features with FFS achieved only 0.8% and 2.0% for arousal and valence respectively.

Testing individual frameworks separately highlighted MIR Toolbox as containing the best suited features, especially for valence with an R^2 of 25.7%. PsySound followed, with a valence accuracy of 21% and Marsyas features set scored the lowest, only 4.6% using FFS and 10.4% with ReliefF, proving to be less effective for valence prediction in this dataset at the time. In terms of arousal, all the frameworks had a close score, ranging from 56% (Marsyas) to 60.3% (MIR Toolbox). A summary of the results is presented in Table 5.2. For unknown reasons, we were unable to replicate the exact results obtained in (Y.-H. Yang, Lin, Su, et al., 2008), using either the Psy15 features or the list of features

[120] https://www.cs.waikato.ac.nz/ml/weka/

[121] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

resulting from the feature selection algorithm[122]. We also conducted the same tests with PCA, normally used to reduce correlation between variables, without any noticeable improvement in results but actually leading to lower R^2 values.

By adding the melodic set of features in our second experiment (Panda, Rocha, et al., 2013), we were able to increase the best results to 67.4% for arousal and 40.6% for valence (although these features alone performed poorly). Such results are a significant improvement over the original study 58.3/28.1 % (Y.-H. Yang, Lin, Su, et al., 2008) as well as our previous results of 63/35.6%. Regarding the different classification strategies tested, SVM was always the best performing algorithm. A summary of the results is condensed in Table 5.2.

Machine Learning	*Feature Set*	*Arousal (R^2)*	*Valence (R^2)*
SVR	Psy15 (all features – all)	58.7%	12.7%
SVR	Psy15 (feature selection – FS)	60.1%	21.1%
SVR	Psy44 (all)	57.3%	7.9%
SVR	Psy44 (FS)	57.3%	19.1%
SVR	MIRT (all)	58.2%	8.5%
SVR	MIRT (FS)	58.7%	25.7%
SVR	MAR (all)	52.9%	3.7%
SVR	MAR (FS)	60.0%	10.4%
SVR	Psy44+MIRT+MAR (all)	57.4%	19.4%
SVR	Psy44+MIRT+MAR (FS)	62.9%	35.6%
SVR	MF (all)	45.2%	2.7%
SVR	MF (FS)	50.0%	2.6%
SVR	Psy44+MIRT+MAR+MF (all)	58.0%	16.3%
SVR	Psy44+MIRT+MAR+MF (FS)	67.4%	40.6%
SLR	Psy44+MIRT+MAR+MF (all)	42.2%	-1.3%
SLR	Psy44+MIRT+MAR+MF (FS)	54.6%	3.3%
KNN	Psy44+MIRT+MAR+MF (all)	56.8%	1.5%
KNN	Psy44+MIRT+MAR+MF (FS)	61.1%	12.0%

Table 5.2: Regression results obtained with different machine learning algorithms and feature set combinations, from (Panda & Paiva, 2011a; Panda, Rocha, et al., 2013).

[122] It is worth mention that, in order to try to replicate Yang et al. results, we employed the SVR parameters described at the paper web page: http://mpac.ee.ntu.edu.tw/~yihsuan/MER/taslp08/.

In addition to the performance improvements, we identified features related with tone color (e.g., MFCCs, RMS energy, and spectral skewness), rhythm (e.g., pulse clarity) or loudness as some of the most relevant to arousal, while valence was more influenced by tone color (spectral dissonance, MFCCs), rhythm (pulse clarity and fluctuation) and harmony (tonality, key strength and clarity) (Panda & Paiva, 2011a).

Considerations on Dimensional Emotion Recognition

Our MER experiments using dimensional models yielded several interesting conclusions. First, although very useful and having diverse applications, the performance achieved with dimensional models is still weak, especially regarding valence prediction. To make this worse, there are no publicly available dimensional MER datasets, which can be said to have remarkable quality (e.g., sizeable, quality controlled annotations), as already noted in Section 3.2.2.

As previously detailed in Section 3.2.1 and supported by Yang's et al. description (Y.-H. Yang, Lin, Su, et al., 2008), the dimensional ground-truth acquisition is a process of high complexity. After all, in addition to the high number of subjects required (e.g., Yang et al. used 253 subjects to annotate 195 audio clips), it is also more error-prone, since there is an extra layer of possible confusion where subjects need to convert the terms usually associated with emotion (e.g., "happiness") to numeric values of an intricate dimensional model (as discussed earlier in Section 2.2).

Regarding the dataset employed in these experiments (Y.-H. Yang, Lin, Su, et al., 2008), one of the few that we could obtain having a well-documented and planned annotation gathering process, several problems were identified. First, according the authors, the users were asked to annotate emotions evoked by music (instead of perceived), which tends to have a much lower inter-subjective agreement, being more context, culture and memory dependent, as discussed in Section 2.1.2. In addition, the subjects were requested to consider both audio and the actual lyrics sang by the performers, which as we found in Section 4.6, can greatly influence the valence of a song, especially for lower arousal songs. Still, the authors did not consider any lyrical information in their experiments.

Other problems with this dataset were already highlighted in the review presented in Section 3.2.2. Namely, the unbalanced nature of the 195 songs over the four AV quadrants. Upon closer analysis, we now know that the 195 clips resulted from a previous study by the authors (Y.-H. Yang et al., 2006), where quadrant annotations of 243 clips where gathered, discarding low agreements (less than 50%) clips. According to these annotations, the 195 were balanced, with 48 to 49 clips assigned to each quadrant. However, in their second study (Y.-H. Yang, Lin, Su, et al., 2008), the same clips were annotated in a significantly different and very unbalanced way by different subjects (e.g., quadrant 2 contains only 21 songs). These inconsistencies raises concerns on the quality

of the dataset and further highlights the difficulties in ground-truth acquisition. Moreover, most of these clips were placed very close to the origin, as illustrated in Figure 5.1, while such emotion models state that emotions are expected to be far from it. One possible hypothesis for this are very different annotations by subjects to the same songs, which when averaged end up close to the center.

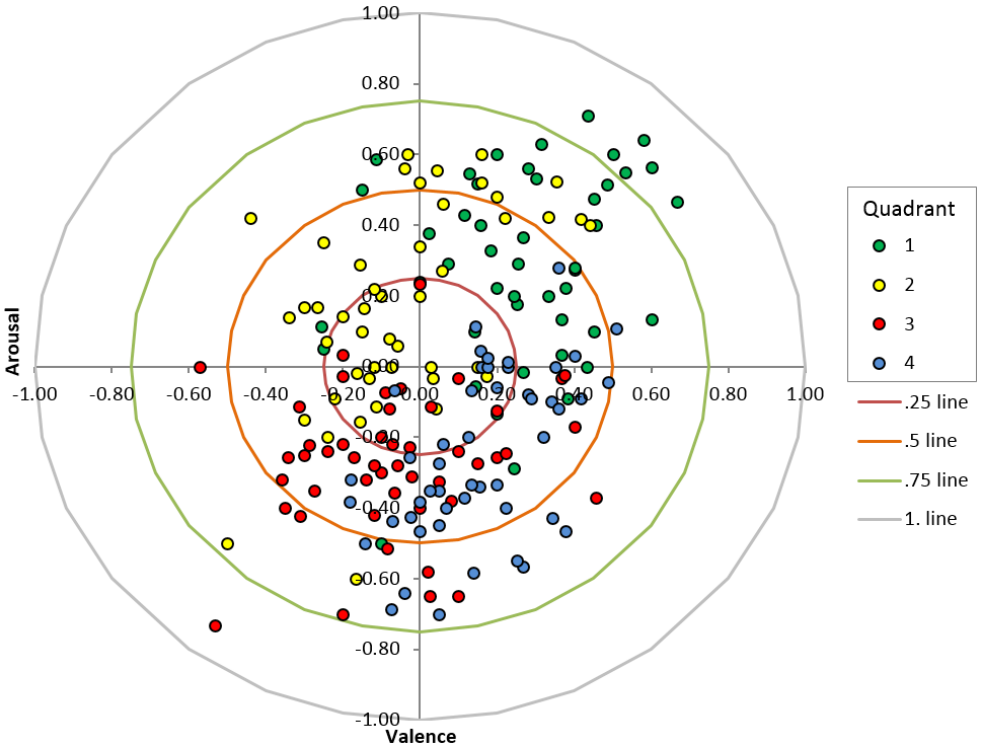


Figure 5.1: Yang et al.'s AV dataset annotations placed on the Russell's emotion model (2008). Different colors indicate the quadrants originally assigned to the clips in the authors' previous study (Y.-H. Yang et al., 2006).

To conclude, given the lack of sizeable high-quality dimensional datasets and the abovementioned difficulties to construct one, added to the above average results of previous dimensional MER studies, especially when predicting valence, we believe it is wiser to first improve MER classification using simpler emotion taxonomies, before tackling these higher complexity problems.

5.1.2. MIREX Categorical Approach

Similarly to the experiments carried out with dimensional models in MER, we also investigated categorical models by reviewing the relevant literature (see Section 3.2.5) and experimenting with MER as a categorical problem. In contrast to the dimensional counterpart of MER, where the AV emotion model has been widely adopted as the *de facto* approach, categorical studies have been much more heterogeneous. In that sense, MER researchers have not agreed on a categorical model, with many studies using different taxonomies containing different ranges of diverse words. As summarized in Table 3.4 of Section 3.2.5, the solutions vary from models with few categories (e.g., selected based on the authors' view, on AV quadrants or Ekman's basic emotions) to a very high number (e.g., derived from dictionaries and online data).

Taking into account the observed dispersion in categorical MER studies and the mentioned lack of quality categorical datasets (see Section 3.2.2), we decided to start by creating a new audio dataset using the MIREX AMC taxonomy. This choice was based on the fact that, despite the limitations highlighted previously, it is widely recognized and has been used in the MER field as the standard benchmark for comparing categorical approaches.

As explained in Section 2.2.1, the taxonomy used in the MIREX AMC task was derived from song metadata provided by AllMusic[123] (Hu & Downie, 2007). At the time, a total of 179 emotional tags were used at the site, said to have been compiled and assigned by experts. This process consisted in three steps: 1) compute the similarity between emotion tags, according to the number of songs containing each of the possible pairs of tags; 2) next, the similarity data was clustered into groups of emotion tags using agglomerative hierarchical clustering; 3) finally, the obtained clusters were manually analyzed, with the authors selecting the five clusters and 29 emotion tags (represented in Table 2.2) that nowadays form the taxonomy. The taxonomy was then used to create the private MIREX AMC dataset, containing 600 30-second audio clips in 22.05 kHz monaural WAV format, annotated by 2 to 3 human judges, as discussed in Section 3.2.2.

MIREX-like Dataset Construction

Having set out to replicate the MIREX dataset organization, it was only natural to also select the AllMusic service as the source of our data. To this end, we built a set of scripts to automate the collection of 30 second audio clips that were tagged with the MIREX emotion tags from the AllMusic website. As a result, a total of 1335 clips belonging to the 29 tags described in MIREX were obtained and organized into the five clusters. Since AllMusic contains multiple tags per song, some of the obtained clips were related to various emotion tags and clusters. Such clips were, hence, removed, reducing the dataset

[123] https://www.allmusic.com

to 903 clips.

As noted, the MIREX AMC dataset clips were labeled based on the agreement between two or, in some cases, three experts (Hu et al., 2008). However, due to the resources required to replicate a manual annotation process, and since the AllMusic emotion tags are said to also have been assigned by music experts (Hu & Downie, 2007; Pao et al., 2008; Y.-H. Yang & Hu, 2012), we decided to use their data directly.

Concerning its organization, the dataset is relatively balanced between clusters, with a slight advantage for clusters 3 and 4, as shown in Figure 5.2, due to the removal of the ambiguous songs. Another relevant aspect of the dataset is that, as previously pointed out in Sections 2.2.1 and 3.2.2, there is a semantic overlap (ambiguity) between clusters 2 and 4, and an acoustic overlap between clusters 1 and 5 (Laurier & Herrera, 2007).

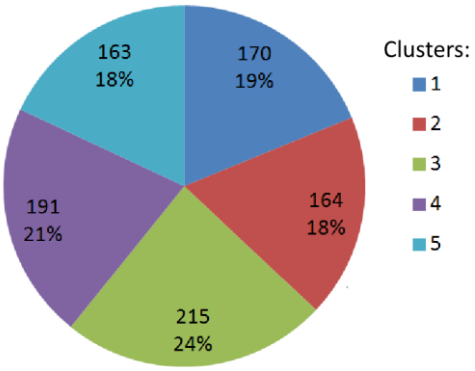


Figure 5.2: MIREX-like dataset audio clips distribution between the five clusters.

The initial version of the dataset (Panda & Paiva, 2012b) was later incremented with lyrics and MIDI files (Panda, Malheiro, et al., 2013). To this end, we developed tools to automatically search for lyrics and MIDI files of the same songs using the Google Search API. In this process, three sites were used for lyrical information (lyrics.com, ChartLyrics and MaxiLyrics), while MIDI versions were obtained from four different sites[124]. After removal of some defective files, the interception of the 903 original audio clips with the lyrics and MIDIs resulted in a total of 764 lyrics and 193 MIDIs. In fact, MIDI files proved harder to acquire automatically. As a result, we formed 3 datasets: an audio-only dataset with 903 clips, an audio-lyrics dataset with 764 audio clips and lyrics (not evaluated here) and a combined multi-modal dataset with 193 audio clips and their corresponding lyrics and MIDIs. All datasets were nearly balanced across clusters (maximum and minimum representativeness of 25% and 13%, respectively).

Even though the final multi-modal dataset is smaller than desired, this approach

[124] freemidi.org, free-midi.org, midiworld.com and cool-midi.com.

demonstrated that we can exploit the specialized human labor of the AllMusic annotations to automatically acquire a music emotion dataset, reducing the required resources. Moreover, the proposed method is sufficiently generic to be employed in the creation of different emotion datasets, with different emotion adjectives than the ones used in the MIREX AMC taxonomy. This MIREX-like dataset is available at our website[125] to any researchers willing to use it in future research.

Although the two datasets (the original MIREX AMC dataset and the proposed MIREX-like one) have similarities in organization, they still differ in not negligible aspects such as the audio clips selection and annotation process. Thus, results obtained with ours must be analyzed or compared with this in mind. To have a possible point of comparison, we used this dataset to develop and run several MER experiments, as described in Section 5.1.2, submitting some of our solutions to the MIREX AMC annual comparison task. As a result, one of our models placed first in that year[126] (in a total of 20 submissions from 10 different research groups).

These experiments served as the learning base to the dataset construction described in Chapter 4.

Music Emotion Classification Experiments

As with our previous experiments, we used Marsyas, the MIR Toolbox and PsySound to extract a total of 253 audio features from the clips. Some compromises had to be done in feature extraction when compared to our regression experiments in order to accomplish our goal of creating a MER solution that could be submitted to MIREX AMC contest. Namely, we used a smaller set of 11 PsySound3 features, since version 2 is no longer usable and the framework is highly resource-intensive, exceeding the contest time limits. In addition, we also reduced the set of features extracted with Marsyas to 65 (centroid, rolloff, flux, MFCCs and tempo) since the remaining were less stable and did not prove relevant.

Support Vector Machines was the preferred classification algorithm, based on results from previous experiments (Panda & Paiva, 2011b; J.-C. Wang et al., 2010). The libSVM library (Chang & Lin, 2011) was the selected implementation, providing a fast and reliable implementation of SVMs. A grid parameter search was also carried out to retrieve the best values for parameters γ and C (cost), used by the radial basis function (RBF) kernel of the SVM model. Some additional tests using other algorithms such as KNN were conducted with lower classification results.

In order to reduce the feature set and achieve a subset of features that are most suitable to our problem, ReliefF was used. For comparison with the MIREX AMC task,

[125] http://mir.dei.uc.pt/

[126] http://www.music-ir.org/nema_out/mirex2012/results/act/mood_report/summary.html

some classification tests were run using 3 and 5-fold cross validation due to the fact that MIREX uses a 3-folds setup. However, our analysis was done with 10-fold cross validation since, according to the literature, "there are more performance estimates, and the training set size is closer to the full data size, thus increasing the possibility that any conclusion made about the learning algorithm(s) under test will generalize to the case where all the data is used to train the learning model" (Refaeilzadeh et al., 2009, p. 536). Using fewer folds did not influence the results significantly.

The best classification results, an F1-Score of 47.21%, were obtained with a subset of features from all the three frameworks, selected by the ReliefF feature selection algorithm. Although the top result requires a high number of features, 39 features are enough to obtain a very close 47.2% F1-Score. In the same direction, only 19 features are needed to obtain 95% of that maximum. Of the three frameworks, the MIR Toolbox obtained the best result and, while PsySound3 score was lower, it is important to highlight that only 11 features were available from it. Finally, using feature selection served to reduce the number of features needed but did not improve the results. A brief summary of these results is presented in Table 5.3.

Feature Set	*F1-Score*	*Precision*	*Recall*
Marsyas (65)	41.52%	40.71%	42.37%
MIR toolbox (177)	44.43%	44.05%	44.81%
PsySound 3 (11)	36.37%	35.64%	37.13%
All Frameworks	47.21%	46.86%	47.60%
All Frameworks (ReliefF)	47.20%	46.82%	47.57%

Table 5.3: Classification results with the MIREX-like audio dataset.

Based on the ranking obtained using the ReliefF algorithm, the 10 most relevant audio features to this classification problem were related with harmony (key strength and clarity, mode, tonal centroid), tone color (MFCCs, rolloff, zero crossing rate, tonal dissonance, high-frequency energy) and rhythm (tempo). As can be noted, some of those were already relevant in our regression experiments.

Finally, the confusion matrix resulting from the best classification model, using features from the three frameworks is listed in Table 5.4. As shown, a considerable percentage of songs were wrongly classified between clusters 1 – 5 and 2 – 4. This seems to go in the direction of the previously identified semantic and acoustic overlap between these same clusters. In Table 5.5, a confusion matrix grouping the overlapped clusters is presented.

		Predicted				
		C1	C2	C3	C4	C5
Annotated	C1	42.36	18.27	8.01	16.66	21.84
	C2	19.79	44.03	14.63	17.33	6.80
	C3	4.08	12.43	54.62	13.98	7.39
	C4	14.13	21.01	17.20	38.99	10.97
	C5	19.64	4.26	5.54	13.04	53.00

Table 5.4: Confusion matrix (results are in %).

		Predicted		
		C1+5	C2+4	C3
Annotated	C1+5	68.82	27.46	13.55
	C2+4	25.34	59.05	31.83
	C3	5.84	13.50	54.62

Table 5.5: Confusion matrix merging the clusters with semantic and acoustic overlap (results are in %).

Our results in this dataset were lower (47.2%) than the top results obtained in the previous editions of MIREX AMC task. Still, it is hard, to draw definitive conclusions from these results, since two different datasets were used. As a possible point of comparison, we conducted tests in our dataset extracting the same feature set used in one of the submissions to MIREX AMC 2008 and 2010 tasks (based on Marsyas, since the code is available in the framework). This set of features achieved only 40.71% precision in our dataset, while obtaining 48.58% to 57.5% precision in MIREX AMC 2010. This fact suggests that our approach would have better accuracy in the MIREX dataset, possibly due to the higher quality of its annotations, which were created by a panel of three experts, while our dataset used AllMusic annotations directly, which were created by experts but very few details exist describing the procedure. This was confirmed by our MIREX 2012 AMC submission, where an accuracy of 67.83% was achieved.

Analysis of the MIREX AMC Taxonomy

As mentioned, the MIREX AMC taxonomy and dataset has been said to contain possible problems related to semantic and acoustic overlaps (Laurier & Herrera, 2007). Moreover, the final taxonomy of 29 adjectives organized into five clusters was built using a

data-driven approach using data from the AllMusic website. This raises questions about its scientific validity, especially since some of the adjectives used (e.g., poignant, literate, autumnal, campy, volatile) are not present in categorical emotion models from psychology (e.g., basic emotions by Ekman) nor are commonly used by people when describing their emotions (or emotions in music). For this reason, and since it is annually used to compare MER progress, we decided to further analyze the taxonomy to assess its quality.

As described in detail in Section 2.2.1, the MIREX Audio Mood Classification (AMC) task taxonomy was automatically derived from the relations found in the music tags of AllMusic service data. To this end, the authors began by gathering a list of songs and albums labelled with each of the 179 mood tags available at the time in AllMusic. Next, the data was reduced to 40 tags, by removing the less frequently used ones. The similarity between pairs of tags was assessed by the number of songs and albums associated with both tags. Finally, this similarity data was fed into clustering algorithms, with the authors selecting the 29 adjectives that were commonly clustered together. The resulting taxonomy is present in Table 5.6.

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
Rowdy	Amiable / Good natured	Literate	Witty	Volatile
Rousing	Sweet	Bittersweet	Whimsical	Visceral
Boisterous	Fun	Autumnal	Wry	Aggressive
Passionate	Rollicking	Brooding	Campy	Tense / Anxious
	Cheerful	Poignant	Quirky	Intense
			Silly	

Table 5.6: Emotion taxonomy used in the MIREX Audio Mood Classification task.

As stated, some authors have pointed possible problems in the dataset. Namely, a semantic overlap (ambiguity) between clusters 2 and 4 (e.g., words fun (cluster 2) and humorous (cluster 4) are close and share the synonym amusing (Laurier & Herrera, 2007). As for songs from clusters 1 and 5, there are acoustic similarities: both tend to be energetic, loud, and many use electric guitar (Laurier & Herrera, 2007). Moreover, the number of words per cluster is unbalanced.

Hence, we devised a set of experiments to further assess the validity of the MIREX taxonomy. Our first experiment tested the hypothesis that, if the taxonomy is good we expect to have words closer in meaning (i.e., synonyms) in the same cluster, while antonyms should be in different clusters. To this end, a list of synonyms and antonyms of

the 29 words and the respective synonym relevance was obtained from Thesaurus dictionary[127]. Next, we analyzed the relations between these words, as illustrated in Figure 5.3. The black arrows represent synonyms and red is used for antonyms, while the thickness represents the relevance of the relation, according to Thesaurus.

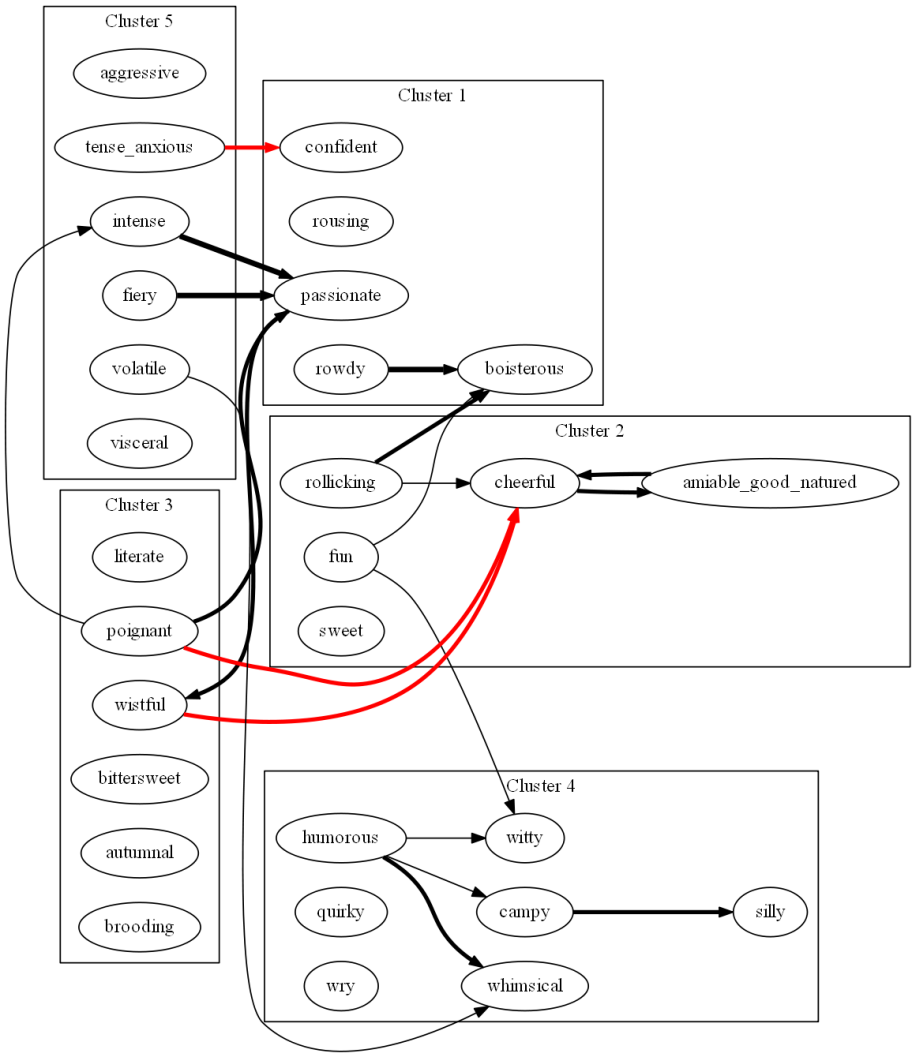


Figure 5.3: MIREX AMC taxonomy cluster similarities based on Thesaurus synonyms and antonyms.

As shown, more synonyms are found between extra-cluster words (10) than in same

[127] https://www.thesaurus.com/

cluster words (8). This is especially visible in cluster 1, where the word passionate has highly relevant synonyms in cluster 5 (intense, fiery) and cluster 3 (poignant, wistful). Still, these results should be taken with caution since the number of synonyms and antonyms found between the 29 words is low. A possible improvement to this method would be to assess similarity not only using direct synonyms but also considering first level common synonyms (i.e., synonyms directly shared by these 29 words but not directly present in the taxonomy).

A second experiment was designed to test the same hypothesis using AllMusic similarity data. Since the creation of the MIREX AMC taxonomy, the emotion tag data in AllMusic has been increased from 179 to 289 words. Furthermore, each AllMusic tag is also associated with a list of similar emotion tags. As an example, the tag fun[128] is said to be similar to boisterous, humorous, quirky, rollicking, rowdy, silly, whimsical, and witty. This supplementary similarity information was used to evaluate the MIREX AMC taxonomy in a similar fashion to experiment 1, i.e., by assessing the number of similar pairs of words (tags) inside and outside each of the five clusters, as illustrated in Figure 5.4. There, the black edges (arrows) represent two words (emotions) said to be similar in the same cluster (intra-cluster), while extra-cluster ones are represented in red. The three numbers below each word represent the number of intra-cluster connections (similar emotion tags), number of extra-cluster similarities and similar moods unavailable in the 29 words of the MIREX taxonomy.

A high-quality taxonomy is expected to have high number of similar emotions inside the clusters and few between different clusters, especially considering that in this case both the taxonomy being analyzed and the similarity data are from the same source. However, a high number of extra-cluster connections was observed (in red). Computing the ratio of intra-cluster to extra-cluster connections for each cluster shows that two of the clusters have a ratio lower than 1, meaning more extra-cluster connections (cluster 1 with 0.7 and cluster 2 with 0.75). The remaining, especially cluster 3, contain a higher number of intra connections (cluster 3 = 10.0, cluster 4 = 2.8 and cluster 5 = 2.1).

[128] https://www.allmusic.com/mood/fun-xa0000001006

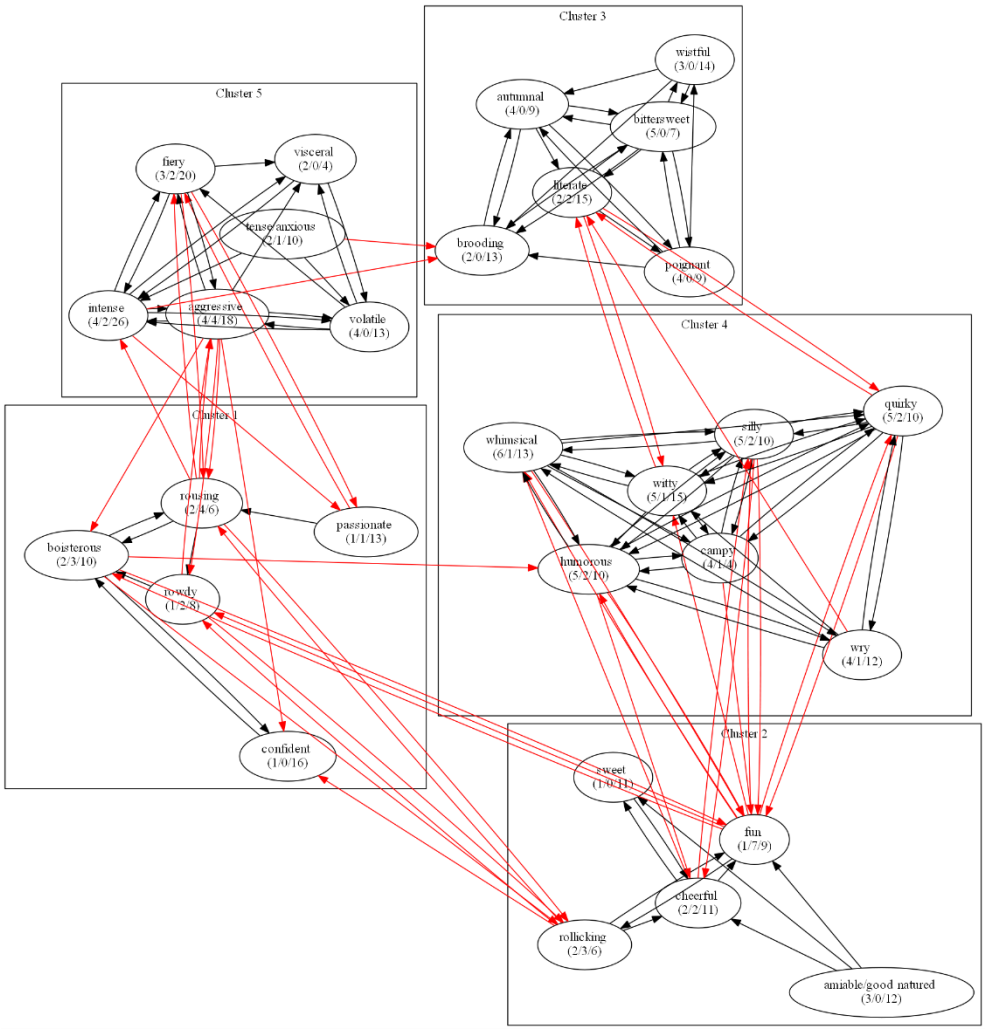


Figure 5.4: Cluster similarities in MIREX Audio Mood Classification taxonomy.

Finally, the third experiment assessed the quality of MIREX taxonomy using data from Warriner et al. (2013), a psychology study containing arousal, valence and dominance (AVD) values for 13915 English words. Here, our aim was to use the AVD values of the 29 MIREX AMC words and cluster them into five groups with standard clustering algorithms based on their proximity in the 3-dimensional space. This would allow us to compare the five clusters of both approaches.

Unfortunately eight of the MIREX words are missing from Warriner's list: boisterous, good natured, rollicking, literate, autumnal, wry, campy and visceral, which reduces the validity of such approach. Still, we conducted the experiment using the k-means

algorithm with Euclidean distance and the five obtained clusters are described in Table 5.7 and illustrated in Figure 5.5. In the figure, the text in each point indicates the word and original MIREX cluster (e.g., "fun:C2"), while the colors indicate the newly generated clusters (e.g., K2 in orange).

k-means	*MIREX*	*Word*	*Arousal*	*Valence*	*Dominance*
K1	C2	Amiable	-0.54	0.43	0.29
K1	C2	Sweet	-0.22	0.69	0.28
K1	C4	Whimsical	-0.06	0.41	0.08
K1	C4	Quirky	-0.16	0.36	0.21
K1	C4	Silly	-0.04	0.43	0.38
K2	C1	Rousing	0.21	0.09	0.14
K2	C5	Intense	0.40	0.16	0.14
K3	C3	Bittersweet	-0.20	0.02	-0.12
K3	C3	Poignant	-0.23	0.03	0.36
K4	C1	Passionate	0.33	0.54	0.41
K4	C2	Fun	0.33	0.84	0.51
K4	C2	Cheerful	0.19	0.75	0.58
K4	C4	Witty	0.16	0.56	0.39
K5	C1	Rowdy	-0.03	-0.21	-0.08
K5	C3	Brooding	-0.25	-0.43	-0.21
K5	C5	Volatile	0.09	-0.39	-0.09
K5	C5	Aggressive	0.22	-0.48	0.12
K5	C5	Tense	0.08	-0.56	-0.07
K5	C5	Anxious	0.30	-0.30	-0.21

Table 5.7: Emotion taxonomy obtained by clustering the MIREX Audio Mood Classification task words using AVD values from Warriner's list.

As illustrated, the clusters obtained with AVD values significantly differ from the original data-driven taxonomy. Words such as passionate, witty and cheerful (C2) or intense and rousing (C5), among others, belong to distinct clusters (MIREX) but were now grouped together due to their spatial proximity. As stated before, these results should be taken with caution since eight words were left out and their addition would change the clustering results. Nonetheless, this does not change the major issue found:

several MIREX AMC words are in different clusters but psychology studies seem to place them close in terms of their emotional value.

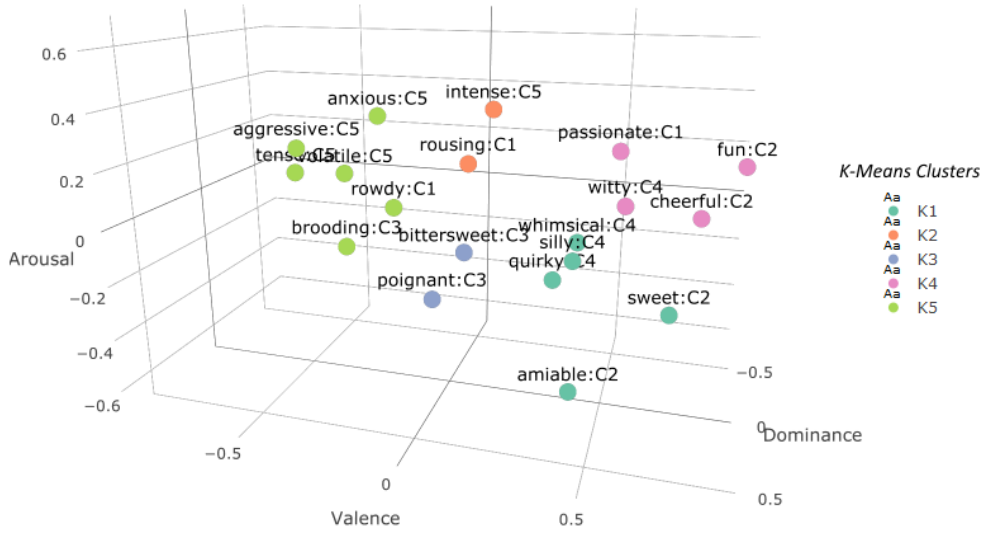


Figure 5.5: The five clusters obtained by clustering MIREX AMC words with Warriner's AVD values (using k-means).

To conclude, our experiments, based on Thesaurus synonyms, the updated AllMusic similarity data and psychology studies, seem to suggest higher than expected inter-cluster similarity in the MIREX AMC taxonomy and, thus, a new taxonomy and dataset should be proposed in future editions of the evaluation exchange.

5.1.3. Conclusions and Uncovered Paths

The experiments described in this section gave us valuable insights, uncovering possible paths that were explored in Chapter 4 in our goal to improve MER. First, it showed us that approaching MER as a dimensional problem using regression is still quite an intricate problem, especially regarding valence regression. Also, the resources needed to create a sizeable, high-quality dataset to it are high and unavailable to many researchers. Considering that the best performing categorical MER systems are still struggling with low number of classes (as shown by the results in the MIREX AMC task), we believe that to improve MER we first need to focus on improving classification in these cases.

Secondly, our analysis also demonstrated several inconsistencies with the MIREX taxonomy, in addition to the previously identified ones, which tells us that: 1) maybe it

is time to update the MIREX AMC task with a newer and higher quality dataset; and 2) we should pursue music emotion classification with a different, better supported categorical taxonomy.

5.2. Emotion-based Playlist Generation

As part of our MER experiments, we built on our work with dimensional models, discussed in Section 5.1.1, and explored the problem of emotion-based automatic playlist generation (APG). As part of this work we built a computational prototype to demonstrate the concept.

Briefly, the rationale behind the approach is to create music playlists based on the Euclidean distance between songs on the Russell's AV space. Such playlists can be generated using a seed song or using an emotional track (set of points in the AR space) obtaining the N songs closer to it/them. This approach and results are presented below.

5.2.1. Feature Extraction and Emotion Modeling

Following our dimensional experiment in Section 5.1.1, our APG experiments used the same dataset of 195 audio samples and AV annotations (Y.-H. Yang, Lin, Su, et al., 2008). As in Section 5.1.1, we used an equal set of features, extracted with Marsyas, the MIR Toolbox and PsySound2 audio frameworks.

The regression strategy was based on Support Vector Regression (SVR), since it achieved the best results in Yang's study (Y.-H. Yang, Lin, Su, et al., 2008). Again, we used the libSVM library (Chang & Lin, 2011). A grid parameter search was also carried out to discover the best SVR parameters. As with the previous experiments, Forward Feature Selection and RReliefF (ReliefF variant for regression) were used for feature selection, while the performance of the regression models was measured with the R^2 statistics.

5.2.2. Experimental Results

As mentioned before, a regressor-based distance strategy was employed to evaluate playlist quality. In this method, distances are calculated using the predicted AV values returned by the regression models. The predicted distances were compared to the reference distances resulting from the real AV annotations.

To this end, the dataset was randomly divided into two groups, balanced in terms of quadrants. The first, representing 75% of the dataset was used to train the regressor.

Next, the resulting model was used to predict AV values for the remaining 25% songs[129]. From this test dataset, a song is selected and serves as the seed for automatic playlist generation. Using the seed's attributes, similarity against other songs is calculated. This originates two playlists ordered by distance to the seed, one based on the predicted and another on the annotated AV values. The annotations playlist is then used to calculate the accuracy of the predicted list, by matching the top 1, 5 and 20 songs. Here, we only count how many songs in each top are the same (e.g., for top5, a match of 60% means that the same three songs are present in both lists). The entire process is repeated 500 times, averaging the results.

Results obtained for playlist generation were very similar between the three audio frameworks. Several tests were run using all the combinations of features referred before. The similarity ranking was calculated using predicted values from the regressor. From all the tests, a slightly higher accuracy was attained using the FFS selection of features from the combination of all frameworks, with 6.2%, 24.8% and 62.3% for top1, top5 and top20 respectively. Detailed results are presented in Table 5.8 (for details on the regression strategy and features see Section 5.1.1). The lower results in smaller playlists were mostly caused by the lack of accuracy when predicting valence. As previously stated, while FFS performed better, the number of features used is higher when compared to the RReliefF algorithm. As expected, best results were obtained with longer playlists, as normally used in a real scenario.

In summary, the playlist generation and similarity analysis for top1 was low, averaging 5% between all frameworks, with top20 presenting some reasonable results, of around 60%. Still, the results are very similar between feature selection algorithms to classify one as better suited. The same is observed in relation to frameworks, where the MIR Toolbox shows a slight advantage.

This experiment demonstrated that such APG approaches are valid. Furthermore, together with the experiments in Section 5.1, it gave us a better understanding of the limitations in MER, namely the identification of the most important problem to address which was the development of novel acoustic features able to capture the relevant musical attributes identified in the literature, especially features better correlated to valence.

As stated in previous studies (Y. E. Kim et al., 2010), the lyrical part of a song can have a great influence in the perceived emotion. The emotional response to the lyrics, obtained through natural language processing and commonsense reasoning, contributes to both the context and emotion classification of the song (Hu & Downie, 2010b; Meyers, 2007). As for playlist creation, it would be interesting to add some constraints regarding song ordering, for example, in terms of balance and progression.

[129] This 75-25 division was necessary so that the validation set was not too short, as we want to evaluate playlists containing up to 20 songs. On the other hand, the 90-10 division was employed in our previous experiment for the sake of comparison with Yang's results.

		Psy15	*Psy44*	*MIR*	*MAR*	*ALL*
Top1	All	4.2 ± 20.7	4.1 ± 18.6	3.6 ± 22.0	4.0 ± 20.7	4.2 ± 20.9
	FFS	5.6 ± 21.0	3.8 ± 18.6	5.2 ± 23.6	4.4 ± 19.8	**6.2 ± 20.7**
	RRF	5.1 ± 22.0	4.6 ± 19.0	5.6 ± 22.0	4.6 ± 22.6	5.2 ± 20.6
Top5	All	21.1 ± 18.1	20.9 ± 17.1	22.8 ± 19.0	18.1 ± 17.6	21.0 ± 17.8
	FFS	21.5 ± 18.3	21.2 ± 17.9	22.0 ± 19.3	19.8 ± 18.5	**24.8 ± 18.3**
	RRF	21.9 ± 18.1	22.1 ± 17.9	23.3 ± 18.4	18.7 ± 17.8	23.3 ± 18.4
Top20	All	61.9 ± 11.6	60.5 ± 12.3	62.7 ± 14.1	58.5 ± 13.6	60.7 ± 14.1
	FFS	62.0 ± 11.9	61.9 ± 12.4	62.5 ± 13.9	60.0 ± 13.6	**62.3 ± 13.6**
	RRF	61.0 ± 12.2	60.8 ± 12.8	61.7 ± 13.7	57.4 ± 13.0	61.6 ± 13.8

Table 5.8: Regression-based automatic playlist generation results (in %).

5.2.3. MOODetector Application

Finally, we have also built a working prototype called MOODetector to analyze music emotion as well as to generate playlists based on a song or an emotion trajectory. The MOODetector application was developed using the Qt Framework[130], a C++ application and UI development framework. It also uses Marsyas as part of the feature extraction logic, as well as the supervised learning algorithms offered by the libSVM library for emotion prediction.

Figure 5.6 illustrates the user interface, which has five main components: i) the main window (where all the other components are); ii) the Russell's plot (region highlighted with red); iii) the playlist view (blue region); iv) the multimedia controls (green region); iv) the instantaneous search bar (yellow region).

[130] https://www.qt.io/

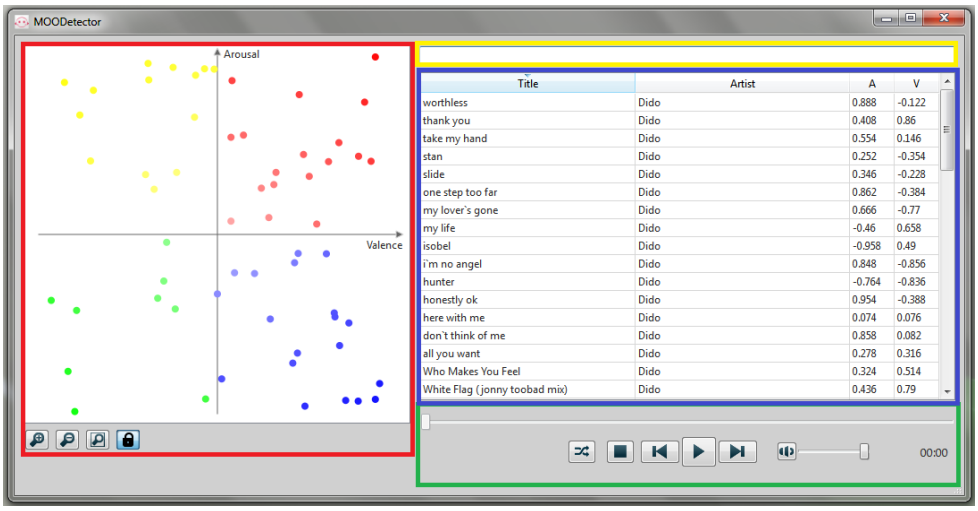


Figure 5.6: MOODetector interface. Red is the Russell's plane, green the multimedia controls, blue the playlist view and yellow the instantaneous search bar.

Regarding the Russell's plane, songs are positioned according to the estimated AV values using the processes described in the previous sections of this chapter. A color code is employed for song presentation in the plot: red, yellow, green and blue for the first, second, third and fourth quadrants, respectively. Additionally, the intensity of the color of each point (song) varies with the distance to the origin: lower if the song is close to the origin and getting higher as the distance increases.

As this application is meant to work as a typical media player, it has the basic usual music playback capabilities:

- Playback control of mp3 songs (pause, stop, forward, backward, shuffle);
- Volume control, mute, seek bar, and time label;
- Double click to play (in Russell's plot and playlist view).

It also has some more advanced music player capabilities like:

- Instantaneous music search (by music name and artist);
- Sort by song name, artist, arousal and valence values;
- Simple management of the library (adding and deleting songs);
- Music library statistics (library size, count by quadrant, analyzed songs, ...).

In the realm of mood detection, among other functions, this system can:

- Automatically estimate AV values for songs added to the library
- Allow the visualization of all or part of the music library in the Russell's plot;
- Navigate in the mentioned visualization using zoom and panning;

- Locate a song in the Russell's plot by clicking on the playlist view;
- See the numeric value of arousal and valence for a song;
- Automatically estimate the emotion variation of a song (an extension of emotion regression by predicting values of smaller song segments, e.g., 1-sec duration);
- Display a visualization of the mood tracking of a song;
- Allow the user to change the AV values of a song by drag and drop in the Russell's plot, besides manual edition of those values;
- Allow the user to reset one or all the changes of AV values.

In regard to playlists we can:

- Generate playlists based on one seed song (or point in the plane);
- Generate playlists based on a desired mood trajectory path drawn by the user, according to the distance to the seed(s) in the plane;
- Filter playlists via instantaneous search (can be combined with the previous);
- Export the playlist to a m3u file.

Playlist Generation

As mentioned, the system allows three types of playlist generation: i) based on a single seed song (or point in the plane); ii) based on a desired mood trajectory path drawn by the user, according to the distance to the seed(s) in the plane; and iii) via instantaneous search (and combined with the previous two).

Single seed mode

In order to generate a playlist based on a single seed song, the desired seed is selected in the Russell's plane. Then, the *N* closest songs, with a distance smaller than *threshold*, are added to the playlist (using the Euclidian distance). By default, the MOODetector system uses *N*=20 and *threshold*=0.35 (these parameters can be configured by the user). In Figure 5.7 we can see the seed song used in the playlist generation (black circumference around it) and the resulting playlist. As it can be observed, it contains only 9 songs, according to the defined threshold.

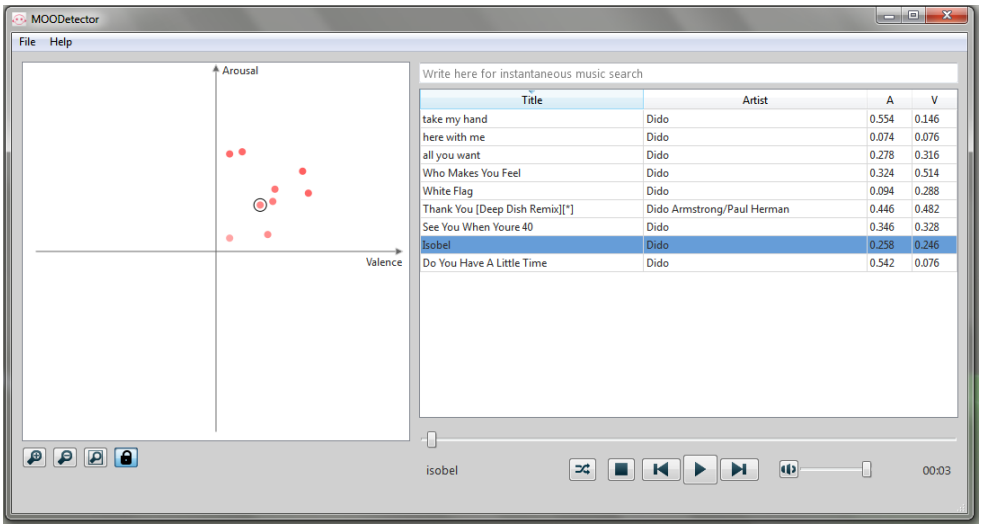


Figure 5.7: MOODetector automatic playlist generation using a seed song.

Path mode

In this mode, the user draws a path in the Russell's plot and the system computes the N closest songs to the path that have a distance less than *threshold*. This is done by evenly dividing the path into N points (i.e., reference points) and then calculating the closest song to each point, with the restriction that the corresponding distance should not exceed *threshold*. As before, by default, the system uses N=20 and *threshold*=0.35. This procedure is illustrated in Figure 5.8, where the black dots represent the reference points marking the path drawn by the user, while the color points are the resulting playlist.

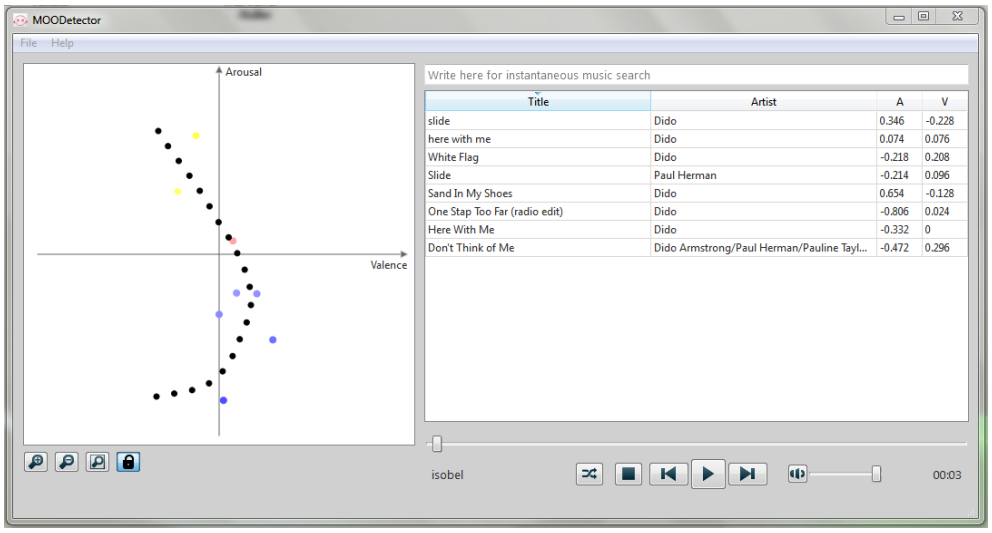


Figure 5.8: MOODetector automatic playlist generation using path mode (reference points in black).

Instantaneous search mode

The user can also write in the search bar, which will automatically create playlists on the fly, by a song filtering process. If a song contains all the written words in any position of its title or artist name fields, the song is kept in the playlist; otherwise, it is removed. This mode can also be combined with both the single seed and path modes.

Music Emotion Variation Visualization

For each song, it is possible to visualize its emotion variations throughout time in the MOODetector prototype. In Figure 5.9, the smaller emotion tracking plot illustrates quadrant changes in the Russell's plane across time. The same color code employed in the visualization of songs in the Russell's plane is kept. However, for simplicity, only quadrant information is encoded in the color code, rather than exact AV information.

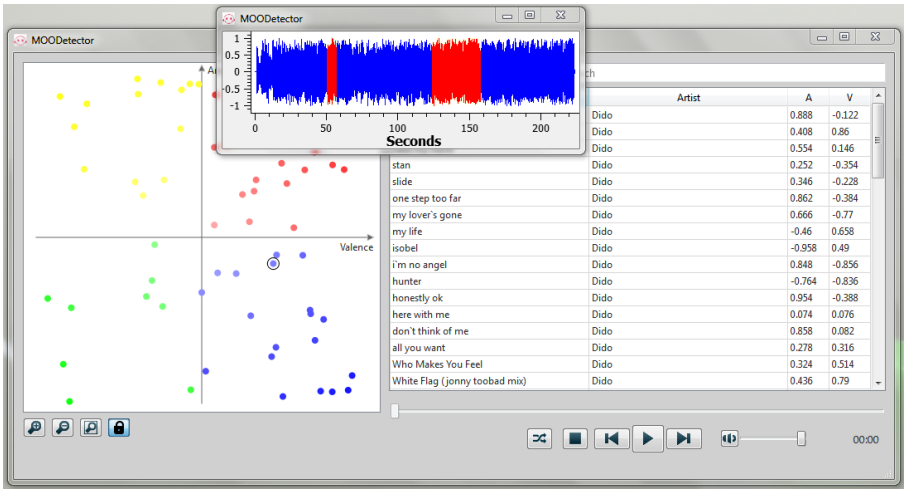


Figure 5.9: Visualization of the emotion variation of a song.

MOODetector Reloaded

The original prototype is currently being rewritten in Java to include a new user interface and a more accurate regression model, as shown in Figure 5.10. Both prototypes are (or will be, in case of the MOODetector Reloaded, still in development) at our website[131].

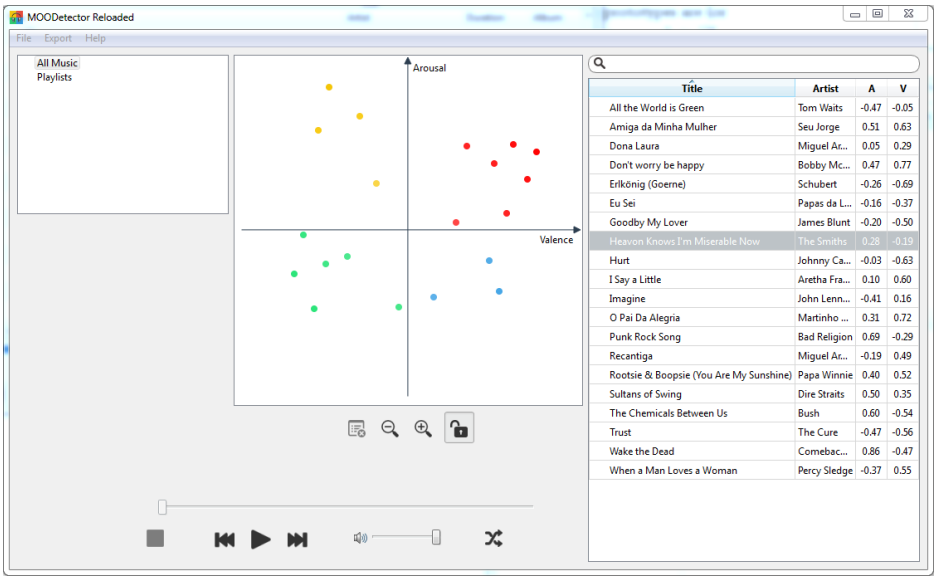


Figure 5.10: Main interface of the new MOODetector Reloaded prototype.

[131] http://mir.dei.uc.pt/downloads.html

5.3. MER Multi-modal Approaches

Several researchers have studied the influence of the lyrical content in music emotion recognition alone (He, Jin, Xiong, Chen, & Sun, 2008; Hu & Downie, 2010b; Malheiro et al., 2013) and in combination with audio signals (Mcvicar & Freeman, 2011; Y.-H. Yang, Lin, Cheng, et al., 2008).

As part of this work, we also investigated multi-modal approaches, by combining different information sources such as audio and lyrics. This section details the process, from multi-modal dataset construction, feature extraction and classification and experimental results.

5.3.1. Dataset Construction

This dataset has been presented in Section 3.2.2 (referred to as the CISUC Bi-modal dataset). To construct it, we started by collecting 200 song lyrics and the corresponding audio (30-sec audio clips). The criteria for selecting the songs were the following: several musical genres and eras; songs distributed uniformly by the 4 quadrants of the Russell emotion model.

The annotation of the dataset was performed by 39 people with different backgrounds. Each annotator classified, for the same song, either the audio (listening) or the lyric (by reading the text). During the process, subjects were instructed to: select the basic predominant emotion expressed by the audio / lyric (if more than one emotion was identified, he/she was told to pick the predominant one); assign values (between -4 and 4 with a granularity of one unit) to valence and arousal.

For both, audio and lyrics dataset, the arousal and valence of each song were obtained by the average of the annotations of all the subjects. We obtained an average of 6 and 8 annotations respectively for audio and lyrics. To improve the consistency of the ground truth, the standard deviation (SD) of the annotations made by different subjects for the same song was evaluated. Using the same methodology as in (Y.-H. Yang, Lin, Su, et al., 2008), songs with an SD above 1.2 were excluded from the original set. As a result, the final audio dataset contains 162 audio clips (quadrant 1 (Q1) – 52 songs; quadrant 2 (Q2) – 45; quadrant 3 (Q3) – 31 and quadrant 4 (Q4) – 34), while the final lyrics dataset contains 180 lyrics (Q1 – 44 songs; Q2 – 41; Q3 – 51 and Q4 – 44). Finally, the consistency of the ground truth was evaluated using Krippendorff's alpha (Krippendorff, 2003, Chapter 11), a measure of inter-coder agreement. This measure achieved, in the range -4 up to 4, 0.69 and 0.72 respectively for valence and arousal. This is considered a substantial agreement among the annotators. As for the lyrics the measure achieved 0.87 and 0.82 respectively for valence and arousal. This is considered a strong agreement among the annotators.

The size of the datasets is not too large, however we think that is acceptable for experiments and is similar to other datasets manually annotated (e.g., (Y.-H. Yang, Lin, Su, et al., 2008) has 195 songs).

Based on the lyrics and audio datasets, we created a bimodal dataset. We considered that a song (audio + lyrics) is a valid song to integrate this bimodal dataset, if the song belongs simultaneously to the audio and lyrics datasets and in both datasets the annotation belongs to the same quadrant, i.e., we can only consider songs in which the classification (quadrant) for the audio sample is equal to the classification for the lyric sample. So we started from a lyrics dataset containing 180 samples and an audio dataset containing 162 clips, obtaining a bimodal dataset containing 133 songs (with audio and lyrics): 37 songs for Q1 and Q2, 30 for Q3 and 29 for Q4.

5.3.2. Feature Extraction

In musical theory, the basic musical elements and characteristics are commonly grouped under broader distinct dimensions such as rhythm, melody, tone color and others as described in Section 2.3. In this experiment, we extracted the same 1701 audio features used in Chapter 4 as standard (baseline).

As for lyric features, we used state-of-the-art features such as: bag-of-words (BOW) – unigrams, bigrams and trigrams – associated or not to a set of transformations, e.g., stemming and stop-words removal; part-of-speech (POS) tagging[132] followed by a BOW analysis; 36 features representing the number of occurrences of 36 different grammatical classes in the lyrics (e.g., number of adjectives). We also used all the features based on existing frameworks like Synesketch (8 features), ConceptNet (8 features), LIWC (82 features) and General Inquirer (182 features). In addition to the previous frameworks, we use features based on known dictionaries such as DAL (Dictionary of Affect in Language) (Whissell, Fournier, Pelland, Weir, & Makarec, 1986) and ANEW (Affective Norms for English Words) (Bradley & Lang, 1999). Finally, we used features proposed by our team (Malheiro et al., 2018): Slang presence, which counts the number of slang words from a dictionary of 17700 words; Structural analysis features, e.g., the number of repetitions of the title and chorus, the relative position of verses and chorus in the lyric; Semantic features, e.g., dictionaries personalized to the employed emotion categories.

[132] They consist in attributing a corresponding grammatical class to each word.

5.3.3. Experimental Results

We conducted one experiment which is classification by quadrants (4 categories – Q1, Q2, Q3 and Q4). We used Support Vector Machines (SVM), since, based on previous evaluations, this technique performed generally better than other methods. The classification results were validated with repeated stratified 10-fold cross validation (with 20 repetitions) and the average obtained performance (F1-Score) is reported.

We constructed first, both for audio and lyrics, the best possible classifiers. We applied, for each of the dimensions, feature selection and ranking using the ReliefF algorithm. Next, we combined the best features of audio and lyrics and constructed, using the same prior terms, the best bimodal classifier.

We can see in Table 5.9 the performance of the best model for lyrics, audio and for the combination of the best lyric and audio features. The fields "# of Features", "Selected Features" and "F1-Score" represent respectively the total number of features, the number of selected features and the F1-Score score attained after feature selection. In the last line, the total number of bimodal features (1065) is the sum of the selected lyrics and audio features (647 and 418), while the while 1057 is the number of the original 1065 feature set, selected using ReliefF.

Classification by Quadrants	*# of Features*	*Selected Features*	*F1-Score*
Audio	1232	647	79.3%
Lyrics	1701	418	72.6%
Bimodal	1065	1057	88.4%

Table 5.9: Summary of the best classification results by quadrants.

As can be seen, the best lyrics-based model achieved better performance than the best audio-based model (79.3% vs 72.6%). This is not the more frequent pattern in the state-of-the-art, where usually the best results are achieved with the audio. This happens for example in (Laurier et al., 2008). To the best of our knowledge, the work by Hu et al. (2009) is one of the only studies where lyrics performance supplants audio performance, but only for some of the emotions. This suggests that our new lyrical features have an important role for these results.

As we can see, both dimensions are important, since bimodal analysis improves significantly (at $p<0.05$ Wilcoxon Test) the results of the lyrics classifier (from 79.3% to 88.4%). Furthermore, the best bimodal classifier, after feature selection, contains almost all the features from the best classifiers of lyrics and audio (1057 features in 1065 possible features). This suggests the importance of the features from both dimensions.

The main conclusion of this experiment is that, unlike most of the similar works in the state-of-the-art, lyrics performed better than audio. This suggests the importance of multi-modal approaches and also exploring new lyric features. Another conclusion is that bimodal analysis is always better than each dimension separately.

Chapter 6

Conclusions and Perspectives

After reaching the end of this work, the most indisputable finding attained is that our contribution is only one more step in the long journey of music emotion recognition. Despite the inherent difficulties of a research topic aiming to bridge subjective fields such as human emotions and music with computer science, we confirmed yet again that the task is feasible and improvements are possible, as demonstrated by our modest contribution. Moreover, despite distant, several possible paths to improve MER have been uncovered, from novel features (e.g., related with musical form), to the exploration of lyrics and voice acoustics in addition to audio as sources.

In this chapter, we present a general overview of the work carried out in this thesis, as well as a summary of the contributions derived from it. Building on these, and considering the problems still open in the field, we draw some of the possible lines of research to advance the music emotion recognition field.

6.1. Summary and Conclusions

Set at the beginning of this dissertation, our general goal was to tackle some of the existing limitations in emotion recognition from music audio signals. The majority of the research in the field has been dedicated to the creation of better datasets and proposing different approaches based on novel machine learning techniques. We chose to tackle the problem from a different vector, by looking into the audio signal, the part that has been somewhat neglected. Thus, our main contributions consist of a knowledge base, which relates musical dimensions with known emotional responses and a set of novel emotionally-relevant audio features, shown to improve emotion classification results; besides, a novel dataset and respective construction methodology were proposed so as to attain the mentioned main goal.

This work began with an assessment on the concept of emotion and how psychologists have defined and classified human emotions over the years. From this, we verified that, despite the inherent subjectivity and lack of consensus, emotion taxonomies can

be divided into two approaches: categorical and dimensional. In the former, emotions are represented by classes (e.g., words), leading to problems such as the choice of classes to use and how different persons may use different but close words to describe similar emotions (or the opposite). Dimensional models were later proposed to solve these issues, by using two or three dimensions (e.g., arousal and valence) to define an emotional space and defining each point there as a different emotional state. Despite the reduced ambiguity, this model introduces additional complexity, since humans use words rather than points in space to describe emotions.

Next, a broad review of the music emotion recognition field was conducted. This gave us a better understanding of the typical approaches, their existing limitations and some possible ideas to be considered. Also, while dimensional MER approaches have gained ground, mostly due to its practicality, the observed results are weak, as are the ones obtained in "simpler" problems with categorical models. Moreover, the features used are mostly from previous information retrieval problems (e.g., Mel-frequency cepstral coefficients used for speech recognition) and the datasets suffer from several problems, from being private, to having low-quality annotations and being small in size. Finally, a great amount of work has focused on different machine learning techniques, with support vector machines or derivatives being generally the best performing algorithms. Two main issues were identified with the review: very few studies have been dedicated specifically to the audio features used; and simpler classification problems with low granularity have not been solved, thus approaching MER as a regression problem may be considered as a rather hasty decision.

To tackle the first abovementioned issue, we reviewed relations between music, emotional responses and computational audio features. This organized knowledge is one of our major contributions. To this end, we first reviewed musical characteristics, consolidating them into one of eight musical dimensions: rhythm, harmony, melody, dynamics, tone color, expressive techniques, musical texture and musical form. Next, we reviewed musicological studies linking these to specific emotional responses. Finally, the audio features available in widely used audio frameworks were analyzed and catalogued into one of the eight dimensions. As a result, we verified that very few audio features are available to musical texture, expressive techniques and musical form dimensions. Still, these dimensions have been related to specific emotions.

Founded on this knowledge, we proposed a set of novel audio features related with melody, dynamics, rhythm, musical texture and expressive techniques. These are built on previous transcription work, going from the audio signal to MIDI notes, capturing information related to articulation, vibrato, glissando, texture types, among others. To assess their importance, a novel dataset of 900 30-sec audio clips was created with a semi-automatic process proposed by us. The dataset is annotated in terms of Russell's quadrants, containing a myriad of metadata (e.g., genres, artist and title, dates, and so on) gathered in a well-documented process.

The classification results demonstrated that our features were emotionally-relevant, being consistently selected among the best features for each of the tested problems – quadrants, arousal and valence classification. Moreover, we were able to extract valuable insights about which features are related with specific emotional states. For illustration, characteristics such as texture type, the signal complexity and characteristics of the voice, extracted without the accompaniment seem very relevant to solve the current major problem: distinguish calm from sad songs (low arousal).

Besides this main research, we also performed other experiments involving the evaluation of different MER strategies (e.g., the dimensional approach) and datasets, the creation of emotion-based music playlists based on emotion regression models and multi-modal approaches, combining audio and lyrics sources.

To sum up, while our proposed features and uncovered insights are interesting and may lead to new ideas to future research, having a good performing solution that is used in real life scenarios is still years ahead.

6.2. Perspectives for Future Research

Regarding the future work, our general goal is to propose additional solutions to improve the current state of the audio music emotion recognition field. To this end, we will use the knowledge acquired during the last years, proposing novel features, namely exploring musical form, as well as other solutions.

Even though the development of novel features is a laborious task, we see it as one of the key requirements in order to improve MER and still understand the existent relations (something lost in a deep learning approach). In that sense, this work already identified musical form as one of the remaining musical dimensions to be captured by computational algorithms. Exploring musical form of a song requires temporal analysis of the data, searching for patterns and similarity between groups of notes, looking for what may be a chorus, intro and outro. Therefore, as this analysis is typically performed in complete songs, it was not exploited at this stage of our research. To capture musical form, some possible solutions could use similarity matrices or autocorrelation between notes close together. One additional strategy that might be worth exploring is the usage of clustering techniques with note information, as well as additional info possibly related to form changes (e.g., rhythm and dynamics). Here, the resulting clusters of notes could indicate similar structures (e.g., a chorus), while the time information of these notes could help split a cluster in different occurrences of such event over time.

Additionally, this work confirmed the importance of other information sources such as the singing voice to the problem and demonstrated that such information may be especially relevant to distinguish sad from calm music. This has been pointed out in

a previous study, in which the authors noted that the voice signal relevance was lost when mixed with the accompaniment (Xu et al., 2014). Moreover, the lyrical content has also been pointed as relevant to discern between negative and positive valence emotions (Malheiro et al., 2018). Given this, we would like to continue exploring our idea of emotion prediction using source separated voice signals and additionally study the possibility of transcribing lyrics directly from the source-separated voice signal. Although being two open problems, some interesting advances have been made recently in both fields with the massification of deep learning strategies. First, several new papers have approached source separation recently. Secondly, a very recent work (Gupta et al., 2018) has built on automatic speech recognition services such as Google Speech-to-Text engine and proposed methods to transcribe and align lyrics to solo-singing vocals.

Furthermore, we want to study the importance of our proposed system in other approaches such as dimensional emotion recognition, music emotion variation detection (where musical form might be particularly relevant) and the construction of linguistically interpretable rule-based models (to make the relations among features more explicit). Also, the created knowledge will be added to our emotion detection prototype.

We cannot conclude without a reference to deep learning algorithms. Over the last years several variations of deep neural networks have been revolutionizing the information and communications technology world, thanks to the advances in computing power, especially GPUs and the amount of data captured nowadays. Such solutions are also improving the MIR field, as can be seen by the influence of the subject in this year's ISMIR (2018) conference publications. Thus, we are also interested in exploring such solutions to emotion recognition. Still, such approaches raise several points that must be considered. First, deep learning solutions require massive amounts of data and such datasets are usually limited to big companies (such as Spotify, Pandora or Last.FM in this field). Secondly, their usage is limited to the quality of data, in other words, a model prediction can be only as good as the data being used. Unfortunately, large MER datasets have been known to be problematic due to the associated subjectivity and complexity of data collection. Finally, deep learning models are opaque in the sense that the extracted features are often difficult to interpret, which, from a music psychology point of view, is limiting.

BIBLIOGRAPHY

Abeles, H. F., & Chung, J.-W. (1996). Physiological responses to music and sound stimuli. In D. A. Hodges (Ed.), *Handbook of Music Psychology* (Second edi, pp. 285–342). MMB Music.

Adams, C. R. (1976). Melodic Contour Typology. *Ethnomusicology*, *20*(2), 179–215. http://doi.org/10.2307/851015

Aggarwal, C. C., & Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications* (1st ed.). CRC Press.

Ahsan, H., Kumar, V., & Jawahar, C. V. (2015). Multi-label annotation of music. In *8th International Conference on Advances in Pattern Recognition – ICAPR* (pp. 1–5). Kolkata, India: IEEE. http://doi.org/10.1109/ICAPR.2015.7050685

Akkermans, V., Serrà, J., & Herrera, P. (2009). Shape-based spectral contrast descriptor. In *6th Sound and Music Computing Conference – SMC 2009* (pp. 143–148). Porto, Portugal.

Aljanaki, A. (2016). *Emotion in Music: representation and computational modeling*. Utrecht University.

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management*, *52*(1), 115–128. http://doi.org/10.1016/j.ipm.2015.03.004

Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PLoS ONE*, *12*(3). http://doi.org/10.1371/journal.pone.0173392

Allamanche, E., Hellmuth, O., Fröba, B., Kastner, T., & Cremer, M. (2001). Content-based Identification of Audio Material Using MPEG-7 Low Level Description. In *2nd International Conference on Music Information Retrieval – ISMIR 2001* (Vol. 8, pp. 197–204). Bloomington, Indiana, USA.

Amado-Boccara, I., Donnet, D., & Olié, J. P. (1972). The concept of mood in psychology. *LEncephale*, *19*(2), 117–122.

Aristotle (IV c B.C.). (1944). Politics. In *Aristotle in 23 Volumes, Vol. 21 (Rackham, H. Trans.)*. Cambridge, MA: Harvard University Press. Retrieved from http://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0086.tlg035

Aucouturier, J.-J., & Pachet, F. (2002). Scaling up music playlist generation. In *IEEE*

International Conference on Multimedia and Expo – ICME 2002 (Vol. 1, pp. 105–108). Lausanne, Switzerland: IEEE. http://doi.org/10.1109/ICME.2002.1035729

Augustine (426 A.D.). (1871). *Works of Aurelius Augustine: The City of God*. (M. Dods, Trans.) (Vol. 2). Edinburgh, UK: T. & T. Clark.

Ayadi, M. El, Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. http://doi.org/10.1016/j.patcog.2010.09.020

Bain, A. (1859). *The Emotions and the Will*. London, UK: John W. Parker and son.

Balkwill, L.-L., & Thompson, W. F. (1999). A Cross-Cultural Investigation of the Perception of Emotion in Music: Psychophysical and Cultural Cues. *Music Perception: An Interdisciplinary Journal*, *17*(1), 43–64. http://doi.org/10.2307/40285811

Barrett, L. F., & Russell, J. A. (1999). The Structure of Current Affect. *Current Directions in Psychological Science*, *8*(1), 10–14. http://doi.org/10.1111/1467-8721.00003

Bartoszewski, M., Kwasnicka, H., Markowska-Kaczmar, U., & Myszkowski, P. B. (2008). Extraction of Emotional Content from Music Data. In *7th Computer Information Systems and Industrial Management Applications - CISIM2008* (pp. 293–299). IEEE. http://doi.org/10.1109/CISIM.2008.46

Bell, C. (1824). *Essays on the Anatomy and Philosophy of Expression*. London, UK: John Murray. Retrieved from https://archive.org/details/essaysonanatomy00bellgoog

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, *41*(3), 407–434. http://doi.org/10.1007/s10844-013-0258-3

Benward, B., & Saker, M. N. (2008). *Music: In Theory and Practice, Vol. I* (8th ed.). McGraw-Hill.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data* (pp. 25–71). Berlin/Heidelberg: Springer-Verlag. http://doi.org/10.1007/3-540-28349-8_2

Berridge, K. C. (2003). Comparing the emotional brains of humans and other animals. *Handbook of Affective Sciences (Series in Affective Science)*, 25–51.

Berry, W. (1976). *Structural Functions in Music*. Prentice-Hall.

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In *12th International Society for Music Information Retrieval Conference – ISMIR 2011* (pp. 591–596). Miami, Florida, USA.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, *19*(8), 1113–

1139. http://doi.org/10.1080/02699930500204250

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., ... Serra, X. (2013). ESSENTIA: An audio analysis library for music information retrieval. In *14th International Society for Music Information Retrieval Conference – ISMIR 2013*. Curitiba, Brazil.

Bogert, B. P., Healy, M. J. R., & Tukey, J. W. (1963). The Quefrency Alanysis [sic] of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed)* (pp. 209–243). Wiley.

Boisen, S., Crystal, M., Schwartz, R. M., Stone, R., & Weischedel, R. M. (2000). Annotating Resources for Information Extraction. In *2nd International Conference on Language Resources and Evaluation - LREC2000*. Athens, Greece: European Language Resources Association.

Bonde, L. O. (2007). Music as Metaphor and Analogy. *Nordic Journal of Music Therapy*, *16*(1), 73–81. http://doi.org/10.1080/08098130709478173

Bradley, M. M., & Lang, P. J. (1999). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. *Psychology*, *Technical*(C-1), 0. http://doi.org/10.1109/MIC.2008.114

Brandt, A. (2011). Musical Form. In *Sound Reasoning* (21.2). Anthony Brandt, Brian Nelson, Robert McClure. Retrieved from https://cnx.org/contents/R21GFBYj@21.2:BbcIIi_q@13/Musical-Form

Brossier, P., Bello, J. P., & Plumbley, M. D. (2004). Real-time temporal segmentation of note objects in music signals. In *International Computer Music Conference – ICMC 2004*. Miami, Florida, USA.

Brown, T. (2010). *Thomas Brown: Selected philosophical writings*. (T. Dixon, Ed.). Exeter, UK: Imprint Academic.

Broze, Y., Paul, B. T., Allen, E. T., & Guarna, K. M. (2014). Polyphonic Voice Multiplicity, Numerosity, and Musical Emotion Perception. *Music Perception: An Interdisciplinary Journal*, *32*(2), 143–159. http://doi.org/10.1525/mp.2014.32.2.143

Burmania, A., Parthasarathy, S., & Busso, C. (2016). Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing*, *7*(4), 374–388. http://doi.org/10.1109/TAFFC.2015.2493525

Cabrera, D. (1999). The Size of Sound: Auditory Volume Reassessed. In *1999 Australasian Computer Music Association Conference* (pp. 26–31).

Cabrera, D., Ferguson, S., & Schubert, E. (2007). "Psysound3": Software for Acoustical and Psychoacoustical Analysis of Sound Recordings. In G. P. Scavone (Ed.), *13th*

International Conference on Auditory Display – ICAD2007 (pp. 356–363). Montreal, Canada: Schulich School of Music, McGill University.

Cacioppo, J. T., Berntson, G., Larsen, J., Poehlmann, K. M., & Ito, T. A. (2000). The Psychophysiology of Emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (Second edi, pp. 173–191). The Guilford Press.

Camacho, A. (2007). *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator*. University of Florida.

Campbell, I. G. (1942). Basal Emotional Patterns Expressible in Music. *The American Journal of Psychology*, *55*(1), 1. http://doi.org/10.2307/1417020

Casey, M. A. (2002). Sound Classification and Similarity Tools. In *Introduction to MPEG-7: Multimedia Content Description Language* (pp. 309–323). http://doi.org/10.1017/S1355771801002126

Cawley, G. C., Talbot, N. L., & Girolami, M. (2007). Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (pp. 209–216). MIT Press.

Celma, Ò., Herrera, P., & Serra, X. (2006). Bridging the Music Semantic Gap. In *Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, held in conjunction with the European Semantic Web Conference* (Vol. 187, pp. 177–190). Budva, Montenegro.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology – TIST*, *2*(3), 1–27. http://doi.org/10.1145/1961189.1961199

Chen, S.-H., Lee, Y.-S., Hsieh, W.-C., & Wang, J.-C. (2015). Music Emotion Recognition Using Deep Gaussian Process. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference – APSIPA 2015* (pp. 495–498). Hong Kong, China.

Chen, Y.-A., Wang, J.-C., Yang, Y.-H., & Chen, H. H. (2014). Linear regression-based adaptation of music emotion recognition models for personalization. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2149–2153). Florence, Italy: IEEE. http://doi.org/10.1109/ICASSP.2014.6853979

Chen, Y.-A., Wang, J.-C., Yang, Y.-H., & Chen, H. H. (2017). Component Tying for Mixture Model Adaptation in Personalization of Music Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing – TASLP*, *25*(7), 1409–1420. http://doi.org/10.1109/TASLP.2017.2693565

Chen, Y.-A., Yang, Y.-H., Wang, J.-C., & Chen, H. H. (2015). The AMG1608 dataset for music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 693–697). IEEE.

http://doi.org/10.1109/ICASSP.2015.7178058

Collier, G. L. (2007). Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, *35*(1), 110–131. http://doi.org/10.1177/0305735607068890

Collier, W. G., & Hubbard, T. L. (2001). Judgments of happiness, brightness, speed and tempo change of auditory stimuli varying in pitch and tempo. *Psychomusicology: A Journal of Research in Music Cognition*, *17*(1–2), 36–55. http://doi.org/10.1037/h0094060

Cook, N. D., & Fujisawa, T. X. (2006). The Psychophysics of Harmony Perception: Harmony is a Three-Tone Phenomenon. *Empirical Musicology Review*, *1*(2), 106–126. http://doi.org/10.18061/1811/24080

Cooke, D. (1959). *The language of music*. Oxford, UK: Oxford University Press. Retrieved from https://philpapers.org/rec/COOTLO-5

Corona, H., & O'Mahony, M. P. (2015). An Exploration of Mood Classification in the Million Songs Dataset. In *12th Sound and Music Computing Conference*. Maynooth, Ireland.

Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. http://doi.org/10.1007/BF00994018

Costa, M., Bitti, P. E. R., & Bonfiglioli, L. (2000). Psychological Connotations of Harmonic Musical Intervals. *Psychology of Music*, *28*(1), 4–22. http://doi.org/10.1177/0305735600281002

Costa, M., Fine, P., & Bitti, P. E. R. (2004). Interval Distributions, Mode, and Tonal Strength of Melodies as Predictors of Perceived Emotion. *Music Perception: An Interdisciplinary Journal*, *22*(1), 1–14. http://doi.org/10.1525/mp.2004.22.1.1

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, *11*(4), 921–937. http://doi.org/10.1037/a0024700

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. http://doi.org/10.1007/BF02310555

Cullen, A., Kane, J., Drugman, T., & Harte, N. (2013). Creaky Voice and the Classification of Affect. In *Workshop in Affective and Social Speech Signals – WASSS 2013*. Grenoble, France.

Dace, W. (1963). The Concept of "Rasa" in Sanskrit Dramatic Theory. *Educational Theatre Journal*, *15*(3), 249. http://doi.org/10.2307/3204783

Daniel, P., & Weber, R. (1997). Psychoacoustical Roughness: Implementation of an Optimized Model. *Acta Acustica United with Acustica*, *83*(1), 113–123.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London, UK: John Murray. http://doi.org/10.1037/h0076058

Davidson, M., & Heartwood, K. (1997). *Songwriting for Beginners: An Easy Beginning Method*. Alfred Music.

Davies, S. (2003). *Themes in the philosophy of music*. Oxford University Press.

Davis, D. (2002). Characteristics of Sound. *Lecture Notes on the Principles of Physics I Course, Chapter 16, Eastern Illinois University, Physics Department*. Retrieved from https://web.archive.org/web/20030920083934/oldsci.eiu.edu/physics/DDavis/1150/16Waves/char.html

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing – TASSP*, *28*(4), 357–366. http://doi.org/10.1109/TASSP.1980.1163420

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917. http://doi.org/10.1121/1.1458024

Degara, N., Rua, E. A., Pena, A., Torres-Guijarro, S., Davies, M. E. P., & Plumbley, M. D. (2012). Reliability-Informed Beat Tracking of Musical Signals. *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, *20*(1), 290–301. http://doi.org/10.1109/TASL.2011.2160854

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M. (2018). Music Mood Detection Based on Audio and Lyrics with Deep Neural Net. In E. Gómez, X. Hu, E. Humphrey, & E. Benetos (Eds.), *19th International Society for Music Information Retrieval Conference – ISMIR 2018* (pp. 370–375). Paris, France.

Deng, L., & O'Shaughnessy, D. (2003). *Speech Processing: A Dynamic and Optimization-Oriented Approach*. Marcel Dekker.

Dilger, D. E. (2011, April 21). iTunes Store quietly generates record revenues of $1.4 billion. *AppleInsider*. Retrieved from https://appleinsider.com/articles/11/04/21/itunes_store_quietly_generates_record_revenues_of_1_4_billion

Dixon, T. (2003). *From Passions to Emotions: The Creation of a Secular Psychological Category*. Cambridge, UK: Cambridge University Press.

Dixon, T. (2012). "Emotion": The History of a Keyword in Crisis. *Emotion Review*, *4*(4), 338–344. http://doi.org/10.1177/1754073912445814

Doets, P. J. O., Gisbert, M. M., & Lagendijk, R. L. (2006). On the comparison of audio fingerprints for extracting quality parameters of compressed audio. In E. J. Delp III & P. W. Wong (Eds.), *SPIE Electronic Imaging Conference 2006*.

http://doi.org/10.1117/12.642968

Dressler, K. (2016). *Automatic Transcription of the Melody from Polyphonic Music*. Ilmenau University of Technology.

Dromey, C., Holmes, S. O., Hopkin, J. A., & Tanner, K. (2015). The Effects of Emotional Expression on Vibrato. *Journal of Voice*, *29*(2), 170–181. http://doi.org/10.1016/j.jvoice.2014.06.007

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Wiley.

Duggan, W. (2018, February 28). Apple Music Is Now Worth $10 Billion. *U.S. News & World Report*. Retrieved from https://money.usnews.com/investing/stock-market-news/articles/2018-02-28/apple-inc-aapl-stock

Eerola, T., Ferrer, R., & Alluri, V. (2012). Timbre and Affect Dimensions: Evidence from Affect and Similarity Ratings and Acoustic Correlates of Isolated Instrument Sounds. *Music Perception: An Interdisciplinary Journal*, *30*(1), 49–70. http://doi.org/10.1525/mp.2012.30.1.49

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*, 487. http://doi.org/10.3389/fpsyg.2013.00487

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *10th International Society for Music Information Retrieval Conference – ISMIR 2009* (pp. 621–626). Kobe, Japan.

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49. http://doi.org/10.1177/0305735610362821

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. (J. Cole, Ed.)*Nebraska Symposium On Motivation*. University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3), 169–200. http://doi.org/10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*. Cambridge, MA: Malor Books.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., … Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717. http://doi.org/10.1037/0022-3514.53.4.712

Ekman, P., Levenson, R., & Friesen, W. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*(4616), 1208–1210. http://doi.org/10.1126/science.6612338

Ellsworth, P. C. (1994). William James and emotion: Is a century of fame worth a century of misunderstanding? *Psychological Review*, *101*(2), 222–229. http://doi.org/10.1037/0033-295X.101.2.222

Epstein, D. (1995). *Shaping time : music, the brain, and performance*. Wadsworth Publishing.

Erickson, R. (1975). *Sound Structure in Music* (1st ed.). Berkeley, California, USA: University of California Press.

Esteller, R., Vachtsevanos, G., Echauz, J., & Litt, B. (2001). A comparison of waveform fractal dimension algorithms. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, *48*(2), 177–183. http://doi.org/10.1109/81.904882

Everett, W. (1999). *The Beatles as Musicians: Revolver through the Anthology*. Oxford University Press.

Eyben, F., Salomão, G. L., Sundberg, J., Scherer, K. R., & Schuller, B. W. (2015). Emotion in the singing voice—a deeperlook at acoustic features in the light ofautomatic classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 19. http://doi.org/10.1186/s13636-015-0057-6

Fan, J., Tatar, K., Thorogood, M., & Pasquier, P. (2017). Ranking-Based Emotion Recognition for Experimental Music. In *18th International Society for Music Information Retrieval Conference - ISMIR 2017*. Suzhou, China.

Fan, Z.-C., Jang, J.-S. R., & Lu, C.-L. (2016). Singing Voice Separation and Pitch Extraction from Monaural Polyphonic Audio Music via DNN and Adaptive Pitch Tracking. In *2016 IEEE Second International Conference on Multimedia Big Data – BigMM 2016* (pp. 178–185). Taipei, Taiwan: IEEE. http://doi.org/10.1109/BigMM.2016.56

Farnsworth, P. R. (1954). A Study of the Hevner Adjective List. *The Journal of Aesthetics and Art Criticism*, *13*(1), 97. http://doi.org/10.2307/427021

Farnsworth, P. R. (1958). *The Social Psychology of Music*. New York, NY, USA: The Dryden Press. Retrieved from https://archive.org/details/socialpsychology00farn

Fastl, H. (1982). Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. *Hearing Research*, *8*(1), 59–69. http://doi.org/10.1016/0378-5955(82)90034-X

Feinstein, H. M. (1970). William James on the Emotions. *Journal of the History of Ideas*, *31*(1), 133. http://doi.org/10.2307/2708376

Feng, Y., Zhuang, Y., & Pan, Y. (2003). Popular Music Retrieval by Detecting Mood. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR 2003* (Vol. 2, pp. 375–376). Toronto, Canada: ACM

Press. http://doi.org/10.1145/860435.860508

Fernández-Sotos, A., Fernández-Caballero, A., & Latorre, J. M. (2016). Influence of Tempo and Rhythmic Unit in Musical Emotion Regulation. *Frontiers in Computational Neuroscience*, *10*, 80. http://doi.org/10.3389/fncom.2016.00080

Flexer, A., Schnitzer, D., Gasser, M., & Widmer, G. (2008). Playlist Generation Using Start and End Songs. In *9th International Society of Music Information Retrieval Conference – ISMIR 2008* (pp. 173–178). Philadelphia, Pennsylvania, USA.

Foote, J. T. (2000). Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo – ICME2000* (Vol. 1, pp. 452–455). New York, NY, USA: IEEE. http://doi.org/10.1109/ICME.2000.869637

Foote, J. T., & Cooper, M. L. (2003). Media segmentation using self-similarity decomposition. In M. M. Yeung, R. W. Lienhart, & C.-S. Li (Eds.), *SPIE Electronic Imaging Conference 2003* (p. 167). Santa Clara, CA, United States. http://doi.org/10.1117/12.476302

Foote, J. T., Cooper, M. L., & Nam, U. (2002). Audio retrieval by rhythmic similarity. In *3rd International Conference on Music Information Retrieval – ISMIR 2002* (Vol. 3, pp. 265–266). Paris, France.

Forte, A. (1979). *Tonal Harmony in Concept and Practice* (3rd editio). Holt, Rinehart and Winston.

Friberg, A. (2008). Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music. In *11th International Conference on Digital Audio Effects – DAFx 2008* (pp. 1–6). Espoo, Finland.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., … Koelsch, S. (2009). Universal Recognition of Three Basic Emotions in Music. *Current Biology*, *19*(7), 573–6. http://doi.org/10.1016/j.cub.2009.02.058

Fritz, T., & Koelsch, S. (2008). The role of semantic association and emotional contagion for the induction of emotion with music. *Behavioral and Brain Sciences*, *31*(05), 579–580. http://doi.org/10.1017/S0140525X08005347

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia*, *13*(2), 303–319. http://doi.org/10.1109/TMM.2010.2098858

Fukunaga, K., & Koontz, W. L. G. (1970). Application of the Karhunen-Loève Expansion to Feature Selection and Ordering. *IEEE Transactions on Computers*, *C-19*(4), 311–318. http://doi.org/10.1109/T-C.1970.222918

Futrelle, J., & Downie, J. S. (2003). Interdisciplinary Research Issues in Music Information Retrieval: ISMIR 2000?2002. *Journal of New Music Research*, *32*(2), 121–

131. http://doi.org/10.1076/jnmr.32.2.121.16740

Gabrielsson, A. (2001a). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, *5*, 123–147. http://doi.org/10.1177/10298649020050S105

Gabrielsson, A. (2001b). Emotions in strong experiences with music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion: Theory and Research (Series in Affective Science)* (pp. 431–449). Oxford University Press.

Gabrielsson, A., & Juslin, P. N. (1996). Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music*, *24*(1), 68–91. http://doi.org/10.1177/0305735696241007

Gabrielsson, A., & Juslin, P. N. (2003). Emotional expression in music. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences (Series in Affective Science)* (pp. 503–534). Oxford, UK: Oxford University Press.

Gabrielsson, A., & Lindström, E. (2001). The Influence of Musical Structure on Emotional Expression. In *Music and Emotion* (Vol. 8, pp. 223–248). Oxford University Press.

Gabrielsson, A., & Lindström, E. (2011). The Role of Structure in the Musical Expression of Emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 367–400). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199230143.003.0014

Gagnon, L., & Peretz, I. (2003). Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies. *Cognition & Emotion*, *17*(1), 25–40. http://doi.org/10.1080/02699930302279

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In *9th International Conference on Speech and Computer – SPECOM 2005* (pp. 191–194). Patras, Greece.

Gerardi, G. M., & Gerken, L. (1995). The Development of Affective Responses to Modality and Melodic Contour. *Music Perception: An Interdisciplinary Journal*, *12*(3), 279–290. http://doi.org/10.2307/40286184

Gold, A. E., MacLeod, K. M., Thomson, K. J., Frier, B. M., & Deary, I. J. (1995). Cognitive function during insulin-induced hypoglycemia in humans: Short-term cerebral adaptation does not occur. *Psychopharmacology*, *119*(3), 325–333. http://doi.org/10.1007/BF02246299

Gómez, E. (2005). Key estimation from polyphonic audio. In *1st Music Information Retrieval Exchange – MIREX 2005, as part of the 6th International Society for Music Information Retrieval Conference – ISMIR 2005*. London, UK.

Gómez, E. (2006a). *Tonal Description of Music Audio Signals*. Universitat Pompeu Fabra. Retrieved from http://mtg.upf.edu/node/472

Gómez, E. (2006b). Tonal Description of Polyphonic Audio for Music Content Processing. *INFORMS Journal on Computing*, *18*(3), 294–304. http://doi.org/10.1287/ijoc.1040.0126

Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, *7*(2), 377–387. http://doi.org/10.1037/1528-3542.7.2.377

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes Problems*. Oxford, UK: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780198510581.001.0001

Gregory, A. H., Worrall, L., & Sarge, A. (1996). The development of emotional responses to music in young children. *Motivation and Emotion*, *20*(4), 341–348. http://doi.org/10.1007/BF02856522

Grey, J. M. (1975). *An Exploration of Musical Timbre*. Stanford University.

Grosche, P., & Müller, M. (2009). A Mid-Level Representation for Capturing Dominant Tempo and Pulse Information in Music Recordings. In *10th International Society for Music Information Retrieval Conference – ISMIR 2009*. Utrecht, Netherlands.

Gundlach, R. H. (1935). Factors Determining the Characterization of Musical Phrases. *The American Journal of Psychology*, *47*(4), 624. http://doi.org/10.2307/1416007

Gupta, C., Tong, R., Li, H., & Wang, Y. (2018). Semi-Supervised Lyrics and Solo-Singing Alignment. In *19th International Society for Music Information Retrieval Conference – ISMIR 2018* (pp. 600–607). Paris,.

Gurney, E. (1880). *The power of sound*. London, UK: Smith, Elder & Co. Retrieved from https://archive.org/details/thepowerofsound00gurnuoft

Hailstone, J. C., Omar, R., Henley, S. M. D., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It's not what you play, it's how you play it: timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology*, *62*(11), 2141–55. http://doi.org/10.1080/17470210902765957

Han, B.-J., Rho, S., Dannenberg, R. B., & Hwang, E. (2009). Smers: Music Emotion Recognition Using Support Vector Regression. In *10th International Society for Music Information Retrieval Conference – ISMIR 2009*. Kobe, Japan.

Hargreaves, D. J., & North, A. C. (1997). *The social psychology of music*. Oxford, UK: Oxford University Press.

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. In *1st ACM workshop on Audio and music computing multimedia - AMCMM 2006* (Vol. C, p. 21). New York, NY, USA: ACM Press. http://doi.org/10.1145/1178723.1178727

Hartmann, W. M. (1997). *Signals, Sound, and Sensation*. American Institute of Physics.

He, H., Jin, J., Xiong, Y., Chen, B., & Sun, W. (2008). Language feature mining for

music emotion classification via supervised learning from lyrics. In *International Symposium on Intelligence Computation and Applications – ISICA 2008 – Advances in Computation and Intelligence* (pp. 426–435). Wuhan, China: Springerverlag Berlin. http://doi.org/10.1007/978-3-540-92137-0_47

Heinlein, C. P. (1928). The affective characters of the major and minor modes in music. *Journal of Comparative Psychology*, *8*(2), 101–142. http://doi.org/10.1037/h0070573

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. http://doi.org/10.1002/bimj.201700067

Herrera, P., Dehamel, A., & Gouyon, F. (2003). Automatic Labeling of Unpitched Percussion Sounds. In *114th Audio Engineering Society Convention – AES 114*. Amsterdam, Netherlands.

Hevner, K. (1935). The Affective Character of the Major and Minor Modes in Music. *The American Journal of Psychology*, *47*(1), 103. http://doi.org/10.2307/1416710

Hevner, K. (1936). Experimental Studies of the Elements of Expression in Music. *The American Journal of Psychology*, *48*(2), 246. http://doi.org/10.2307/1415746

Hevner, K. (1937). The Affective Value of Pitch and Tempo in Music. *The American Journal of Psychology*, *49*(4), 621. http://doi.org/10.2307/1416385

Hopkins, B., & Skellam, J. G. (1954). A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany*, *18*(70), 213–227.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. http://doi.org/10.1037/h0071325

Hsu, J.-L., Zhen, Y.-L., Lin, T.-C., & Chiu, Y.-S. (2017). Affective content analysis of music emotion through EEG. *Multimedia Systems*, 1–16. http://doi.org/10.1007/s00530-017-0542-0

Hu, X., & Downie, J. S. (2007). Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata. In *8th International Society for Music Information Retrieval Conference – ISMIR 2007* (pp. 67–72). Vienna, Austria: Österreichische Computer Gesellschaft.

Hu, X., & Downie, J. S. (2010a). Improving mood classification in music digital libraries by combining lyrics and audio. In *10th Annual Joint Conference on Digital Libraries - JCDL '10* (p. 159). Gold Coast, Australia: ACM Press. http://doi.org/10.1145/1816123.1816146

Hu, X., & Downie, J. S. (2010b). When lyrics outperform audio for music mood classification: a feature analysis. In J. S. Downie & R. C. Veltkamp (Eds.), *11th International Society for Music Information Retrieval Conference – ISMIR 2010* (pp. 619–

624). Utrecht, Netherlands.

Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. In *10th International Society for Music Information Retrieval Conference – ISMIR 2009* (pp. 411–416). Kobe, Japan.

Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 Mirex Audio Mood Classification Task: Lessons Learned. In *9th International Society of Music Information Retrieval Conference – ISMIR 2008* (pp. 462–467). Philadelphia, Pennsylvania, USA.

Hu, X., & Yang, Y.-H. (2014). Cross-cultural mood regression for music digital libraries. In *IEEE/ACM Joint Conference on Digital Libraries – JCDL* (pp. 471–472). London, UK: IEEE. http://doi.org/10.1109/JCDL.2014.6970230

Hu, X., & Yang, Y.-H. (2017). Cross-Dataset and Cross-Cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs. *IEEE Transactions on Affective Computing – TAFFC*, *8*(2), 228–240. http://doi.org/10.1109/TAFFC.2016.2523503

Huron, D. (2000). Perceptual and cognitive applications in music information retrieval. *Cognition*, *10*(1), 83–92.

Huron, D. (2001). What is a Musical Feature? Forte's Analysis of Brahms's Opus 51, No. 1, Revisited. *The Online Journal of the Society for Music Theory*, *7*(4).

Hutchinson, W., & Knopoff, L. (1978). The acoustic component of western consonance. *Interface*, *7*(1), 1–29. http://doi.org/10.1080/09298217808570246

IFPI. (2017, April 25). IFPI Global Music Report 2017. *International Federation of the Phonographic Industry*. Retrieved from http://www.ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2017

Ilie, G., & Thompson, W. F. (2006). A Comparison of Acoustic Cues in Music and Speech for Three Dimensions of Affect. *Music Perception*, *23*(4), 319–330. http://doi.org/10.1525/mp.2006.23.4.319

Imberty, M. (1979). *Entendre la Musique: Sémantique Psychologique de la Musique*. Dunod.

Imbrasaite, V., Baltrusaitis, T., & Robinson, P. (2013). Emotion tracking in music using continuous conditional random fields and relative feature representation. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (pp. 1–6). San Jose, CA, USA: IEEE. http://doi.org/10.1109/ICMEW.2013.6618357

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, *57*(S1), S35–S35. http://doi.org/10.1121/1.1995189

Izard, C. E. (2010a). More Meanings and More Questions for the term "Emotion." *Emotion Review*, *2*(4), 383–385. http://doi.org/10.1177/1754073910374670

Izard, C. E. (2010b). The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation, and Regulation. *Emotion Review*, *2*(4), 363–370. http://doi.org/10.1177/1754073910374661

James, W. (1884). What is an Emotion? *Mind*, *9*(34), 188–205. http://doi.org/10.1093/mind/os-IX.34.188

James, W. (1994). The physical basis of emotion (1894 reprint). *Psychological Review*, *101*(2), 205–210. http://doi.org/10.1037/0033-295X.101.2.205

Jensen, K. (1999). *Timbre Models of Musical Sounds*. Københavns Universitet, Datalogisk Institut.

Jensen, K., & Andersen, T. H. (2003). Beat estimation on the beat. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (Vol. 2003–Janua, pp. 87–90). New Paltz, NY, USA: IEEE. http://doi.org/10.1109/ASPAA.2003.1285826

Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H., & Cai, L.-H. (2002). Music type classification by spectral contrast feature. In *IEEE International Conference on Multimedia and Expo – ICME 2002* (Vol. 1, pp. 113–116). Lausanne, Switzerland: IEEE. http://doi.org/10.1109/ICME.2002.1035731

Johnson-Laird, P. N., & Oatley, K. (1992). Basic emotions, rationality, and folk theory. *Cognition & Emotion*, *6*(3–4), 201–223. http://doi.org/10.1080/02699939208411069

Johnson, M. L., & Larson, S. (2003). "Something in the Way She Moves"-Metaphors of Musical Motion. *Metaphor and Symbol*, *18*(2), 63–84. http://doi.org/10.1207/S15327868MS1802_1

Juslin, P. N. (1997). Perceived Emotional Expression in Synthesized Performances of a Short Melody: Capturing the Listener's Judgment Policy. *Musicae Scientiae*, *1*(2), 225–256. http://doi.org/10.1177/102986499700100205

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: relating performance to perception. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(6), 1797–813.

Juslin, P. N. (2013). What does music express? Basic emotions and beyond. *Frontiers in Psychology*, *4*, 596. http://doi.org/10.3389/fpsyg.2013.00596

Juslin, P. N., & Laukka, P. (2004). Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research*, *33*(3), 217–238. http://doi.org/10.1080/0929821042000317813

Juslin, P. N., & Timmers, R. (2011). Expression And Communication of Emotion in Music Performance. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of Music and*

Emotion: Theory, Research, Applications (pp. 452–489). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199230143.003.0017

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(05), 559–621. http://doi.org/10.1017/S0140525X08005293

Kagan, J. (2003). Behavioral inhibition as a temperamental category. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences (Series in Affective Science)* (pp. 320–331). Oxford University Press.

Kaiser, F., & Peeters, G. (2012). Adaptive Temporal Modeling of Audio Features in the Context of Music Structure Segmentation. In *International Workshop on Adaptive Multimedia Retrieval* (Vol. 8382, pp. 248–261). Springer.

Kartomi, M. J. (1990). *On Concepts and Classifications of Musical Instruments*. University of Chicago Press.

Kastner, M. P., & Crowder, R. G. (1990). Perception of the Major/Minor Distinction: IV. Emotional Connotations in Young Children. *Music Perception: An Interdisciplinary Journal*, *8*(2), 189–201. http://doi.org/10.2307/40285496

Katayose, H., Imai, M., & Inokuchi, S. (1988). Sentiment extraction in music. In *9th International Conference on Pattern Recognition* (pp. 1083–1087). Rome, Italy: IEEE Comput. Soc. Press. http://doi.org/10.1109/ICPR.1988.28447

Khvedelidze, B. V. (1990). Hilbert transform. In M. Hazewinkel (Ed.), *Encyclopaedia of Mathematics*. Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-009-5991-0

Kim, J., Lee, S., Kim, S., & Yoo, W. Y. (2011). Music mood classification model based on arousal-valence values. In *13th International Conference on Advanced Communication Technology – ICACT 2011* (pp. 292–295). Gangwon-Do, South Korea.

Kim, Y. E., Schmidt, E. M., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection. In *9th International Society of Music Information Retrieval Conference – ISMIR 2008* (pp. 231–236). Philadelphia, Pennsylvania, USA.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., … Turnbull, D. (2010). Music Emotion Recognition: A State of the Art Review. In *11th International Society for Music Information Retrieval Conference – ISMIR 2010* (pp. 255–266). Utrecht, Netherlands.

Kira, K., & Rendell, L. A. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. In W. R. Swartout (Ed.), *10th National Conference on Artificial Intelligence* (pp. 129–134). San Jose, CA, USA: AAAI Press.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. In

1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258) (pp. 3089–3092 vol.6). IEEE. http://doi.org/10.1109/ICASSP.1999.757494

Kleinen, G. (1968). *Experimentelle Studien zum musikalischen Ausdruck [Experimental studies on musical expression]*. Hamburg, Germany: Universität Hamburg.

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and Emotion*, *5*(3), 263–291. http://doi.org/10.1007/BF00993889

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., … Patras, I. (2012). DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing – TAFFC*, *3*(1), 18–31. http://doi.org/10.1109/T-AFFC.2011.15

Konečni, V. J. (2008). Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 115–129. http://doi.org/10.1037/1931-3896.2.2.115

Konečni, V. J., & Karno, M. P. (1994). Empirical investigations of the hedonic and emotional effects of musical structure. *Musik Psychologie*, *11*, 119–137.

Konishi, T., Imaizumi, S., & Niimi, S. (2000). Vibrato and emotion in singing voice. In *6th International Conference for Music Perception and Cognition – ICMPC 2000*.

Kononenko, I., Šimec, E., & Robnik-Šikonja, M. (1997). Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence*, *7*(1), 39–55. http://doi.org/10.1023/A:1008280620621

Korhonen, M. D., Clausi, D. A., & Jernigan, M. E. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *36*(3), 588–599. http://doi.org/10.1109/TSMCB.2005.862491

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caracterisation du timbre des sons complexes II: Analysis acoustiques et quantificata psychophysics. *Journal de Physique*, *C5*, 625–628.

Krippendorff, K. H. (2003). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Krumhansl, C. L. (1996). A Perceptual Analysis of Mozart's Piano Sonata K. 282: Segmentation, Tension, and Musical Ideas. *Music Perception: An Interdisciplinary Journal*, *13*(3), 401–432. http://doi.org/10.2307/40286177

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology*, *51*(4), 336–353. http://doi.org/10.1037/1196-1961.51.4.336

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. http://doi.org/10.1007/BF02289565

Lagrange, M., Martins, L. G., & Tzanetakis, G. (2008). A Computationally Efficient Scheme for Dominant Harmonic Source Separation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 165–168). IEEE. http://doi.org/10.1109/ICASSP.2008.4517572

Laitz, S. G. (2007). *The Complete Musician* (2nd ed.). Oxford University Press, USA.

Lakoff, G., & Johnson, M. (1980a). Conceptual Metaphor in Everyday Language. *The Journal of Philosophy*, *77*(8), 453. http://doi.org/10.2307/2025464

Lakoff, G., & Johnson, M. (1980b). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*(2), 195–208. http://doi.org/10.1016/S0364-0213(80)80017-6

Lamere, P. (2008). Social Tagging and Music Information Retrieval. *Journal of New Music Research*, *37*(2), 101–114. http://doi.org/10.1080/09298210802479284

Lane, D. M. (2009). Introduction To Linear Regression. In *Introduction to Statistics* (pp. 462–468). Houston, Texas. Retrieved from http://onlinestatbook.com/

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biological Psychiatry*, *44*(12), 1248–63.

Langer, S. K. (1957). *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite, and Art*. Harvard Univ Press.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. Clark (Ed.), *Emotion (The Review of Personality and Social Psychology - Vol. 3)* (pp. 25–59). Thousand Oaks, CA, US: Sage Publications.

Lartillot, O. (2010). mirtempo: Tempo estimation through advanced frame-by-frame peaks tracking. In *6th Music Information Retrieval Exchange – MIREX 2010, as part of the 11th International Society for Music Information Retrieval Conference – ISMIR 2010*. Ustrón, Poland.

Lartillot, O. (2018). MIR Toolbox 1.7.1 User's Manual. Oslo, Norway: University of Oslo.

Lartillot, O., & Cereghetti, D. (2013). A Simple, High-Yield Method for Assessing Structural Novelty. In *3rd International Conference of Music and Emotion – ICME 2013*. Jyväskylä, Finland.

Lartillot, O., Cereghetti, D., Eliard, K., & Trost, W. J. (2013). Estimating Tempo and Metrical Features by Tracking the Whole Metrical Hierarchy. In G. Luck & O. Brabant (Eds.), *3rd International Conference on Music & Emotion – ICME 2013*.

Jyväskylä, Finland.

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation, and optimization. In *9th International Society of Music Information Retrieval Conference – ISMIR 2008* (pp. 521–526). Philadelphia, Pennsylvania, USA.

Lartillot, O., & Toiviainen, P. (2007). A Matlab Toolbox for Musical Feature Extraction from Audio. In *10th International Conference on Digital Audio Effects – DAFx 2007* (pp. 237–244). Bordeaux, France.

Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633–653. http://doi.org/10.1080/02699930441000445

Laurier, C. (2011). *Automatic Classification of Musical Mood by Content-Based Analysis*. Universitat Pompeu Fabra. Retrieved from http://mtg.upf.edu/node/2385

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal Music Mood Classification Using Audio and Lyrics. In *7th International Conference on Machine Learning and Applications – ICMLA 2008* (pp. 688–693). San Diego, CA, USA. http://doi.org/10.1109/ICMLA.2008.96

Laurier, C., & Herrera, P. (2007). Audio Music Mood Classification Using Support Vector Machine. In *8th International Society for Music Information Retrieval Conference – ISMIR 2007* (pp. 2–4). Vienna, Austria.

Laurier, C., Lartillot, O., Eerola, T., & Toiviainen, P. (2009). Exploring relationships between audio features and emotion in music. In *7th Triennial Conference of European Society for the Cognitive Sciences of Music – ESCOM 2009* (Vol. 3, pp. 260–264). Jyväskylä, Finland: European Society for Cognitive Sciences of Music. http://doi.org/10.3389/conf.neuro.09.2009.02.033

Law, E., Ahn, L. von, Dannenberg, R. B., & Crawford, M. J. (2007). TagATune: A Game for Music and Sound Annotation. In *8th International Society for Music Information Retrieval Conference – ISMIR 2007*. Vienna, Austria.

Law, E., West, K., Mandel, M., Bay, M., & Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. In *10th International Society for Music Information Retrieval Conference - ISMIR 2009* (pp. 387–392). Kobe, Japan.

Lee, J. H., & Downie, J. S. (2004). Survey of Music Information Needs, Uses, and Seeking Behaviours: Preliminary Findings. In *5th International Conference on Music Information Retrieval – ISMIR 2004* (pp. 441–446). Barcelona, Spain. http://doi.org/10.1109/ISM.2009.123

Leman, M., Vermeulen, V., de Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of musical affect using a combination of acoustic structural cues. *Journal of New Music Research*. http://doi.org/10.1080/09298210500123978

Li, T., & Ogihara, M. (2003). Detecting emotion in music. In *4th International Symposium on Music Information Retrieval – ISMIR 2003* (pp. 239–240). Baltimore, Maryland, USA.

Li, T., & Ogihara, M. (2004). Content-based music similarity search and emotion detection. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP 2004* (Vol. 5, p. V-705-8). Montreal, Quebec, Canada: IEEE. http://doi.org/10.1109/ICASSP.2004.1327208

Lichte, W. H. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, *28*(6), 455–480. http://doi.org/10.1037/h0053526

Lin, Y.-C., Yang, Y.-H., Chen, H. H., Liao, I.-B., & Ho, Y.-C. (2009). Exploiting genre for music emotion classification. In *2009 IEEE International Conference on Multimedia and Expo – ICME 2009* (pp. 618–621). New York, NY, USA: IEEE. http://doi.org/10.1109/ICME.2009.5202572

Lindström, E. (1999). Expression in Music: Interaction between Performance and Melodic Structure. In *Meeting of the Society for Music Perception and Cognition*. Evanston, Illinois, USA.

Lindström, E. (2006). Impact of melodic organization on perceived structure and emotional expression in music. *Musicae Scientiae*, *10*(1), 85–117. http://doi.org/10.1177/102986490601000105

Liu, D., & Lu, L. (2003). Automatic Mood Detection from Acoustic Music Data. *Int. J. on the Biology of Stress*, *8*(6), 359–377.

Liu, J.-Y., Liu, S.-Y., & Yang, Y.-H. (2014). LJ2M dataset: Toward better understanding of music listening behavior and user mood. In *2014 IEEE International Conference on Multimedia and Expo – ICME 2014* (pp. 1–6). Chengdu, China: IEEE. http://doi.org/10.1109/ICME.2014.6890172

Livingstone, S. R., & Brown, A. R. (2005). Dynamic Response: Real-Time Adaptation for Music Emotion. In *2nd Australasian Conference on Interactive Entertainment* (pp. 105–111). Sydney, Australia.

Lu, L., Liu, D., & Zhang, H.-J. (2006). Automatic Mood Detection and Tracking of Music Audio Signals. *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, *14*(1), 5–18. http://doi.org/10.1109/TSA.2005.860344

Lu, Q., Chen, X., Yang, D., & Wang, J. (2010). Boosting for Multi-Modal Music Emotion Classification. In J. S. Downie & R. C. Veltkamp (Eds.), *11th International Society for Music Information Retrieval Conference - ISMIR 2010* (pp. 105–110). Utrecht, Netherlands: International Society for Music Information Retrieval.

Lusa. (2009, January 29). Concertos de música ligeira geraram 30 milhões em 2007. *RTP Notícias*. Retrieved from https://www.rtp.pt/noticias/cultura/concertos-de-musica-ligeira-geraram-30-milhoes-em-2007_n168841

Ma, A., Sethi, I., & Patel, N. (2009). Multimedia Content Tagging Using Multilabel Decision Tree. In *2009 11th IEEE International Symposium on Multimedia* (pp. 606–611). San Diego, CA, USA: IEEE. http://doi.org/10.1109/ISM.2009.87

MacDorman, K. F., Ough, S., & Ho, C. C. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, *36*(4), 281–299. http://doi.org/10.1080/09298210801927846

Madsen, J., Jensen, B. S., & Larsen, J. (2015). Learning Combinations of Multiple Feature Representations for Music Emotion Prediction. In *1st International Workshop on Affect & Sentiment in Multimedia – ASM 2015* (pp. 3–8). Brisbane, Australia. http://doi.org/10.1145/2813524.2813534

Maher, T. F. (1980). A Rigorous Test of the Proposition That Musical Intervals Have Different Psychological Effects. *The American Journal of Psychology*, *93*(2), 309. http://doi.org/10.2307/1422235

Maher, T. F., & Berlyne, D. E. (1982). Verbal and Exploratory Responses to Melodic Musical Intervals. *Psychology of Music*, *10*(1), 11–27. http://doi.org/10.1177/0305735682101002

Makris, D., Karydis, I., & Sioutas, S. (2015). The Greek Music Dataset. In *16th International Conference on Engineering Applications of Neural Networks (INNS) - EANN '15* (pp. 1–7). Rhodes, Greece: ACM Press. http://doi.org/10.1145/2797143.2797175

Makris, D., Kermanidis, K. L., & Karydis, I. (2014). The Greek Audio Dataset. In *Artificial Intelligence Applications and Innovations - AIAI 2014* (pp. 165–173). Rhodos, Greece: Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-662-44722-2_18

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2013). Music Emotion Recognition from Lyrics: A Comparative Study. In *6th International Workshop on Music and Machine Learning – MML 2013 – in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2013*. Prague, Czech Republic.

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016a). Bi-modal music emotion recognition: Novel lyrical features and dataset. In *9th International Workshop on Music and Machine Learning – MML 2016 – in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2016*. Riva del Garda, Italy.

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016b). Classification and Regression of Music Lyrics: Emotionally-Significant Features. In *8th International Conference on Knowledge Discovery and Information Retrieval – KDIR 2016*. Porto,

Portugal. http://doi.org/10.5220/0006037400450055

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2018). Emotionally-Relevant Features for Classification and Regression of Music Lyrics. *IEEE Transactions on Affective Computing – TAFFC*, *9*(2), 240–254. http://doi.org/10.1109/TAFFC.2016.2598569

Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., & Jarina, R. (2017). Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition. In *14th Sound & Music Computing Conference – SMC 2017* (pp. 208–213). Espoo, Finland.

Malloch, S. N. (1997). *Timbre and Technology: an analytical partnership*. University of Edinburgh. Retrieved from http://hdl.handle.net/1842/21394

Malm, W. P. (1995). *Music cultures of the Pacific, the Near East, and Asia* (3rd edition). Prentice Hall.

Margulis, E. H. (2007). Silences in Music are Musical Not Silent: An Exploratory Study of Context Effects on the Experience of Musical Pauses. *Music Perception: An Interdisciplinary Journal*, *24*(5), 485–506. http://doi.org/10.1525/mp.2007.24.5.485

Masri, P., & Bateman, A. (1996). Improved modelling of attack transients in music analysis-resynthesis. In *International Computer Music Conference – ICMC 1996* (pp. 100–103).

McAdams, S., & Bregman, A. (1979). Hearing Musical Streams. *Computer Music Journal*, *3*, 26–43, 60. http://doi.org/10.2307/4617866

McCosh, J. (1880). *The emotions*. New York, US: Charles Scribner's Sons. Retrieved from https://archive.org/details/theemotions00mccouoft

McCulloch, R. (1999). *Modality and children's affective responses to music [Undergraduate project for Perception and Performance course (Ian Cross, instructor)]*. Retrieved from http://www.mus.cam.ac.uk/~ic108/PandP/%0AMcCulloch99/McCulloch99.html%0A

Mcennis, D., Mckay, C., Fujinaga, I., & Depalle, P. (2005). JAudio: A feature extraction library. In *6th International Conference on Music Information Retrieval – ISMIR 2005*. London, UK.

Mcfee, B., Raffel, C., Liang, D., Ellis, D. P. W., Mcvicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. In *14th Python in Science Conference – SciPy 2015* (pp. 18–25). Austin, Texas.

Mckay, C. (2010). *Automatic Music Classification with jMIR*. McGill University.

McKinney, J. C. (1994). *The diagnosis and correction of vocal faults: a manual for teachers of singing and for choir directors*. Genevox Music Group.

Mcvicar, M., & Freeman, T. (2011). Mining the Correlation Between Lyrical and Audio

Features and the Emergence of Mood. In *12th International Society for Music Information Retrieval Conference – ISMIR 2011* (pp. 783–788). Miami, Florida, USA.

Melanson, D. (2011, April 10). Apple: 16 billion iTunes songs downloaded, 300 million iPods sold. *Engadget*. Retrieved from https://www.engadget.com/2011/10/04/apple-16-billion-itunes-songs-downloaded-300-million-ipods-sol/

Meng, A., Ahrendt, P., Larsen, J., & Hansen, L. K. (2007). Temporal Feature Integration for Music Genre Classification. *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, *15*(5), 275–9. http://doi.org/10.1155/2007/36409

Meyer, L. B. (1973). *Explaining Music: Essays and Explorations*. University of California Press.

Meyers, O. C. (2007). *A Mood-Based Music Classification and Exploration System*. MIT Press. Retrieved from https://dspace.mit.edu/handle/1721.1/39337

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 301–312. http://doi.org/10.1109/34.990133

Moffat, D., Ronan, D., & Reiss, J. D. (2015). An Evaluation of Audio Feature Extraction Toolboxes. In *18th Int. Conference on Digital Audio Effects - DAFx-15* (pp. 1–7). Trondheim, Norway.

Montaigne, M. de. (1603). *The essayes, or, Morall, politike and millitarie discourses of Lo. Michaell de Montaigne (Florio, J. Translation)*. London, UK: Printed by Val. Sims for Edward Blount dwelling in Paules churchyard. Retrieved from https://archive.org/details/essayesormorallp00mont

Myint, E. E. P., & Pwint, M. (2010). An approach for multi-label music mood classification. In *2010 2nd International Conference on Signal Processing Systems* (pp. 290–294). IEEE. http://doi.org/10.1109/ICSPS.2010.5555619

Noll, A. M. (1967). Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, *41*(2), 293–309. http://doi.org/10.1121/1.1910339

Oliveira, A. P., & Cardoso, A. (2010). A musical system for emotional expression. *Knowledge-Based Systems*, *23*(8), 901–913. http://doi.org/10.1016/j.knosys.2010.06.006

Oliveira, J. L., Gouyon, F., Martins, L. G., & Reis, L. P. (2010). IBT: A Real-Time Tempo and Beat Tracking System. In *11th International Society for Music Information Retrieval Conference - ISMIR 2010*. Utrecht, Netherlands.

Oppenheim, A. V. (1965). *Superposition in a class of nonlinear systems*. MIT Research Laboratory of Electronics. Retrieved from https://dspace.mit.edu/handle/1721.1/4393

Ortony, A., & Turner, T. J. (1990). What's Basic About Basic Emotions? *Psychological Review*, *97*(3), 315–331.

Owen, H. (2000). *Music theory resource book*. Oxford University Press.

Pachet, F., & Zils, A. (2003). Evolving Automatically High-Level Music Descriptors From Acoustic Signals. In *1st International Symposium on Computer Music Modeling and Retrieval – CMMR 2003* (Vol. 2771, pp. 1–13). Montpellier, France: Springer.

Paiva, R. P. (2006). *Melody Detection in Polyphonic Audio*. University of Coimbra. Retrieved from http://hdl.handle.net/10316/10159

Paiva, R. P., Mendes, T., & Cardoso, A. (2006). Melody Detection in Polyphonic Musical Signals: Exploiting Perceptual Rules, Note Salience, and Melodic Smoothness. *Computer Music Journal*, *30*(4), 80–98. http://doi.org/10.1162/comj.2006.30.4.80

Pampalk, E., Rauber, A., & Merkl, D. (2002). Content-based organization and visualization of music archives. In *10th ACM International Conference on Multimedia – MULTIMEDIA 2002*. Juan-les-Pins, France: ACM Press. http://doi.org/10.1145/641007.641121

Panda, R., Malheiro, R., & Paiva, R. P. (2018). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing – TAFFC*, 1–1. http://doi.org/10.1109/TAFFC.2018.2820691

Panda, R., Malheiro, R., Rocha, B., Oliveira, A. P., & Paiva, R. P. (2013). Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. In *10th International Symposium on Computer Music Multidisciplinary Research – CMMR 2013* (pp. 570–582). Marseille, France.

Panda, R., & Paiva, R. P. (2011a). Automatic Creation of Mood Playlists in the Thayer Plane: A Methodology and a Comparative Study. In *8th Sound and Music Computing Conference – SMC 2011*. Padova, Italy. http://doi.org/10.5281/zenodo.849887

Panda, R., & Paiva, R. P. (2011b). Using Support Vector Machines for Automatic Mood Tracking in Audio Music. In *130th Audio Engineering Society Convention – AES 130*. London, UK.

Panda, R., & Paiva, R. P. (2012a). MIREX 2012: Mood Classification Tasks Submission. In *8th Music Information Retrieval Exchange – MIREX 2012, as part of the 13th International Society for Music Information Retrieval Conference – ISMIR 2012*. Porto, Portugal.

Panda, R., & Paiva, R. P. (2012b). Music Emotion Classification: Dataset Acquisition and Comparative Analysis. In *15th International Conference on Digital Audio Effects – DAFx 2012*. York, UK.

Panda, R., Rocha, B., & Paiva, R. P. (2013). Dimensional music emotion recognition:

Combining standard and melodic audio features. In *10th International Symposium on Computer Music Multidisciplinary Research – CMMR 2013* (pp. 583–593). Marseille, France.

Panda, R., Rocha, B., & Paiva, R. P. (2015). Music Emotion Recognition with Standard and Melodic Audio Features. *Applied Artificial Intelligence – AAI*, *29*(4), 313–334. http://doi.org/10.1080/08839514.2015.1016389

Panksepp, J. (1992). A Critical Role for "Affective Neuroscience" in Resolving What Is Basic About Basic Emotions. *Psychological Review*, *99*(3), 554–60.

Panksepp, J. (1998). *Affective neuroscience: the foundations of human and animal emotions*. Oxford University Press.

Pannese, A., Rappaz, M.-A., & Grandjean, D. (2016). Metaphor and music emotion: Ancient views and future directions. *Consciousness and Cognition*, *44*, 61–71. http://doi.org/10.1016/j.concog.2016.06.015

Pao, T.-L., Cheng, Y.-M., Yeh, J.-H., Chen, Y.-T., Pai, C.-Y., & Tsai, Y.-W. (2008). Comparison between weighted D-KNN and other classifiers for music emotion recognition. In *3rd International Conference on Innovative Computing Information and Control – ICICIC 2008* (pp. 2–5). Dalian, China. http://doi.org/10.1109/ICICIC.2008.679

Parkinson, C., Kohler, P. J., Sievers, B., & Wheatley, T. (2012). Associations between Auditory Pitch and Visual Elevation Do Not Depend on Language: Evidence from a Remote Population. *Perception*, *41*(7), 854–861. http://doi.org/10.1068/p7225

Parncutt, R. (1989). *Harmony: A Psychoacoustical Approach* (Vol. 19). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-74831-8

Patra, B. G., Das, D., & Bandyopadhyay, S. (2013). Unsupervised Approach to Hindi Music Mood Classification. In *Mining Intelligence and Knowledge Exploration - MIKE 2003* (pp. 62–69). Tamil Nadu, India: Springer, Cham. http://doi.org/10.1007/978-3-319-03844-5_7

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. http://doi.org/10.1080/14786440109462720

Peckham, A. (2005). *Vocal workouts for the contemporary singer*. Berklee Press.

Peckham, A., Crossen, J., Gebhardt, T., & Shrewsbury, D. (2010). *The Contemporary Singer: Elements of Vocal Technique*. Berklee Press.

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. *CUIDADO IST Project Report*.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of*

the Acoustical Society of America, *130*(5), 2902–2916. http://doi.org/10.1121/1.3642604

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., & Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, *49*(2), 1685–1689. http://doi.org/10.1103/PhysRevE.49.1685

Plack, C. J., & Oxenham, A. J. (2005). Overview: The Present and Future of Pitch. In C. J. Plack, A. J. Oxenham, & R. R. Fay (Eds.), *Pitch: Neural Coding and Perception* (pp. 1–6). New York: Springer-Verlag. http://doi.org/10.1007/0-387-28958-5_1

Plato (375 B.C.). (1969). Republic III. In *Plato in Twelve Volumes, Vols. 5 & 6 (Shorey, P. Trans.)*. Cambridge, MA: Harvard University Press. Retrieved from http://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0059.tlg030

Plewa, M., & Kostek, B. (2012). A Study on Correlation between Tempo and Mood of Music. In *133th Audio Engineering Society Convention – AES 133*. San Francisco, California, USA.

Plomp, R., & Levelt, W. J. M. (1965). Tonal Consonance and Critical Bandwidth. *The Journal of the Acoustical Society of America*, *38*(4), 548–560. http://doi.org/10.1121/1.1909741

Pohle, T., Pampalk, E., & Widmer, G. (2005). Evaluation of Frequently Used Audio Features for Classification of Music Into Perceptual Categories. In *4th International Workshop on Content-Based Multimedia Indexing – CBMI 2005* (Vol. 162). Riga, Latvia.

Pollard, H. F., & Jansson, E. V. (1982). A Tristimulus Method for the Specification of Musical Timbre. *Acta Acustica United with Acustica*, *51*(3), 162–171.

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., & Serra, X. (2018). End-to-end learning for music audio tagging at scale. In *19th International Society for Music Information Retrieval Conference – ISMIR 2018* (pp. 637–644). Paris, France.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–34. http://doi.org/10.1017/S0954579405050340

Powers, H. S. (1958). Mode and Raga. *The Musical Quarterly*, *44*(4), 448–460. http://doi.org/10.1093/mq/XLIV.4.448

Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., & McGonegal, C. A. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing – TASSP*, *24*(5), 399–418. http://doi.org/10.1109/TASSP.1976.1162846

Rault, L. (2000). *Musical Instruments: A Worldwide Survey of Traditional Music-Making*. Thames & Hudson.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning From Crowds. *The Journal of Machine Learning Research*, *11*, 1297–1322.

Read, G. (1953). *Thesaurus of Orchestral Devices*. New York, Pitman Pub. Corp. Retrieved from https://archive.org/details/thesaurusoforche00read

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In M. Tamer & L. Liu (Eds.), *Encyclopedia of Database Systems* (2nd Editio, Vol. 25, pp. 532–538). Boston, MA: Springer US. http://doi.org/10.1007/978-0-387-39940-9_565

Rentfrow, P. J., & McDonald, J. A. (2010). Preference, personality, and emotion. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 669–695). Oxford University Press.

Ricard, J. (2004). *Towards computational morphological description of sound*. Universitat Pompeu Fabra. Retrieved from http://mtg.upf.edu/node/2231

Rigg, M. G. (1940). Speed as a determiner of musical mood. *Journal of Experimental Psychology*, *27*(5), 566–571. http://doi.org/10.1037/h0058652

Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. In *14th International Conference on Machine Learning - ICML1997* (pp. 296–304). Nashville, Tennessee, USA: Morgan Kaufmann Publishers.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, *53*(1–2), 23–69. http://doi.org/10.1023/A:1025667309714

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326. http://doi.org/10.1126/science.290.5500.2323

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. http://doi.org/10.1037/h0077714

Russell, J. A. (1983). Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, *45*(6), 1281–1288. http://doi.org/10.1037/0022-3514.45.6.1281

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–72.

Russell, J. A., & Fehr, B. (1994). Fuzzy concepts in a fuzzy hierarchy: varieties of anger. *Journal of Personality and Social Psychology*, *67*(2), 186–205.

Saari, P., & Eerola, T. (2014). Semantic Computing of Moods Based on Tags in Social Media of Music. *IEEE Transactions on Knowledge and Data Engineering – TKDE*, *26*(10), 2548–2560. http://doi.org/10.1109/TKDE.2013.128

Saarni, C. (1999). *The development of emotional competence*. Guilford Press.

Salamon, J., & Gómez, E. (2012). Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, *20*(6), 1759–1770. http://doi.org/10.1109/TASL.2012.2188515

Salamon, J., Rocha, B., & Gómez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing –ICASSP 2012* (pp. 81–84). Kyoto, Japan: IEEE. http://doi.org/10.1109/ICASSP.2012.6287822

Sanden, C., & Zhang, J. Z. (2011). An Empirical Study of Multi-label Classifiers for Music Tag Annotation. In *12th International Society for Music Information Retrieval Conference – ISMIR 2011* (pp. 717–722). Miami, Florida, USA.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends in Information Retrieval*, *8*(2–3), 127–261. http://doi.org/10.1561/1500000042

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, *103*(1), 588–601. http://doi.org/10.1121/1.421129

Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality & Social Psychology*, *5*, 37–63.

Scherer, K. R. (2004). Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them? *Journal of New Music Research*, *33*(3), 239–251. http://doi.org/10.1080/0929821042000317822

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. http://doi.org/10.1177/0539018405058216

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences (Series in Affective Science)* (pp. 433–456). Oxford, UK: Oxford University Press.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346. http://doi.org/10.1007/BF00992539

Scherer, K. R., Sundberg, J., Tamarit, L., & Salomão, G. L. (2015). Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, *29*(1), 218–235. http://doi.org/10.1016/J.CSL.2013.10.002

Schimmack, U., & Grob, A. (2000). Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. *European Journal of*

Personality, *14*(4), 325–345. http://doi.org/song

Schimmack, U., & Reisenzein, R. (2002). *Experiencing activation: energetic arousal and tense arousal are not mixtures of valence and activation. Emotion* (Vol. 2). American Psychological Association. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12899373

Schmidt, E. M., & Kim, Y. E. (2009). Projection of acoustic features to continuous valence-arousal mood labels via regression. In *10th International Society for Music Information Retrieval Conference – ISMIR 2009* (Vol. 14, pp. 2009–2009). Kobe, Japan.

Schmidt, E. M., & Kim, Y. E. (2010a). Prediction of Time-Varing Musical mood distributions from audio. In *11th International Society for Music Information Retrieval Conference – ISMIR 2010*. Utrecht, Netherlands.

Schmidt, E. M., & Kim, Y. E. (2010b). Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. In *9th International Conference on Machine Learning and Applications – ICMLA 2010* (pp. 655–660). Washington, DC, USA: IEEE. http://doi.org/10.1109/ICMLA.2010.101

Schmidt, E. M., & Kim, Y. E. (2011). Learning emotion-based acoustic features with deep belief networks. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics - WASPAA 2011* (pp. 65–68). New Paltz, NY, USA: IEEE. http://doi.org/10.1109/ASPAA.2011.6082328

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *11th ACM International Conference on Multimedia Information Retrieval – MIR 2010* (pp. 267–274). Philadelphia, Pennsylvania, USA: ACM Press. http://doi.org/10.1145/1743384.1743431

Schneider, M. (2018, July 26). Spotify Reports 83 Million Subscribers as Losses Widen and Revenue Increases. *Billboard*. Retrieved from https://www.billboard.com/articles/business/8467162/spotify-quarterly-financial-results-subscribers-revenue-losses

Schonfeld, E. (2009, January 6). iTunes Sells 6 Billion Songs, And Other Fun Stats From The Philnote. *TechCrunch*. Retrieved from https://techcrunch.com/2009/01/06/itunes-sells-6-billion-songs-and-other-fun-stats-from-the-philnote/

Schubert, E. (1999). *Measurement and Time Series Analysis of Emotion in Music*. Awarded by: University of New South Wales. School of Music and Music Education. Retrieved from http://handle.unsw.edu.au/1959.4/18268

Schubert, E. (2003). Update of the Hevner Adjective Checklist. *Perceptual and Motor Skills*, *96*(3_suppl), 1117–1122. http://doi.org/10.2466/pms.2003.96.3c.1117

Sethares, W. A. (2005). *Tuning, Timbre, Spectrum, Scale*. Springer.

Seyerlehner, K., Widmer, G., Schedl, M., & Knees, P. (2010). Automatic music tag classification based on blocklevel features. In *7th Sound and Music Computing Conference - SMC2010*. Barcelona, Spain.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*. http://doi.org/10.1145/584091.584093

Shao, Y., Jin, Z., Wang, D., & Srinivasan, S. (2009). An auditory-based feature for robust speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2009* (pp. 4625–4628). Taipei, Taiwan: IEEE. http://doi.org/10.1109/ICASSP.2009.4960661

Shen, J.-L., Hung, J.-W., & Lee, L.-S. (1998). Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments. In *5th International Conference on Spoken Language Processing – ICSLP 1998, Incorporating The 7th Australian International Speech Science and Technology Conference*. Sydney, Australia.

Skovenborg, E., & Nielsen, S. H. (2004). Evaluation of different loudness models with music and speech material. In *117th Audio Engineering Society Convention – AES 117*. San Francisco, California, USA.

Skowronek, J., McKinney, M. F., & Par, S. van de. (2006). Ground-truth for automatic music mood classification. In *7th International Conference on Music Information Retrieval – ISMIR 2006* (pp. 4–5). Victoria, BC, Canada.

Skowronek, J., McKinney, M. F., & Par, S. van de. (2007). A Demonstrator for Automatic Music Mood Estimation. In *8th International Society for Music Information Retrieval Conference – ISMIR 2007* (pp. 345–346). Vienna, Austria.

Smith, L. D., & Williams, R. N. (1999). Children's Artistic Responses to Musical Intervals. *The American Journal of Psychology*, *112*(3), 383. http://doi.org/10.2307/1423638

Soleymani, M., Aljanaki, A., & Yang, Y.-H. (2016). *DEAM: MediaEval Database for Emotional Analysis in Music*. Geneva, Switzerland.

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., & Yang, Y.-H. (2013). 1000 Songs for Emotional Analysis of Music. In *2nd ACM International Workshop on Crowdsourcing for Multimedia – CrowdMM '13* (pp. 1–6). Barcelona, Spain. http://doi.org/10.1145/2506364.2506365

Soleymani, M., & Larson, M. (2010). Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. In *Workshop on Crowdsourcing for Search Evaluation - SIGIR 2010*. Geneva, Switzerland.

Song, L., Smola, A. J., Gretton, A., Borgwardt, K. M., & Bedo, J. (2007). Supervised feature selection via dependence estimation. In *24th international conference on Machine learning - ICML '07* (pp. 823–830). New York, New York, USA: ACM Press. http://doi.org/10.1145/1273496.1273600

Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation Of Musical Features For Emotion Classification. In *13th International Society for Music Information Retrieval Conference – ISMIR 2012* (pp. 523–528). Porto, Portugal.

Sorabji, R. (2002). *Emotion and Peace of Mind: From Stoic Agitation to Christian Temptation*. Oxford, UK: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199256600.001.0001

Soulodre, G. A. (2004). Evaluation of Objective Loudness Meters. In *116th Audio Engineering Society Convention – AES 116*. Berlin, Germany.

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. In *12th International Society for Music Information Retrieval Conference – ISMIR 2011* (pp. 549–554). Miami, FL, USA.

Steinbeis, N., & Koelsch, S. (2008). Comparing the Processing of Music and Language Meaning Using EEG and fMRI Provides Evidence for Similar and Distinct Neural Representations. *PLoS ONE*, *3*(5), e2226. http://doi.org/10.1371/journal.pone.0002226

Stevens, S. S. (1934). The Volume and Intensity of Tones. *The American Journal of Psychology*, *46*(3), 397. http://doi.org/10.2307/1415591

Stevens, S. S. (1975). Psychophysics: Introduction to its perceptual, neural, and social prospects. *Transaction Publishers*.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190. http://doi.org/10.1121/1.1915893

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. (M. I. T. Press, Ed.)*The MIT Press* (Vol. 08). MIT Press.

Strapparava, C., & Valitutti, R. (2004). WordNet-Affect: an Affective Extension of WordNet. In *4th International Conference on Language Resources and Evaluation* (Vol. 2004, pp. 1083–1086).

Streich, S. (2007). *Music Complexity a multi-faceted description of audio content*. Universitat Pompeu Fabra. Retrieved from http://mtg.upf.edu/node/507

Streich, S., & Herrera, P. (2005). Detrended Fluctuation Analysis of Music Signals Danceability Estimation and further Semantic Characterization. In *118th Audio Engineering Society Convention – AES 118*. Barcelona, Spain.

Swain, J. P. (2002). *Harmonic Rhythm: Analysis and Interpretation*. Oxford, UK: Oxford University Press.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review.

In C. C. Aggarwal (Ed.), *Data Classification: Algorithms and Applications* (1st ed., pp. 37–64). CRC Press.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. http://doi.org/10.5116/ijme.4dfb.8dfd

TechWhack.com. (2005, January 25). Apple iTunes touches an impressive 250 Million sales figure. *TechWhack.Com*. Retrieved from https://web.archive.org/web/20080830070628/http://news.techwhack.com/690/apple-itunes-250-million/

Tellegen, A., Watson, D., & Clark, L. A. (1999). On the Dimensional and Hierarchical Structure of Affect. *Psychological Science*, *10*(4), 297–303. http://doi.org/10.1111/1467-9280.00157

Temperley, D. (1999). What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered. *Music Perception: An Interdisciplinary Journal*, *17*(1), 65–100. http://doi.org/10.2307/40285812

Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290*(5500), 2319–2323. http://doi.org/10.1126/science.290.5500.2319

Terhardt, E., Stoll, G., & Seewann, M. (1982). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, *71*(3), 679–688. http://doi.org/10.1121/1.387544

Thammasan, N., Fukui, K., & Numao, M. (2017). Multimodal Fusion of EEG and Musical Features Music-emotion Recognition. In *31st AAAI Conference on Artificial Intelligence – AAAI 2017* (pp. 4991–4992). San Francisco, California, USA. Retrieved from http://arxiv.org/abs/1611.10120

Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal*. Oxford University Press, USA.

Thompson, W. F., & Robitaille, B. (1992). Can Composers Express Emotions through Music? *Empirical Studies of the Arts*, *10*(1), 79–89. http://doi.org/10.2190/NBNY-AKDK-GW58-MTEL

Tillmann, B., & Bigand, E. (1996). Does Formal Musical Structure Affect Perception of Musical Expressiveness? *Psychology of Music*, *24*(1), 3–17. http://doi.org/10.1177/0305735696241002

Timmers, R., & Ashley, R. (2007). Emotional Ornamentation in Performances of a Handel Sonata. *Music Perception: An Interdisciplinary Journal*, *25*(2), 117–134. http://doi.org/10.1525/mp.2007.25.2.117

Toh, A. M., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. In *Computer Engineering* (pp. 22–25).

Toiviainen, P., & Krumhansl, C. L. (2003). Measuring and Modeling Real-Time Responses to Music: The Dynamics of Tonality Induction. *Perception*, *32*(6), 741–766. http://doi.org/10.1068/p3312

Tolos, M., Tato, R., & Kemp, T. (2005). Mood-based navigation through large collections of musical data. In *2nd IEEE Consumer Communications and Networking Conference – CCNC 2005* (pp. 71–75). Las Vegas, NV, USA: IEEE. http://doi.org/10.1109/CCNC.2005.1405146

Tomkins, S. S. (1962). *Affect, Imagery, Consciousness - Volume I: The Positive Affects*. Springer Publishing Company.

Tomkins, S. S. (1963). *Affect, Imagery, Consciousness - Volume II: The Negative Affects*. Springer Publishing Company.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multi-Label Classification of Music Into Emotions. In *9th International Society of Music Information Retrieval Conference – ISMIR 2008* (pp. 325–330). Philadelphia, Pennsylvania, USA.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2011). Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, *2011*(1), 4. http://doi.org/10.1186/1687-4722-2011-426793

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2007). Towards musical query-by-semantic-description using the CAL500 data set. In *30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* (p. 439). New York, New York, USA: ACM Press. http://doi.org/10.1145/1277741.1277817

Tzanetakis, G. (2002). *Manipulation, Analysis and Retrieval Systems for Audio Signals*. Princeton University.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing – TSA*, *10*(5), 293–302. http://doi.org/10.1109/TSA.2002.800560

Tzanetakis, G., & Percival, G. (2013). An effective, simple tempo estimation method based on self-similarity and regularity. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 241–245). IEEE. http://doi.org/10.1109/ICASSP.2013.6637645

Vakkuri, A., Yli-Hankala, A., Talja, P., Mustola, S., Tolvanen-Laakso, H., Sampson, T., & Viertio-Oja, H. (2004). Time-frequency balanced spectral entropy as a measure of anesthetic drug effect in central nervous system during sevoflurane, propofol, and thiopental anesthesia. *Acta Anaesthesiologica Scandinavica*, *48*(2), 145–153. http://doi.org/10.1111/j.0001-5172.2004.00323.x

Vale, P. (2017). *The Role of Artist and Genre on Music Emotion Recognition*. Universidade

Nova de Lisboa.

Variety. (2018, May 15). Apple Music Passes 50 Million Subscribers, Including Free Trials. *Variety*. Retrieved from https://variety.com/2018/biz/news/apple-music-passes-50-million-subscribers-including-free-trials-1202811061/

Vickers, E. (2001). Automatic Long-term Loudness and Dynamics Matching. In *111th Audio Engineering Society Convention – AES 111*.

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, *22*(4), 720–752. http://doi.org/10.1080/02699930701503567

Wallin, N. L., Merker, B., & Brown, S. (1999). *The origins of music*. MIT Press.

Wang, J.-C., Lo, H.-Y., Jeng, S.-K., & Wang, H.-M. (2010). Mirex 2010: Audio Classification Using Semantic Transformation And Classifier Ensemble. In *6th Music Information Retrieval Exchange – MIREX 2010, as part of the 11th International Society for Music Information Retrieval Conference – ISMIR 2010* (pp. 2–5). Utrecht, Netherlands.

Wang, J.-C., Yang, Y.-H., Chang, K., Wang, H.-M., & Jeng, S.-K. (2012). Exploring the relationship between categorical and dimensional emotion semantics of music. In *2nd International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies - MIRUM '12* (pp. 63–68). New York, NY, USA: ACM Press. http://doi.org/10.1145/2390848.2390865

Wang, J.-C., Yang, Y.-H., Jhuo, I.-H., Lin, Y.-Y., & Wang, H.-M. (2012). The acousticvisual emotion guassians model for automatic generation of music video. In *20th ACM international conference on Multimedia - MM '12* (p. 1379). Nara, Japan: ACM Press. http://doi.org/10.1145/2393347.2396494

Wang, J.-C., Yang, Y.-H., Wang, H.-M., & Jeng, S.-K. (2012). The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In *20th ACM International Conference on Multimedia – ACM MM 2012* (p. 89). Nara, Japan: ACM Press. http://doi.org/10.1145/2393347.2393367

Wang, M., Zhang, N., & Zhu, H. (2004). User-adaptive music emotion recognition. In *7th International Conference on Signal Processing – ICSP 2004* (Vol. 2, pp. 1352–1355). Beijing, China: IEEE. http://doi.org/10.1109/ICOSP.2004.1441576

Wang, S.-Y., Wang, J.-C., Yang, Y.-H., & Wang, H.-M. (2014). Towards time-varying music auto-tagging based on CAL500 expansion. In *2014 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). Chengdu, China: IEEE. http://doi.org/10.1109/ICME.2014.6890290

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. http://doi.org/10.3758/s13428-012-0314-x

Watson, D., & Clark, L. A. (1992). On Traits and Temperament: General and Specific Factors of Emotional Experience and Their Relation to the Five-Factor Model. *Journal of Personality*, *60*(2), 441–476. http://doi.org/10.1111/j.1467-6494.1992.tb00980.x

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form. *Iowa Research Online*. Iowa, USA: The University of Iowa. Retrieved from http://ir.uiowa.edu/psychology_pubs/11/

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–70.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235. http://doi.org/10.1037/0033-2909.98.2.219

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, *76*(5), 820–838. http://doi.org/10.1037/0022-3514.76.5.820

Watson, K. B. (1942). The nature and measurement of musical meanings. *Psychological Monographs*, *54*(2), i-43. http://doi.org/10.1037/h0093496

Webster, G. D., & Weir, C. G. (2005). Emotional Responses to Music: Interactive Effects of Mode, Texture, and Tempo. *Motivation and Emotion*, *29*(1), 19–39. http://doi.org/10.1007/s11031-005-4414-0

Wedin, L. (1972). A Multidimensional Study of Perceptual-Emotional Qualities in MusicIC. *Scandinavian Journal of Psychology*, *13*(1), 241–257. http://doi.org/10.1111/j.1467-9450.1972.tb00072.x

Westfall, P. H. (2014). Kurtosis as Peakedness, 1905–2014. R.I.P. *The American Statistician*, *68*(3), 191–195. http://doi.org/10.1080/00031305.2014.917055

Whissell, C., Fournier, M., Pelland, R., Weir, D., & Makarec, K. (1986). A Dictionary of Affect in Language: IV. Reliability, Validity, and Applications. *Perceptual and Motor Skills*, *62*(3), 875–888. http://doi.org/10.2466/pms.1986.62.3.875

Wieczorkowska, A. A. (2004). Towards Extracting Emotions from Music. In *Intelligent Media Technology for Communicative Intelligence – IMTCI 2004* (pp. 228–238). Warsaw, Poland. http://doi.org/10.1007/11558637_23

Wieczorkowska, A. A., Synak, P., Lewis, R., & Raś, Z. W. (2005). Extracting emotions from music data. In *International Symposium on Methodologies for Intelligent Systems 2005: Foundations of Intelligent Systems* (pp. 456–465). http://doi.org/10.1007/11425274_47

Wieczorkowska, A. A., Synak, P., & Raś, Z. W. (2006). Multi-label classification of

emotions in music. In *Intelligent Information Processing and Web Mining – IIPWM 2006* (pp. 307–315). Ustrón, Poland. http://doi.org/10.1007/3-540-33521-8_30

Wu, B., Horner, A., & Lee, C. (2014a). Musical timbre and emotion: The identification of salient timbral features in sustained musical instrument tones equalized in attack time and spectral centroid. In *40th International Computer Music Conference – ICMC 2014, Joint with the 11th Sound and Music Computing Conference – SMC 2014* (p. 928). Athens, Greece.

Wu, B., Horner, A., & Lee, C. (2014b). The Correspondence of Music Emotion and Timbre in Sustained Musical Instrument Sounds. *Journal of the Audio Engineering Society – JAES*, *62*(10), 663–675. http://doi.org/10.17743/jaes.2014.0037

Wu, B., Zhong, E., Horner, A., & Yang, Q. (2014). Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning. In *22th ACM International Conference on Multimedia – ACM MM 2014* (pp. 117–126). Orlando, Florida, USA. http://doi.org/10.1145/2647868.2654904

Wu, T.-L., & Jeng, S.-K. (2006). Automatic emotion classification of musical segments. In *9th International Conference on Music Perception & Cognition – ICMPC 2006*. Bologna, Italy.

Wu, T.-L., & Jeng, S.-K. (2008). Probabilistic Estimation of a Novel Music Emotion Model. In *14th International Multimedia Modeling Conference* (pp. 487–497). Kyoto, Japan: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-77409-9_46

Wu, W., & Xie, L. (2008). Discriminating Mood Taxonomy of Chinese Traditional Music and Western Classical Music with Content Feature Sets. In *2008 Congress on Image and Signal Processing - CISP 2008* (pp. 148–152). Sanya, Hainan, China: IEEE. http://doi.org/10.1109/CISP.2008.272

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2008). What is the best segment duration for music mood analysis? In *2008 International Workshop on Content-Based Multimedia Indexing* (pp. 17–24). IEEE. http://doi.org/10.1109/CBMI.2008.4564922

Xu, J., Li, X., Hao, Y., & Yang, G. (2014). Source Separation Improves Music Emotion Recognition. In *International Conference on Multimedia Retrieval – ICMR 2014* (pp. 423–426). Glasgow, United Kingdom: ACM Press. http://doi.org/10.1145/2578726.2578784

Yang, D., & Lee, W. (2004). Disambiguating Music Emotion Using Software Agents. In *5th International Conference on Music Information Retrieval – ISMIR 2004* (pp. 52–58). Barcelona, Spain.

Yang, L., Chew, E., & Rajab, K. Z. (2013). Vibrato Performance Style: A Case Study Comparing Erhu and Violin. In *10th International Symposium on Computer Music Multidisciplinary Research – CMMR 2013* (pp. 904–919). Marseille, France.

Yang, X., Dong, Y., & Li, J. (2017). Review of data features-based music emotion recognition methods. *Multimedia Systems*, 1–25. http://doi.org/10.1007/s00530-017-0559-4

Yang, Y.-H., & Chen, H. H. (2011a). *Music Emotion Recognition*. CRC Press.

Yang, Y.-H., & Chen, H. H. (2011b). Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Transactions on Audio, Speech, and Language Processing – TASLP 2011*, *19*(4), 762–774. http://doi.org/10.1109/TASL.2010.2064164

Yang, Y.-H., & Chen, H. H. (2012). Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology – TIST*, *3*(3), 1–30. http://doi.org/10.1145/2168752.2168754

Yang, Y.-H., & Hu, X. (2012). Cross-Cultural Music Mood Classification: A Comparison On English And Chinese Songs. In *13th International Society for Music Information Retrieval Conference – ISMIR 2012* (pp. 19–24). Porto, Portugal.

Yang, Y.-H., Lin, Y.-C., & Chen, H. H. (2009). Personalized music emotion recognition. In *32nd international ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '09* (pp. 748–749). Boston, MA, USA: ACM Press. http://doi.org/10.1145/1571941.1572109

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., & Chen, H. H. (2008). Toward multi-modal music emotion classification. In *Pacific-Rim Conference on Multimedia* (Vol. 5353, pp. 70–79). Springer.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2007). Music Emotion Classification: A Regression Approach. In *Multimedia and Expo, 2007 IEEE International Conference on* (pp. 208–211). Beijing, China: IEEE. http://doi.org/10.1109/ICME.2007.4284623

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2008). A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, *16*(2), 448–457. http://doi.org/10.1109/TASL.2007.911513

Yang, Y.-H., Liu, C.-C., & Chen, H. H. (2006). Music emotion classification: a fuzzy approach. In *14th Annual ACM International Conference on Multimedia – ACM MM 2006* (pp. 81–84). Santa Barbara, CA, USA: ACM.

Yang, Y.-H., Su, Y.-F., Lin, Y.-C., & Chen, H. H. (2007). Music Emotion Recognition: The Role of Individuality. In *ACM SIGMM International Workshop on Human-centered Multimedia – HCM 2007, in conjunction with ACM Multimedia* (pp. 13–22). Augsburg, Germany: ACM. http://doi.org/10.1145/1290128.1290132

Yeh, C.-C., Tseng, S.-S., Tsai, P.-C., & Weng, J.-F. (2006). Building a Personalized Music Emotion Prediction System. In *7th Pacific Rim Conference on Multimedia - PCM2006* (pp. 730–739). Hangzhou, China: Springer, Berlin, Heidelberg.

http://doi.org/10.1007/11922162_84

Zapata, J., Davies, M. E. P., & Gómez, E. (2014). Multi-Feature Beat Tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing – TASLP*, *22*(4), 816–825. http://doi.org/10.1109/TASLP.2014.2305252

Zentner, M., & Eerola, T. (2010). Self-report measures and models. In *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 187–221). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199230143.003.0008

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8*(4), 494–521. http://doi.org/10.1037/1528-3542.8.4.494

Zhang, J. L., Huang, X. L., Yang, L. F., Xu, Y., & Sun, S. T. (2017). Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods. *Multimedia Systems*, *23*(2), 251–264. http://doi.org/10.1007/s00530-015-0489-y

Zhao, Z., & Liu, H. (2007). Semi-supervised Feature Selection via Spectral Analysis. In *2007 SIAM International Conference on Data Mining* (pp. 641–646). Philadelphia, PA: Society for Industrial and Applied Mathematics. http://doi.org/10.1137/1.9781611972771.75

Zheng, F., Song, Z., Li, L., Yu, W., & Wu, W. (1998). The distance measure for line spectrum pairs applied to speech recognition. In *5th International Conference on Spoken Language Processing 1998 – ICSLP 1998* (pp. 1123–1126). Sydney, Australia.

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, *33*(2), 248–248. http://doi.org/10.1121/1.1908630

APPENDIX A

MUSICAL DIMENSIONS ANALYSIS

A condensed version of the existing musical dimensions was presented in Section 2.3. This section extends it with additional details on the subject.

A.1. Melody

Melody can be defined as a horizontal succession of pitches (perceptual correlate of fundamental frequency) or musical tones, perceived by listeners as a single musical line. For a condensed introduction to melody see Section 2.3.1.

Melodic Arrangement

Melodies are usually composed of several musical phrases, which are then repeated and arranged to compose the musical piece. The melodic arrangement describes this organization of melodies in the song. The most basic arrangement in a musical piece is to have one melody followed by another after that (one after another). It is also possible to have multiple melodies at the same type. One of this cases is the counter-melody, where a secondary melody is played at the same time as the primary, without clashing musically.

Melodic Movement and Contour

The arranged melody contains movement – the direction followed by the patterns of notes that compose it, which on its core are sequences of higher, lower or the same pitch. This movement can be smooth, with transitions between notes close in the scale, or rough, with longer intervals, sequences of the same notes or ascending and descending patterns, originating a specific shape or contour. Regarding the movement (or motion), several distinct types exist. Namely, stepwise motion based on small steps (e.g., using major 2^{nd} and minor 2^{nd} intervals), intervallic leaps or skips (e.g., series of repeating larger intervals) or stepwise and skipwise leaps.

Melodic contours are the shapes created by the melody of a musical piece, as illustrated in Figure A.1. Achieving a specific melodic contour is not usually an objective the composer aims for. Nonetheless, they can be analyzed and compared when discussing the melody of songs.

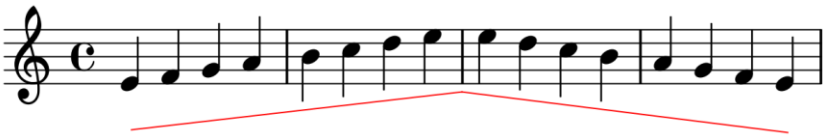


Figure A.1: Example of a melodic contour which can be described as an arch, ascending and then descending.

The typology of melodic contours was studied by Charles Adams (1976). In his work, the author defines melodic contours "as the product of distinctive relationships among the minimal boundaries of a melodic segment", where melodic segments are "any series of differentiated pitches", while minimal boundaries refer to the "pitches which are considered necessary and sufficient to delineate a melodic segment, with respect to its temporal aspect (beginning-end) and its tonal aspect (tonal range)". (Adams, 1976, p. 196).

As a result, Adams defines three features that are the basis to classify the 15 melodic contour shapes shown in Figure A.2, which are: 1) the slope of the contour (S); 2) a deviation (change of direction) in the slope of the contour (D); and 3) the reciprocal (or inverse) of deviation in the slope of the contour (R).

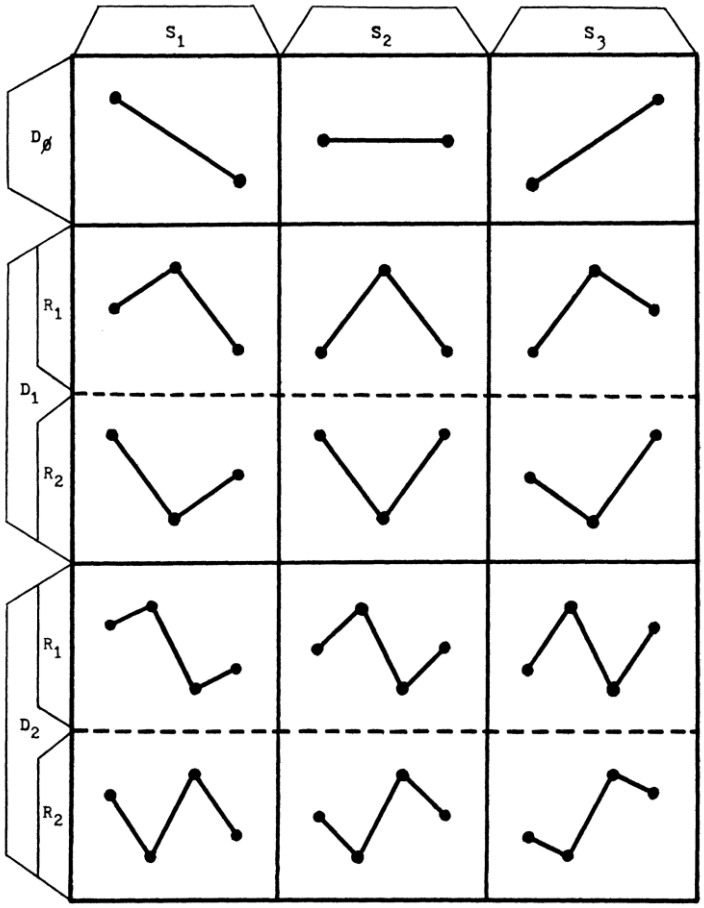


Figure A.2: The 15 melodic contour types according to Adams (Adams, 1976, p. 199).

Pitch

Pitch is a perceptual property of sounds which makes it possible to rate them as higher or lower in terms of frequency (Plack & Oxenham, 2005, p. 2). Pitch is related with frequency but not a direct synonym, rather a subjective psychoacoustical attribute of sound (Hartmann, 1997, pp. 145, 284, 287).

An instrument or voice has a definite pitch when its sound is tuned and distinct, such as the singing voice. On the other hand, a sound with indefinite pitch is an untuned sound such as the speaking voice; it is harder to identify and notate since it does not follow specific notes.

Pitch Range

The pitch range of a musical instrument is the interval of notes (or distance in terms of pitch), from higher to lowest, that such instrument can play. As an example, a tuba has a range typically from D1 to F4, while a flute (in C) from C4 to D7 and a piano from A0 to C8 (Read, 1953, pp. 11–28).

When describing a musical piece and its melody, the pitch range is the distance between its lowest and highest note. This range can be narrow or wide, as illustrated in Figure A.3.

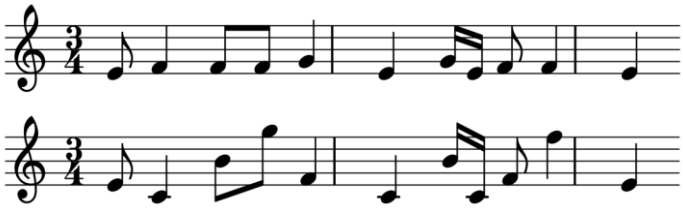


Figure A.3: Examples of different pitch ranges: narrow range in the above musical score and wide below.

Register

The register is the relative "height" of a sound, note, melody or instrument (including the human voice). Different instruments have different registers, which can be roughly classified from high to low, where higher registers indicate a higher pitch. The combination of different instruments and their ranges in a musical piece enriches it by increasing contrast and variety.

As for the human voice, its register can go from high, attained by whistling or singing in falsetto to vocal fry register (creaky voice), the lowest register possible. The singing voice is normally classified into distinct voice types, from Soprano to Bass as presented in Table A.1.

Melodic Embellishments

Several features are also used to embellish the melodies in a musical piece. This is often accomplished by using the ornamentation techniques previously described as part of expressive techniques (e.g., trills, glissandos or mordents).

Moreover, other features are also used to connect melodies. As an example, riffs and ostinatos consist in repeated melodic patterns throughout the piece. Other melodic fragments are sometimes used to unify different parts of the musical piece, namely motifs,

sequences and repetitions, each with subtle differences. Such embellishments are created using characteristics similar (or equal) to the ones described in other dimensions (e.g., using ornamentations, described Sections 2.3.6 and A.6).

Category	*Sex*	*Classical* (McKinney, 1994)	*Non-classical* (Peckham, 2005)
Soprano	Female	C4-C6	C4-C6
Mezzo-soprano	Female	A3-A5	A3-A5
Contralto	Female	F3-F5	F3-E5
Countertenor	Male	E3-E5	-
Tenor	Male	C3-C5	B2-A4
Baritone	Male	A2-A4	G2-F4
Bass	Male	E2-E4	E2-E4

Table A.1: Different voice types for classical and non-classical singers.

A.2. Harmony

If melody is said to be the horizontal part of music, harmony refers to its "vertical" aspect. That is, the sound produced by the combination of various pitches (notes or tones) in chords. For a condensed introduction to harmony see Section 2.3.2.

Harmonic Rhythm or Tempo

Harmonic rhythm is the rate at which the chords change (how fast the harmony moves), in a musical composition, in relation to the rate of notes. That is, the "perception of rhythm that depends on changes in aspects of harmony" (Swain, 2002, p. 4). A song may have a slow or fast harmonic rhythm, which may not be directly related with its melodic rhythm. As an example, a musical piece with chords changing only every measure, but with several notes per measure will have a slow harmonic rhythm and fast melodic rhythm.

Harmonic Progression

Harmonic (or chord) progressions are series of musical chords, serving as the foundation of harmony in western music. These progressions are typically expressed using either roman numerals or the name of the chords (e.g., iii, IV, V, in the key of C representing

E minor, F major, G major).

The most commonly used chords in western music are called triads, which contain three notes:

- the root note
- the third – a second note with an interval (the difference in pitch between the two notes) of a third above the root (minor third being 3 semitones and major third being four semitones)
- the fifth – a third note, in this case at an interval of a third from the second note

Simple chord progressions are usually based on sequences of these triads and present in all types of music. Other progressions, using chords with other seventh and extended chords (triad chords with related notes added on top) are often used in complex progressions, namely in Jazz music.

Modulation

Modulation consists in altering the key (or key center) in a musical piece, i.e., the tonic key of the song. The modulation to a new key can be used to create interest in the song and thus is often used during the climax of a song.

Harmonic Perception

Harmony is based on consonance, often described in terms of its relative harshness. Several reasons contribute to create a consonant chord: the lack of perceptual roughness, which happens when the frequency components of the sound (partials) fall outside a specific bandwidth related to the human ear's ability to distinguish frequencies; when the chord spectrum is similar to a harmonic series (the spectrum of each sound is a multiple of the lowest frequency), creating a perceptual fusion of the chord; and finally, the familiarity, with the chords heard more frequently sounding more consonant.

While consonant sounds are pleasing, due to the abovementioned reasons, dissonant sounds are heard as a "clash" of notes, resultant for instance of a harsh-sounding harmonic combination. Such dissonant sounds are sometimes used (e.g., in contemporary art music) to create "tension" in a song, which is often "released" by changing to consonant chords.

The perception of dissonant sounds as unpleasant is mostly found in Western music and listeners, where diatonic scales, major and minor tonalities dominate. The composition of music using harmonies arranged in progressions is "one of the major differences between Western and much non-Western music" (Malm, 1995, p. 15).

Similarly to the previous dimensions, composers can add features related with harmony to enrich the musical piece, namely, the repetition of chordal patterns (e.g., harmonic ostinatos and riffs) or sustaining a specific note (e.g., drone effect).

A.3. Rhythm

Rhythm represents the element of "time" in music, the patterns of long and short sounds and silences found in music. For a condensed introduction to rhythm see Section 2.3.3.

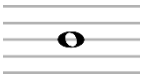
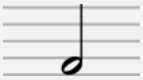
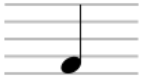
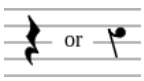
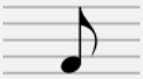
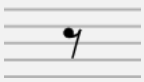
Rhythm Types

As the name implies, rhythm types describes how the rhythm "is". Typically it can be characterized with words such as simple or complex, but also as regular or irregular.

As an example, a steady beat at a constant interval is said to have a "regular rhythm". Such a feature is commonly found in many of the mainstream music and helps the audience to anticipate and remember the song. Irregular rhythm on the other hand is often used for expression, breaking from the expected pattern.

Note Values and Rests

A note value in music indicates its length or duration. These note values are not absolute, but relative to each other, either longer or shorter. The rhythm of a musical piece is constructed by the arrangements of the various notes and silence values. Table A.2 shows the representation of the most common notes and silence values.

	English Name (American)	*Value*	*Note*	*Rest*
Long	Breve (double note)	8 beats		
Long	Semibreve (whole note)	4 beats		
Long	Minim (half note)	2 beats		
Short	Crotchet (quarter note)	1 beat		
Short	Quaver (eighth note)	½ beat		

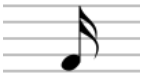
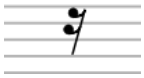
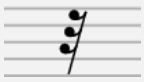
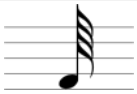
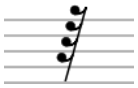
Short	Semiquaver (sixteenth note)	¼ beat		
Short	Demisemiquaver (thirty-second note)	⅛ beat		
Short	Hemidemisemiquaver (sixty-fourth note)	$^1/_{16}$ beat		

Table A.2: Common notes' and rests' values and symbols.

Rhythmic Devices

In music, rhythmic devices are characteristics used in musical pieces that define their shape. Several distinct rhythmic devices exist, some of which are presented in Table A.3.

Rhythmic device	*Description*
Ostinato	A constantly repeated musical phrase in the same instrument or voice.
Syncopation	An interruption of the general flow of the rhythm by a beat that falls in an unexpected place.
Accent	Accentuation (or stress) of particular notes for expression.
Polyrhythm	Two or more rhythms with distinct pulses played simultaneously.
Cross rhythms	Two or more rhythmic patterns with conflicting metres played simultaneously.
Hemiola	Specific syncopation pattern with two beats played in the time of three (or the opposite).
Riff	A repeated (melodic or chordal) pattern that is usually a few bars in length and typically observed in jazz or rock genres.
Repetition	Reappearance of a specific pattern during the course of the piece of music, serving as a unifying feature.
Diminution	Repetition of a pattern where half duration notes are used.
Augmentation	Repetition of a pattern where double duration notes are used.
Rock beat	Typically stressed drum pattern, very common in pop rock songs.

Table A.3: Examples of some known rhythmic devices and their descriptions.

Some rhythmic devices are more common in specific genres and thus often help

defining them. As an example, rock and jazz usually contain ostinato, riff, repetition, syncopation, accents, rock beat and others. Composers up to the baroque commonly used *hemiola*, repetition and accents. Still, many devices such as repetition, accents, cross rhythms, polyrhythms, diminution and augmentation are common in most genres.

Rhythmic Layers

Rhythmic layers consist of groups of different "performing media" such as instrumental or vocals. One of the first steps in music analysis starts by identifying the number of rhythmic layers, followed by the differentiation of the instruments present in each one.

Duration

The duration, or the amount of time, a specific sound or silence lasts (e.g., how long a note, phrase or composition is. "Duration is the length of time a pitch, or tone, is sounded" (Benward & Saker, 2008, p. xiv).

Bar

Also known as measure, the bar is a division of music. It represents a segment of time corresponding to a specific number of beats, where each beat is represented by a specific note value. In music notation, the bar limits are indicated by bar lines and the number of beats is normally specified at the beginning of the score by the time signature (see metre definition, below).

Beat

The beat is the underlying steady pulse (regularly repeating event) in a piece of music. It is the basic unit of time of the mensural or beat level (Berry, 1976, p. 349).

Metre

As Benward et al. defines it, metre is "a regular, recurring pattern of strong and weak beats. This recurring pattern of durations is identified at the beginning of a composition by a meter signature (time signature)." (Benward & Saker, 2008, p. 10). The time signature is normally represented as two numbers, one above the other. The top number indicates the number of beats in a bar, while the bottom one is an indication of what is considered a beat (based on note values indicated in Table A.2). For instance, a time signature of 3/4, indicates a metre of three beats per bar, divided using crotchet beats (based on the American name – quarter note).

The most common metre types are: duple metre, with two pulses per bar, where the

first beat is accented (e.g., time signatures such as 2/2 or 2/4); triple metre, with three pulses and accentuation on the first beat (e.g., 3/4); and quadruple metre, with four pulses per bar where the first and third beats are accented (e.g., 4/4). Time signatures can be either simple or compound, depending on the division of beats per bar. As an example, a time signature of 3/4 is considered a simple triple. "Triple" refers to the metre as having three beats (pulses) per bar, while "simple" states that each of these beats can be divided into two notes. On the other hand, beats in a compound meter are divided into three notes, always using a dotted note as its beat. As an example, a time signature of 6/8 is considered compound duple, meaning it has two beats per bar, and each of these beats can be divided in three notes. Both cases are illustrated in Figure A.4. Typically, compound time signatures will have 6 (duple compound), 9 (triple compound) and 12 (quadruple compound) as the top number. A musical piece may also have mixed meter, where different time signatures are present or no meter at all.

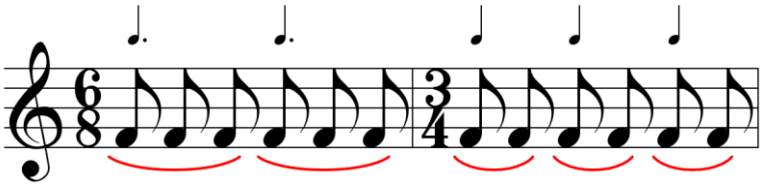


Figure A.4: Mixed meter showing compound versus simple time signatures[133].

Tempo

Tempo represents the speed of the beat, and thus the speed or pace of a musical piece. It is described in several different ways such as with a range of words (e.g., "fast", "adagio", "slowly" or becoming faster or slower). In classical music it is often designated by Italian terms such as *largo*, *adagio*, *andante*, *moderato*, *allegro* or *presto*, all relative to each other, indicated at the beginning of the musical piece. Tempo is usually measured in beats per minute (bpm).

A.4. Dynamics

Dynamics represents the variation in loudness or softness of notes in a musical piece. For a condensed introduction to dynamics see Section 2.3.4.

[133] Image source: lesson 15 (Simple and Compound Meter) from musictheory.net

Dynamic levels

The different volume levels in a musical piece are called its dynamic levels. When writing a song, the composer inserts dynamic markings in a music score to inform the performer how a specific passage should be played (e.g., louder). These markings are typically Italian terms that represent varying degrees from very soft to very loud. The two basic indications are *forte*, represented with an *f*, meaning "strong" and *piano*, *p*, for "soft". Steps between these two levels can be signaled with the word *mezzo*, meaning "half" (e.g., *mezzo-forte*). Extremer values are indicated by multiple *f* or *p* values, such as *fff* for *fortississimo*. Table A.4 summarizes some of these values.

Name	*Letters*	*Level*
fortississimo	*fff*	very very loud
fortissimo	*ff*	very loud
forte	*f*	loud
mezzo-forte	*mf*	semi-loud
mezzo-piano	*mp*	semi-soft
piano	*p*	soft
pianissimo	*pp*	very soft
pianississimo	*ppp*	very very soft

Table A.4: Scale of common dynamic markings.

Accents and changes in dynamic levels

In addition to changing the dynamics for a specific part using the abovementioned terms, it is also common to have progressive subtle changes to louder or softer levels. These changes are usually named *crescendo*, for a gradual increase in loudness and *decrescendo*, for the inverse (gradually getting softer).

It is also possible to change the dynamic level of a single note or sound instead of a larger section. This is accomplished by using accents such as *sforzando*, meaning forced, played with sudden emphasis. Table A.5 presents some of the available markings, which are then illustrated in the musical score in Figure A.5.

Name	*Symbol*	*Description*
Crescendo	cresc. or 	Increasing in loudness
Decrescendo	decres. or 	Decreasing / getting softer
Sforzando, sforzato, forzando	sfz, sf or fz	Sudden emphasis

or forzato		
Subito	sub.	Suddently

Table A.5: Most common accents and dynamic level changes.

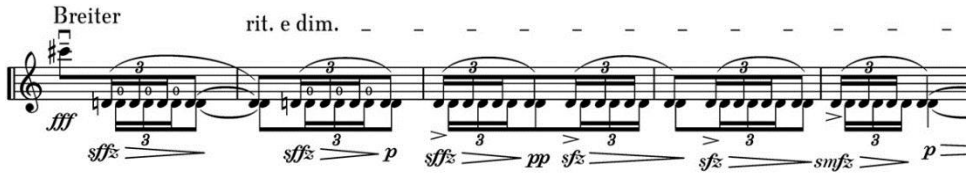


Figure A.5: Passage of String Quartet op. 3 (2nd movement of violin II) by Alban Berg, containing *sforzatos* (sfz) and *sforzattissimos* (sffz), as well as decrescendos and accent marks in specific notes.

It can be argued that elements such as accents can be classified as articulation mechanisms and thus should be placed in the expression techniques dimension (see Section 2.3.6). As stressed before, many of the musical attributes touch several musical categories and thus this organization reflects our view.

A.5. Tone Color or Timbre

Tone color, also known as timbre, refers to the perceived sound quality (properties) of a sound (e.g., a musical note). For a condensed introduction to tone color see Section 2.3.5.

Sound Envelope

Sound is the result of moving air (or other medium) particles. More precisely, playing an instrument (such as hitting a drum) generates vibration which creates waves of air particles that travel from the sound source through the air to the listener's eardrum.

Two key components make the sound wave: 1) frequency – related to the length of the wave, or in other words, the number of cycles it completes in a second; and 2) amplitude – related to the height of the sound wave. The frequency defines the sound's pitch, which can be low frequency, caused by slower vibrations, and high pitch, with shorter, faster vibrations. The amplitude is related to the sound's volume, as in a sound being classified from quiet to loud. Volume is largely a subjective term, related with level, which is an objective measure of sound pressure – the difference from normal air

pressure caused by the sound travelling in the air.

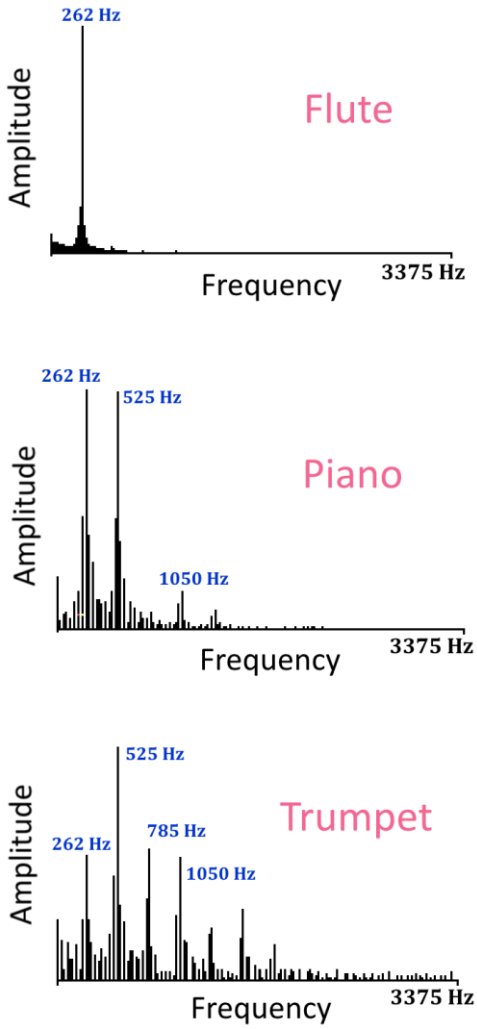


Figure A.6: Spectra of middle C played on a flute, piano and trumpet (D. Davis, 2002).

Sounds are defined by their pitch. For example the A4 musical note has a frequency of 440Hz, or 440 oscillations per second. However, such pure sounds are synthetic, and are not usually found in the natural world. The same note produced by a musical instrument is richer, made of a sum of distinct frequencies. The lowest frequency produced is normally the fundamental frequency, used to identify the note. This frequency is many times, but not always, the frequency heard by the listener as dominant. Above it we have overtones, frequencies that are greater than the fundamental, some of which are harmonics. The harmonics are series of frequencies which have frequencies multiple of the

fundamental. As previously noted, this harmonic series forms a spectral envelope which varies across instruments and helps distinguish them, by influencing the tone color, as illustrated in Figure A.6.

Waveform Envelope

The second aspect influencing tone color is the envelope of the sound waveform, which can be described as the overall amplitude structure of the sound – the "envelope" in which the sound wave fits, as illustrated in Figure A.7 in red. The envelope comprises four stages: the attack, decay, sustain and release, thus named ADSR envelope.

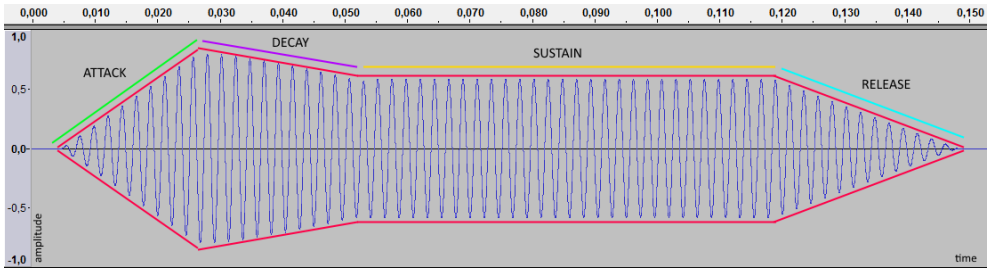


Figure A.7: Representation of the ADSR envelope, in red.

The tone color of a sound is influenced by three key factors: the material of the instrument or voice; the techniques employed in producing the sound; and the layers of sound and the effects the sound has on the music.

Instrument Materials

The materials, shape and dimensions of a musical instrument have great implications in the produced sound, changing its tone color. Since the beginning of humankind, humans have used any possible material found in nature to create music, from wood or stones to animal parts such as shells, bones and skin (Rault, 2000, p. 9).

Most instruments can be organized into six groups according to their built material, as represented in Table A.6, within which certain sound similarities are present: electronic, metal, skin, string, wood and the vocal tract (not a material in the sense of the remaining five).

Material	*Description / examples*
Electronic	Sound is produced using electronic circuits, such as the ones in

	a musical keyboard and other musical controllers and synthesizers.
Metal	Instruments made of metal include (almost) all brass instruments such as trumpets, horns, trombones or tubas. Many other instruments in different families use metal, such as bells, xylophones, cymbals or even the musical triangle.
Skin	Skin is mostly associated with percussion instruments such as drums, tympani, djembes or tambourines.
String	Sound is produced by vibrating strings. Some well-known examples are guitars, harps, pianos and violins.
Wood	Woodwind instruments are a subgroup of instruments from the wind instruments, which includes wood flutes, pan flutes or bagpipes. Wood is not exclusive to wind instruments; other examples are percussion instruments, such as wooden xylophones, claves, musical wood blocks and others.
Vocals	All the sounds generated by the vocal tract.

Table A.6: Different materials used to produce sound and music.

Many musical instruments are complex, made of multiple materials from distinct groups. As an example, a drum can be made of skin, wood and metal, while a violin also uses wood and metal in addition to strings. Still, in each of them the sound is produced and primarily influenced by one of the materials.

Playing Methods

The sound produced by a musical instrument is not only influenced by its material. Other factors, such as the method used to produce the sound are also important. The sound produced by a guitar, violin or piano is generated by the vibration of its strings, however, the three use distinct playing methods. While guitar strings are plucked by the guitar player, a violinist uses a bow, scrapping the violin's strings. A piano player, on the other hand, presses a key, which drives a hammer to hit one of the piano strings.

The methods used to produce sound from instruments can be grouped as follows:

- hitting (e.g., drums)
- blowing (e.g., flute)
- shaking (e.g., maracas)
- scraping (e.g., violin)
- plucking (e.g., guitar)

Instruments and Voices' Types

The source of a sound can be a musical instrument, a voice and also an unconventional source (whistling, traffic, and so on). Regarding musical instruments, several distinct classification schemes have been devised over the centuries (Kartomi, 1990), based on their materials and more recently on the sound production methods.

Nowadays, one of the reigning classification systems is the western classification, applying mostly to musical instruments of western tradition. Western musical instruments can be broadly classified into three groups:

- Percussion (e.g., drums)
- Strings (e.g., violin)
- Wind (e.g., flute)

These categories are sometimes further divided into subgroups, for instance dividing strings according to the playing method into plucked strings or bowed strings. A common example is the differentiation made in classical music of wind instruments into either: woodwind instruments, for those containing a reed (mouthpiece) to produce the sound (e.g., flute); or brass instruments, where the air is set in motion directly by the lips of the performer (e.g., trumpet).

Some instruments are hard to properly classify into these three groups (e.g., piano and some non-western traditional instruments). As an example, the piano contains strings, which are struck by hammers, thus opening the debate of whether it should be considered a percussion or strings instrument.

Both instruments and voices can also be further classified by their musical ranges when compared to other instruments in the same family, as previously described in the melody/register section. As an example, saxophones can be classified into several groups such as soprano, alto, tenor, baritone or bass saxophones.

Combinations and Types of Sounds

In music, sounds can be divided into acoustic or electronic. An acoustic sound is non-electric, a sound that is not created, modified or enhanced electronically. The sound is mechanical and thus vibration is needed to create it. On the other hand, electronic sounds are either produced or modified by electronic means. They can be created using: raw sounds that are modified electronically, such as an electric guitar; or from a source that creates the sound electronically, such as a synthetizer or MIDI.

Most of the musical pieces are a combination of several sound sources such as different musical instruments and voices. These instruments are often organized in layers, each containing one or more instruments, which have distinct roles (e.g., melodic or rhythmic). These are the basis of musical texture, the dimension described in Sections 2.3.7 and A.7. Various well known combinations of instruments and voices exist and

have specific names such as ensembles (group of several instruments), orchestras (large groups), bands, choirs and smaller specific ensembles such as Jazz trios, quartets and quintets.

A.6. Expressive Techniques

Expressive techniques refer to the way a performer plays a musical piece, specifically the techniques used by him to create the musical detail that articulates a style or interpretation of a style. For a condensed introduction to expressive techniques see Section 2.3.6.

Tempo changes

Although tempo was presented as an element related to rhythm, we know that many dimensions of music overlap. Here, the tempo, and especially its changes over time, may also affect the expressivity of music. Changes in tempo can normally be described as:

- Gradually getting faster or slower
- Instantly slowing down or getting faster
- Returning to the original speed

Stylistic indications

Stylistic indications consist of terms used by composers to indicate how a musical piece should be played by the performer. These indications affect the style of music, helping defining its genre.

Such indications are normally given by single words or even phrases describing how specific sections of a musical piece are to be played. Some common Italian terms are *legato*, meaning "smoothly", and *rubato*, indicating the piece may be played "with freedom", in the performer's own time/style. Some composers may even write longer phrases in the musical score as stylistic indications. For example, stating "marked / with accent", "slowly, with expression", "medium funk" or "moderate jazzy beat", which of them are more common in specific genres or instruments.

Articulation

Articulation is a performance technique affecting the way transitions (or continuity) between notes are performed. Several articulation techniques exist, each influencing differently on how a note is played. Some of these techniques are specific to particular instruments and may sound different or not be possible in other instrument types. As

an example, techniques such as *pizzicato*, which are specific to string instruments, are obviously impossible in wind instruments (e.g., flute).

Typical articulation styles include *legato* and *staccato* but can also include accent techniques (e.g., *sforzando*) which are also part of the dynamics element. A summary of the articulation methods is presented in Table A.7.

Articulation	*Description*
Legato	Notes should be played without separation, smoothly and connected (indicated with a slur).
Staccato	Short and detached note (opposite of *legato*).
Portato	Also called *mezzo-legato* or articulated *legato*, *non-legato* or *portamento*. From the Italian word *portare*, "to carry", consists in a smooth but pulsing articulation.
Tenuto	From the Italian word *tenere*, "to hold". The meaning varies with context from holding the note for its full length to playing it slightly louder.
Marcato	Meaning "marked", it indicates a note or passage to be accentuated, played louder than the surrounding ones.
Martellato	Italian word for "hammered", more common in bowed string instruments, indicates that notes should be played explosively.
Pizzicato	Technique for string instruments (e.g., violin) where the strings are plucked.

Table A.7: Listing of common articulation techniques.

These instructions are typically indicated in the musical score by a symbol above or below the musical note, as exemplified in Figure A.8.

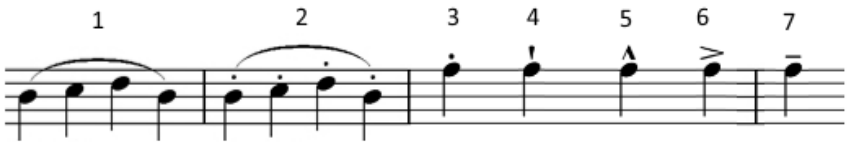


Figure A.8: Example of articulation techniques indicated in a musical score: 1) *legato*, indicated with a slur (arch); 2) *portato*, mixing *staccato* markers and the slur; 3) *staccato*; 4) *stacattissimo*; 5) *martellato*; 6) *marcato*; 7) *tenuto*.

Ornamentation

Ornamentation is the decoration or embellishment of specific notes in a melody or harmony with special features to provide interest and give the performer opportunity to add variety expressiveness to a piece. A description of common ornaments in Western music is presented in Table A.8. While most of the described ornaments are more prevalent in western classical music, some exist that are specific to other genres (e.g., pop/rock hammer-ons on electric guitar) or music types. An additional example is the ornamentation in Indian music, or *Gamak* (meaning "ornamented note" in Sanskrit, the name given to ornamentations), which contains several types of slides, stresses and deflections (Powers, 1958).

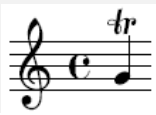
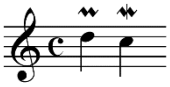
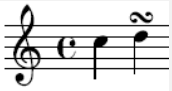
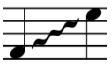
Ornament	*Description*	*Symbol*
Trill	Rapid alternation between the indicated note and the one above.	tr. mark. 
Mordent	Two variations exist: upper and lower mordent. Consists on a rapid alternation between: the note, the note above (upper mordent) or below (lower mordent) and returning to the original note again.	Indicated with a tilt, the lower mordent adds a vertical line to it. 
Turn	Playing the notes around a note: the indicated note, then the upper note, the lower note, and the first note again.	Indicated by an inverted and 90º rotated S. 
Glissando	A slide from one note to another. The intermediate notes are only briefly heard, thus differentiating it from *portato*.	A wavy line connecting both notes. 

Table A.8: Ornaments used in western music.

Instrumental, Vocal and Electronic Techniques

Over time, creative composers and performers have explored both instruments and vocals, attempting to produce new types of sounds and effects. Such efforts gave rise to several new expressive techniques that are nowadays used to enrich musical pieces across all genres.

Regarding musical instruments, some of the existent effects are generic, available to most instruments. Some of these are similar to the techniques described previously, such as *legato*, *staccato*, slurs (playing two notes simultaneously) or accents. Vibrato is a widely used technique consisting in a note rapidly changing its pitch (vibrating). *Tremolo* is somewhat similar, but with the note changing in amplitude.

Specific techniques exist that are exclusive to families of instruments. As an example, *pizzicato* (plucking the strings) is used in violin and other orchestral string instruments. Guitar and bass techniques include strumming and finger picking, slapping (bass related) and distortion effects using pedals.

The voice is also a very powerful instrument and composers have explored multiple techniques to enhance it. Some of the possible techniques are using *vibrato*, singing in *falsetto* (in an upper register), rapping (speaking or singing following the rhythm and beat of the song), improvising nonsense sounds and syllables (named scat, common in jazz music).

Finally, it is also possible to use electronic manipulation to alter the sound of an instrument or the voice. Two very common examples of electronic manipulation nowadays are panning, moving the sound (stereo) from one channel to another (left to right or vice-versa), and the usage of effects pedals and amplifiers on guitars, distorting its original sound. Other techniques related with voice are the vocoder (voice encoder) and auto-tune, which have been widely used by pop artists in the last two decades to distort the voice or alter the pitch of poorly sung notes.

A.7. Musical Texture

Musical texture refers to the way the rhythmic, melodic and harmonic information produced by musical instruments and voices is combined in a musical composition. For a condensed introduction to musical texture see Section 2.3.7.

Number of Layers, Density and Range

The number of layers in a musical piece influences the texture density, which can be either thin or thick. Musical layers in a piece can be broadly categorized as: a single melodic layer (or line); a melodic layer with some additional accompaniment layer; multiple melodic lines; a primary melodic line with a counter-melody (an accompanying melody played at the same time); and others.

In most musical pieces, the composer selects distinct instruments (or groups of instruments) to play different roles such as melodic, harmonic and rhythmic. A layer with the melody role usually draws the listener's attention, usually containing an instrument

or voice leading the music (e.g., the main melody). A layer with a harmonic role is used to hold the harmony in the piece, for instance by using instruments such as a bass or electric guitar playing some chordal accompaniment.

A layer with rhythmic role is used to hold the underlying rhythm in the piece by providing beat to the other layers, such as the rhythm section in a band (e.g., guitars, bass or drums). The human voice is sometimes used with the same purpose, as is the case of a beatboxer playing with a rapper.

Furthermore, these roles are not exclusive since a single layer can have multiple roles. A rhythm and a bass guitar can have harmonic and rhythmic roles. The same guitar may play, for instance, a secondary melody (e.g., counter-melody) in a section of the musical piece.

Texture Types

The combination of musical layers in a musical piece and their relation forms different types of textures. Some of the most common types of texture are: monophonic, polyphonic, homophonic and heterophonic.

Monophonic Texture

A monophonic texture refers to music with only one musical (melodic) line or layer at a time, thus no harmony or accompaniment (Benward & Saker, 2008, p. 147) as illustrated in Figure A.9. This can be a single instrument or line, or a group in unison (even if at octaves apart). This type was common in medieval music and some mainstream pop music.

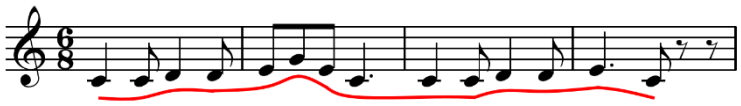


Figure A.9: Monophonic texture from the English kids song "Pop Goes the Weasel". The melodic line is drawn in red.

Polyphonic Texture

A polyphonic texture consists of two or more melodic lines moving independently or in some cases in imitation with one another (Benward & Saker, 2008, p. 148), as illustrated in Figure A.10. The most intricate types of polyphonic texture – *canon* and *fugue* – may introduce three, four, five or more independent melodies simultaneously. The polyphonic texture type was predominant in the Renaissance period, where each (usually vocal) part had a melody. The melodies of all parts were performed at the same time, yet

all fitted together harmonically (Benward & Saker, 2008, p. 150). Imitative (Imitation) texture is sometimes used to distinguish a special type of polyphonic texture where a musical idea is echoed from "voice" to "voice". It was especially common in Renaissance and Baroque periods.

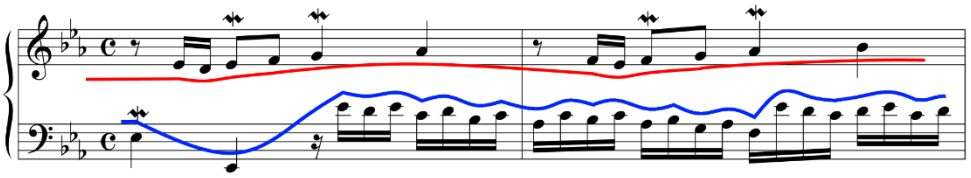


Figure A.10: Polyphonic texture, two independent lines (red and blue) from Bach's "Invention no. 5 in E-flat Major, BWV 776, mm. 1–2".

Homophonic Texture

Homophonic texture is the most common texture type heard in Western music, based on melody and accompaniment. It consists in two or more simultaneous musical layers moving in sync (as opposed to polyphonic texture), with a prominent melody (melodic layer) in the upper part, supported by a less intricate group of layers forming a background of harmonic accompaniment (harmonic layer) (Benward & Saker, 2008, p. 149), as shown in Figure A.11.

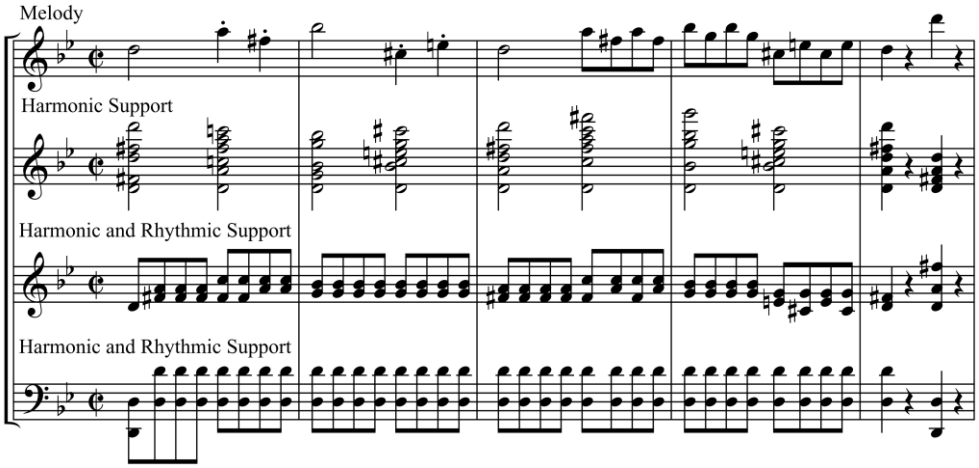


Figure A.11: Homophonic texture with a melody line and separated harmonic and rhythmic support. From Mozart's "Symphony no. 40 in G Minor, K. 550, I: Molto Allegro, mm. 221–225".

Homophonic texture was the predominant texture type during the romantic and

impressionistic periods. Nowadays, nearly all pop music, as well as most Jazz are homophonic (Benward & Saker, 2008, p. 151). Typical examples are a singer with instrument accompaniment or a performer playing a solo, accompanied by a group of other instruments such as a band or orchestra.

Other less common texture types exist, such as heterophonic texture, which consists in the two or more voices simultaneously performing variations of the same melody (melodic line).

A.8. Musical Form

Musical form or musical structure refers to the overall structure of a musical piece, and describes the layout of a composition as divided into sections (Brandt, 2011). For a condensed introduction to musical form see Section 2.3.8.

Song Elements

Introduction

The initial section that appears at the beginning of a musical piece, usually containing only music and no lyrics. Some of the most common forms are: based on the chords of the verse or chorus; using only percussion parts, such as drums setting the rhythm; or a solo vocal or instrumental solo.

Verse

The *verse* is usually made of a group of lines of a poem (*stanza*), where its musical structure "nearly always recurs at least once with a different set of lyrics" (Everett, 1999, p. 15). The lyrical part of the *verse* is used by the song writer to transmit the message, it is used together with the accompaniment to transmit "the story, the events, images and emotions[134] that the writer wishes to express" (M. Davidson & Heartwood, 1997, p. 6).

Regarding the musical structure of the song, the song's *verses* have the primary role of supporting the *chorus* message, both in terms of its music and lyrical message.

Chorus

Sometimes called *refrain*, the chorus is composed of passages of unchanging music and

[134] The influence of each part of a song to emotion (lyrics and audio) varies greatly between songs and genres.

lyrics that provide a periodic sense of return. It "contains the main idea, or big picture, of what is being expressed lyrically and musically. It is repeated throughout the song, and the melody and lyric rarely vary" (M. Davidson & Heartwood, 1997, p. 6).

Musically, the chorus usually contains a thicker musical texture, caused by the addition of backing vocals or instruments (hence its name). The thicker texture, combined with the fact that it is repeated during the song, confers to the chorus a stronger musical and emotional intensity than the verse and thus becoming the most memorable element of the song, used to transmit the main message (Everett, 1999, p. 16).

Bridge

A section that many times contrasts with the verse and, as the name implies, is used to link the verse and chorus, used to "break up the repetitive pattern of the song and keep the listeners attention" (M. Davidson & Heartwood, 1997, p. 7).

Conclusion

The conclusion, sometimes referred as *coda* or *outro* (mainly in pop music) of a musical piece is the final section or ending of the song. Frequently used forms are based on a repeat and fade-out or instrumental part.

Instrumental Solo

Sometimes songs contain a solo section where a melodic line is played most of the times by a single performer (e.g., by a guitar player), without lyrics. This performed melody is often based on the chorus with added ornamentations and embellishments (described in expressive techniques, Section 2.3.6) such as riffs and scale runs (i.e., navigating scales in an upward or downward momentum). In some genres such as Jazz, the solo may be improvised, showcasing the performer creativity and quality.

In Table A.9 we present the song "Billie Jean" by Michael Jackson to illustrate the abovementioned notions:

She was more like a beauty queen from a movie scene I said, "Don't mind, but what do you mean, I am the one Who will dance on the floor in the round?" She said I am the one who will dance on the floor in the round	1st Verse
She told me her name was Billie Jean As she caused a scene Then every head turned with eyes that dreamed of being the one	2nd Verse

Who will dance on the floor in the round	
People always told me, "Be careful of what you do And don't go around breaking young girls' hearts" And mother always told me, "Be careful of who you love And be careful of what you do, 'cause the lie becomes the truth"	Bridge
Billie Jean is not my lover She's just a girl who claims that I am the one But the kid is not my son She says I am the one, but the kid is not my son	Chorus
For forty days and for forty nights, the law was on her side But who can stand when she's in demand? Her schemes and plans 'Cause we danced on the floor in the round So take my strong advice, just remember to always think twice (Don't think twice) Do think twice! (hoo)	3rd Verse
She told my baby we'd danced till three, then she looked at me Then showed a photo of a baby crying, his eyes were like mine (oh no) 'Cause we danced on the floor in the round, baby	4th Verse
People always told me, "Be careful of what you do And don't go around breaking young girls' hearts" She came and stood right by me Then the smell of sweet perfume This happened much too soon She called me to her room	Bridge
Billie Jean is not my lover She's just a girl who claims that I am the one But the kid is not my son Billie Jean is not my lover She's just a girl who claims that I am the one But the kid is not my son She says I am the one, but the kid is not my son	Chorus
(...)	Instrumental Solo
She says I am the one, but the kid is not my son	Chorus

Billie Jean is not my lover She's just a girl who claims that I am the one But the kid is not my son She says I am the one, but the kid is not my son She says I am the one She says she is the one She says I am the one	
Billie Jean is not my lover Billie Jean is not my lover Billie Jean is not my lover Billie Jean is not my lover Billie Jean is not my lover Billie Jean is not my lover	Outro

Table A.9: Song structure of “Billie Jean” by Michael Jackson.

Organization Levels

The structure of a musical piece may be analyzed at different levels, from simple pulses arranged in beats to further and more complex organizations. Music form can be roughly divided into three levels designated as passage, piece and cycle. In the analysis of musical form, any components that can be defined on the time axis (such as sections and units) are conventionally designated by letters (e.g., component A, B, C).

Passage

The passage is the lowest level of construction and concerns the way musical phrases are organized into musical sentences and "paragraphs" such as the verse of a song. As an example, the first verse of the song “Twinkle, twinkle, little star” is composed of two (A, B) differently-rhymed couplets, and can be described as AABB, as illustrated in the example of the section below (“Binary Form”).

Piece

The level above passage is called piece and concerns the structure of an entire musical piece. The musical form analysis that was previously used in passage can be employed at each level. As an example, if the song simply repeats the same musical sections, it is said to be in strophic form. However, if two sections repeat with distinct changes, such as the verse and chorus or a fast and slow sections, it is said to be in binary form.

Cycle

The higher level is concerned about the organization of pieces in larger compositions. This organization varies across genres related with classical music, with the pieces (sometimes called *movements*) originating *suites*, representing a set of Baroque dances, *symphonies*, *concertos* and *sonatas*.

Basic Musical Forms

The different musical sections, made of the repetition of melodic material or the presentation of new, contrasting material are usually identified by letters (i.e., A, B, C). The most common organization methods are described next.

Through-composed

A structure in which there is no repeat or return of any large-scale musical section. An example of such piece is Schubert's composition of Goethe's poem "Erlkönig", which can be described as:

A B C D E...

Strophic Form

Strophic form (also called "verse-repeating" or chorus form) is the opposite of through-composed. In such songs, all verses or *stanzas* of the text are sung to the same music. It is interesting to note that songs with verse-chorus form are binary (ABAB), thus contrasting with strophic. However they may also be interpreted as strophic form at a higher level (AA). Many folk and popular songs are strophic in form. Some examples are the "Old McDonald had a farm" kids' song and Bob Dylan's "Blowin' in the Wind".

A A A A A...

Binary Form

This form consists of two different albeit related sections (AB), which are often repeated over the course of the piece, normally as "A-A-B-B". The basic premise of such form is the contrast between both sections. Examples of such form are the "Twinkle, Twinkle, Little Star", or the "Minuet in G Major" by Johann Sebastian Bach. Below, we present an excerpt of the former to illustrate the AABB structure:

Twinkle, twinkle, little star, (A)

How I wonder what you are. (A)

Up above the world so high, (B)

Like a diamond in the sky. (B)

Ternary Form

A ternary structure is, as the name implies, a three-part form usually with two sections (AB), featuring a return of the initial section after the contrasting section (used in the form ABA). This return to the initial section is used to achieve balance and symmetry. However, this is not always the case, since in many cases the initial section is repeated (as AABA).

The ternary form was the base to the thirty-two-bar form songs. Thirty-two-bar form uses four sections, most often eight measures long each (4 x 8 = 32), composed of two verses (A sections), a contrasting section (B section, the bridge or "middle-eight") and a return to one last A section (thus, AABA). One example of thirty-two-bar AABA form is the Christmas carol "Deck the Halls":

Deck the halls with boughs of holly! Fa-la-la-la-la, la-la-la-la. (A)

'Tis the season to be jolly. Fa-la-la-la-la, la-la-la-la. (A)

Don we now our gay apparel. Fa-la-la, la-la-la, la-la-la. (B)

Troll the ancient Yuletide carol. Fa-la-la-la-la, la-la-la-la. (A)

Other forms exist such as medley, rondo and others. Medley, similarly to thought-composed form, is the extreme opposite of strophic form. It is an indefinite sequence of self-contained sections (ABCD...), often with repetitions (AABBCCDD...), normally composed of parts of existing pieces played sequentially sometimes with overlaps. Examples are pop mega-mixes or orchestral overtures. Rondo form can be roughly described as an extra-long version of the ternary form, having a recurring theme (A) alternating with different (usually contrasting) sections called "episodes", such as (ABACADAE...).

APPENDIX B

NOVEL DATASET DETAILS

This section contains additional information about our novel dataset besides the details presented in Section 4.1.

B.1. Emotion Tags per Quadrant

The number of songs tagged with each emotion tag, both total and divided per cluster, is presented in Table B.1. This information is illustrated in different forms in Section 4.1, such as the tag cloud format in Figure 4.15

Similar information regarding genre tags is provided in Section 4.1, e.g., Table 4.3 and Figure 4.16.

Emotion tag	*Q1*	*Q2*	*Q3*	*Q4*	*Total*
Acerbic	0	7	6	0	13
Aggressive	0	10	0	0	10
Agreeable	2	0	1	18	21
Amiable	6	0	0	22	28
Angry	0	71	3	0	74
Angst-Ridden	0	11	10	0	21
Anguished	0	11	42	0	53
Anxious	0	47	2	0	49
Atmospheric	0	0	1	50	51
Austere	0	0	3	6	9
Autumnal	1	0	5	9	15
Belligerent	0	3	0	0	3

Bitter	0	29	45	0	74
Bittersweet	1	0	2	18	21
Bleak	0	18	55	1	74
Boisterous	18	1	0	2	21
Brash	2	2	2	0	6
Brassy	14	0	0	1	15
Bravado	5	1	0	3	9
Bright	9	0	0	16	25
Brittle	0	0	5	0	5
Brooding	2	2	28	2	34
Calm	0	0	0	84	84
Campy	8	3	0	3	14
Carefree	20	0	0	15	35
Cathartic	12	12	4	2	30
Celebratory	17	0	0	0	17
Cerebral	1	0	26	15	42
Cheerful	52	0	0	16	68
Circular	0	0	9	7	16
Clinical	0	1	18	1	20
Cold	0	4	18	0	22
Complex	3	3	20	13	39
Confident	12	0	0	16	28
Confrontational	0	17	2	0	19
Crunchy	0	2	0	1	3
Cynical	0	17	8	0	25
Delicate	0	0	5	51	56
Detached	1	1	23	8	33
Difficult	0	0	5	0	5
Distraught	0	11	42	0	53
Dramatic	14	0	0	6	20
Dreamy	0	0	1	46	47
Druggy	0	2	0	2	4

Earnest	2	0	0	17	19
Earthy	4	0	1	20	25
Eccentric	11	0	12	3	26
Eerie	0	14	7	4	25
Effervescent	4	0	0	3	7
Elaborate	6	2	4	20	32
Elegant	3	0	0	35	38
Energetic	35	0	0	2	37
Enigmatic	0	0	0	2	2
Epic	0	0	1	4	5
Erotic	3	0	0	0	3
Ethereal	0	0	0	41	41
Euphoric	5	0	0	0	5
Exciting	56	6	1	0	63
Explosive	5	11	1	1	18
Exuberant	14	0	0	3	17
Fierce	0	56	0	0	56
Fiery	23	23	3	0	49
Flowing	1	0	0	3	4
Fractured	0	1	1	0	2
Freewheeling	10	0	0	3	13
Fun	34	0	0	4	38
Gentle	1	0	0	51	52
Giddy	18	0	0	0	18
Gleeful	9	0	0	1	10
Gloomy	1	2	81	0	84
Good-Natured	6	0	0	22	28
Greasy	0	1	61	0	62
Gritty	6	3	10	1	20
Gutsy	18	3	1	1	23
Happy	54	0	0	7	61
Harsh	0	70	5	0	75

Hedonistic	3	5	1	0	9
Hostile	0	72	1	0	73
Humorous	25	2	1	3	31
Hungry	1	3	0	0	4
Hypnotic	4	2	4	38	48
Indulgent	1	1	12	2	16
Innocent	8	0	0	29	37
Insular	0	1	2	1	4
Intense	20	11	4	0	35
Intimate	3	0	0	29	32
Ironic	0	1	0	1	2
Irreverent	3	5	0	0	8
Jittery	0	44	0	0	44
Jovial	1	0	0	0	1
Joyous	65	0	0	12	77
Knotty	1	0	2	1	4
Laid-Back	0	0	0	28	28
Lazy	0	1	40	1	42
Light	5	0	0	41	46
Literate	1	0	4	10	15
Lively	41	0	0	4	45
Lonely	0	0	19	2	21
Lush	7	0	3	32	42
Majestic	1	0	0	2	3
Malevolent	0	33	1	0	34
Manic	1	6	5	0	12
Marching	4	1	0	0	5
Meandering	0	0	0	3	3
Meditative	0	0	1	1	2
Melancholy	2	0	12	14	28
Mellow	0	0	0	28	28
Menacing	0	35	8	0	43

Messy	1	2	27	0	30
Mighty	2	0	0	0	2
Mystical	4	0	2	0	6
Naive	0	0	9	3	12
Negative	0	39	1	0	40
Nervous	0	44	0	0	44
Nihilistic	0	21	0	0	21
Nocturnal	3	1	6	20	30
Nostalgic	1	0	6	22	29
Ominous	0	29	12	1	42
Optimistic	5	0	0	8	13
Organic	2	1	2	32	37
Outraged	0	25	0	1	26
Outrageous	3	47	0	0	50
Paranoid	0	18	3	0	21
Passionate	20	0	0	8	28
Pastoral	0	0	0	14	14
Peaceful	0	0	0	84	84
Plaintive	0	0	1	13	14
Playful	26	0	0	9	35
Poignant	1	0	2	24	27
Positive	25	0	0	9	34
Powerful	7	0	0	1	8
Precious	1	0	3	37	41
Provocative	9	8	1	4	22
Pulsing	9	0	0	8	17
Pure	1	0	1	6	8
Quiet	0	0	7	39	46
Quirky	4	0	3	9	16
Rambunctious	11	2	0	0	13
Ramshackle	0	2	1	0	3
Raucous	1	12	0	0	13

Rebellious	0	37	2	1	40
Reckless	0	54	4	0	58
Refined	2	0	0	36	38
Reflective	2	0	0	16	18
Regretful	1	0	0	0	1
Relaxed	0	0	0	40	40
Reserved	0	0	3	43	46
Restrained	1	0	10	23	34
Reverent	3	0	0	32	35
Rollicking	12	0	0	4	16
Romantic	12	0	0	22	34
Rousing	26	0	0	5	31
Rowdy	4	13	3	0	20
Rustic	4	0	6	11	21
Sad	0	0	76	5	81
Sarcastic	0	17	8	0	25
Sardonic	0	0	1	0	1
Searching	2	0	1	7	10
Self-Conscious	0	0	4	5	9
Sensual	26	0	0	22	48
Sentimental	2	0	0	23	25
Serious	2	0	7	4	13
Sexual	19	13	1	0	33
Sexy	47	0	0	3	50
Silly	7	4	0	4	15
Sleazy	2	18	1	0	21
Slick	9	0	0	12	21
Smooth	4	0	0	31	35
Snide	0	19	8	1	28
Soft	0	0	7	39	46
Somber	0	3	35	11	49
Soothing	1	0	1	72	74

Sophisticated	7	0	0	23	30
Spacey	0	0	1	12	13
Sparkling	11	1	0	35	47
Sparse	0	0	13	10	23
Spicy	44	0	6	3	53
Spiritual	6	0	1	44	51
Spooky	0	10	0	0	10
Sprawling	0	0	0	3	3
Springlike	1	0	0	8	9
Stately	1	1	1	10	13
Street-Smart	4	1	1	5	11
Striding	1	0	1	8	10
Strong	22	0	0	1	23
Stylish	8	0	0	15	23
Suffocating	0	6	0	0	6
Sugary	0	0	3	1	4
Summery	12	0	1	15	28
Swaggering	6	0	0	2	8
Sweet	7	0	0	32	39
Swinging	7	0	0	4	11
Technical	1	0	2	0	3
Tender	1	0	0	0	1
Tense	0	47	2	0	49
Theatrical	17	0	1	14	32
Thoughtful	0	0	0	2	2
Threatening	0	2	1	0	3
Thrilling	10	0	0	0	10
Thuggish	2	13	4	1	20
Tragic	0	1	0	0	1
Trashy	1	18	0	0	19
Trippy	0	1	3	9	13
Uncompromising	3	36	10	0	49

Unsettling	0	20	3	0	23
Uplifting	6	0	0	5	11
Urgent	8	14	0	0	22
Visceral	1	12	1	0	14
Volatile	1	41	4	0	46
Warm	3	0	0	24	27
Weary	0	0	46	5	51
Whimsical	8	1	1	12	22
Wintry	0	0	4	8	12
Wistful	0	0	0	21	21
Witty	13	0	1	8	22
Wry	2	0	0	6	8
Yearning	2	0	0	9	11
Total[135]	*1266*	*1355*	*1076*	*2143*	*5840*

Table B.1: Number of songs per emotion tag and quadrant.

[135] The total number of songs is higher than the dataset size (900) because each song can be associated with more than one genre tag.

APPENDIX C

FEATURES RELEVANCE VISUALIZATION

Besides the information presented in Section 4.6, we also designed additional visualizations of the features relevance and weight. These were used to better understand how the existent features performed for each of the studied problems.

In this section we present some of these illustrations containing additional information.

C.1. Feature Weights by Musical Dimension

The figures in the following sections illustrate the weight of the various features for each musical dimension, organized by problem.

A. Feature Weight to Arousal and Valence Classification

Figure C.1 and Figure C.2 show the features' weight grouped by musical dimension for arousal and for valence classification. As can be seen, while many features for arousal have weights above 0.1, for valence classification most were considered much less relevant, with only some novel expressive techniques ones above the same 0.1 threshold.

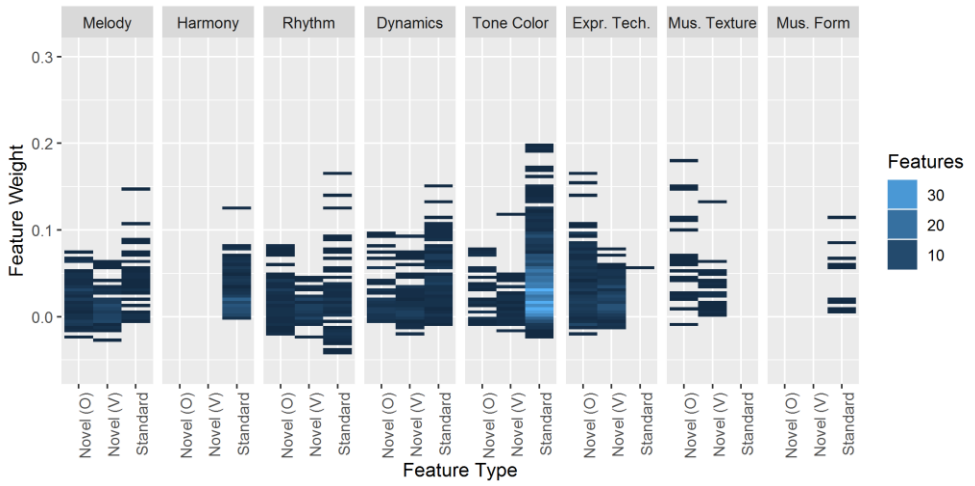


Figure C.1. Feature weights related to arousal classification, split by musical dimension.

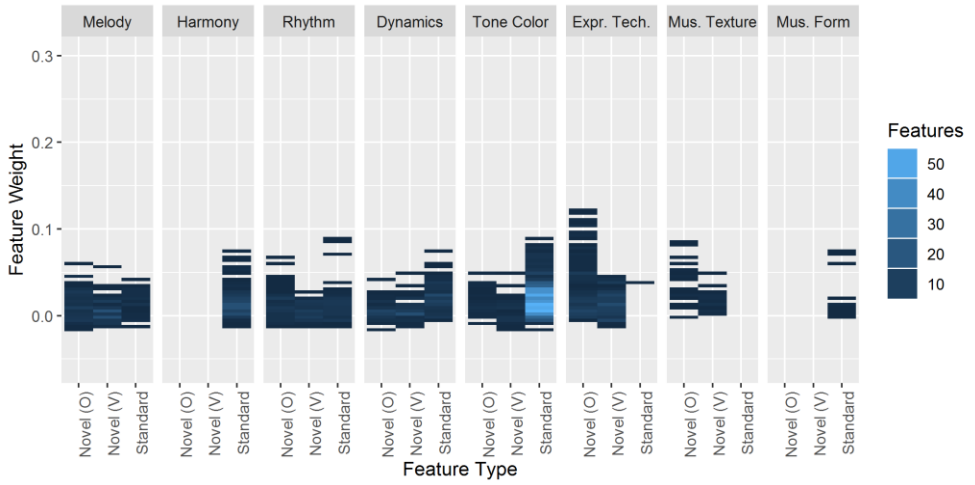


Figure C.2. Feature weights related to valence classification, split by musical dimension.

B. Feature Weight to Arousal Classification using Positive or Negative Valence Songs

Figure C.3 and Figure C.4 show the features' weight to discriminate between positive and negative arousal, considering only the positive valence songs or negative valence songs. As illustrated, the features considered discriminate arousal remarkably better for

negative valence songs.

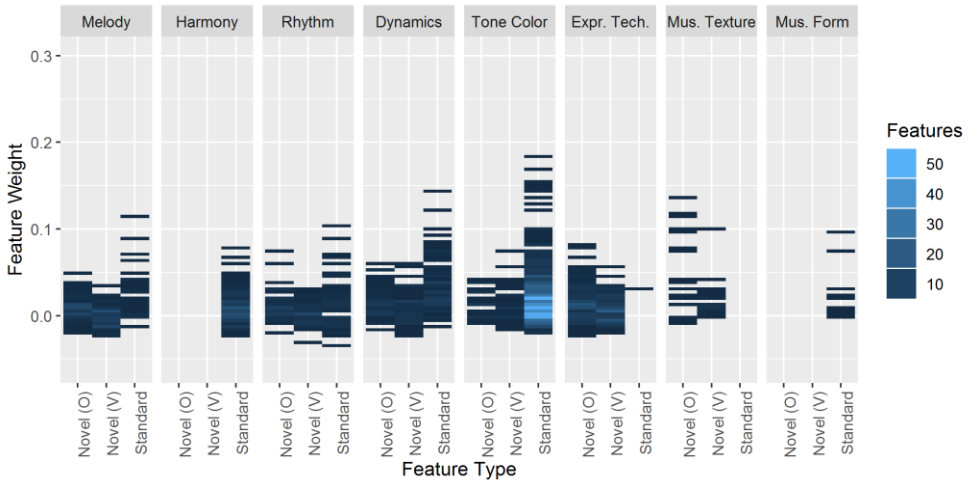


Figure C.3. Feature weights related to arousal classification for positive valence songs only, divided by musical dimension.

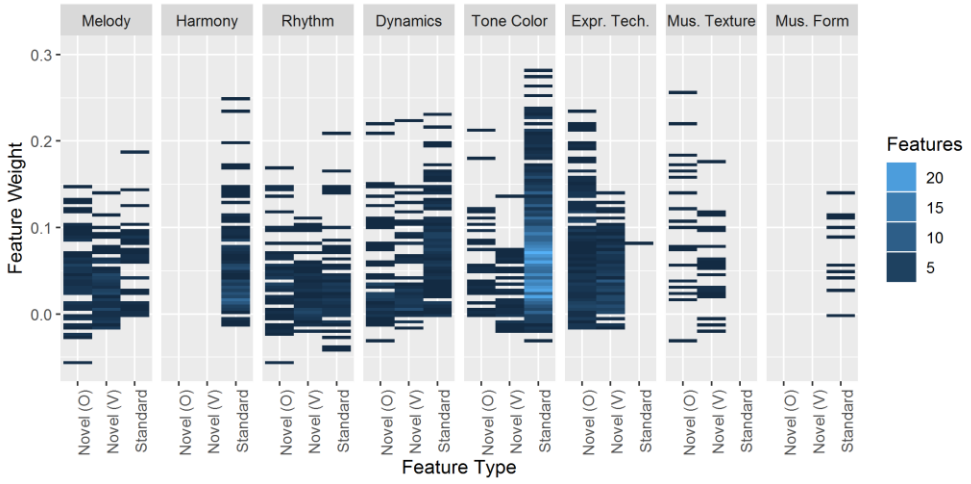


Figure C.4. Feature weights related to arousal classification for negative valence songs only, divided by musical dimension.

C. Feature Weight to Valence Classification using Positive or Negative Arousal Songs

Figure C.5 and Figure C.6 show the features' weight to discriminate between positive and negative valence, considering only the positive or negative arousal songs. As shown, the features weight is much lower than the previous, especially for low arousal songs.

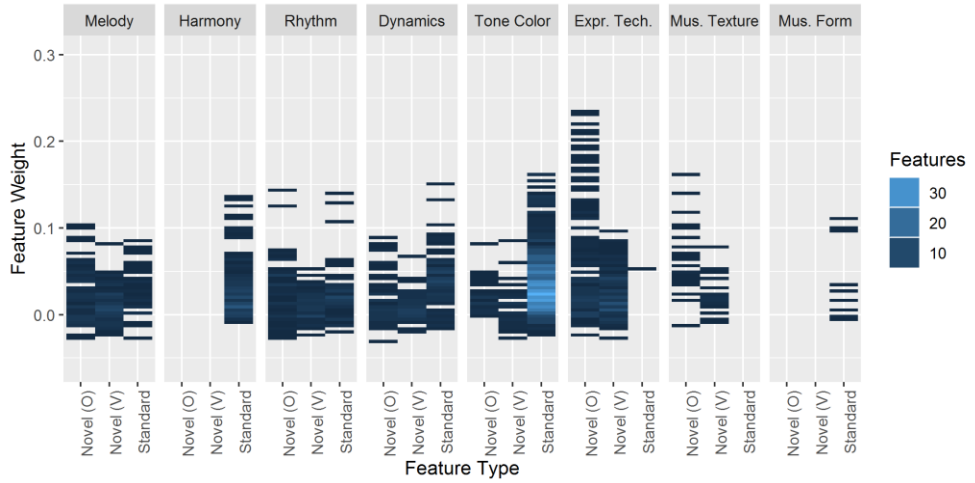


Figure C.5. Feature weights related to valence classification for positive arousal songs only, divided by musical dimension.

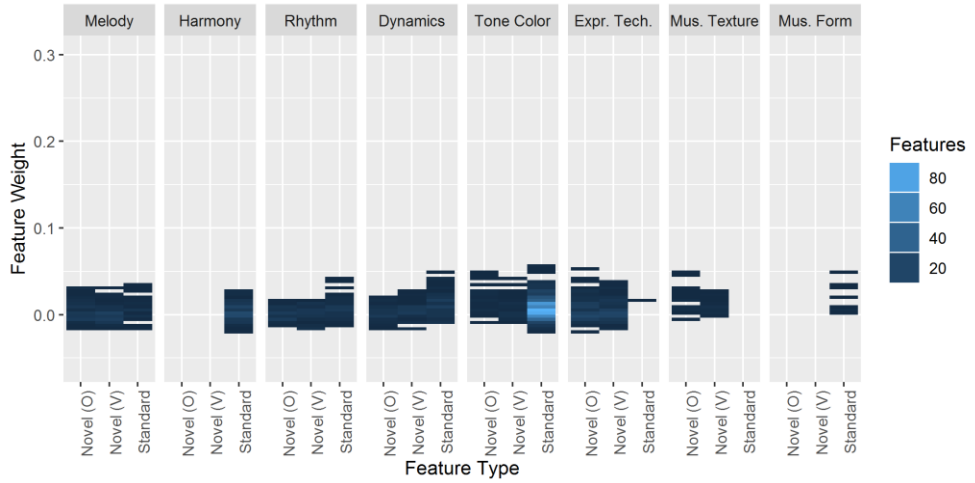


Figure C.6. Feature weights related to valence classification for negative arousal songs only, divided by musical dimension.

D. Feature Weights to Discriminate each of the Four Quadrants

Finally, Figure C.7 to Figure C.10 show the features' weight to discriminate between each specific quadrant against the remaining.

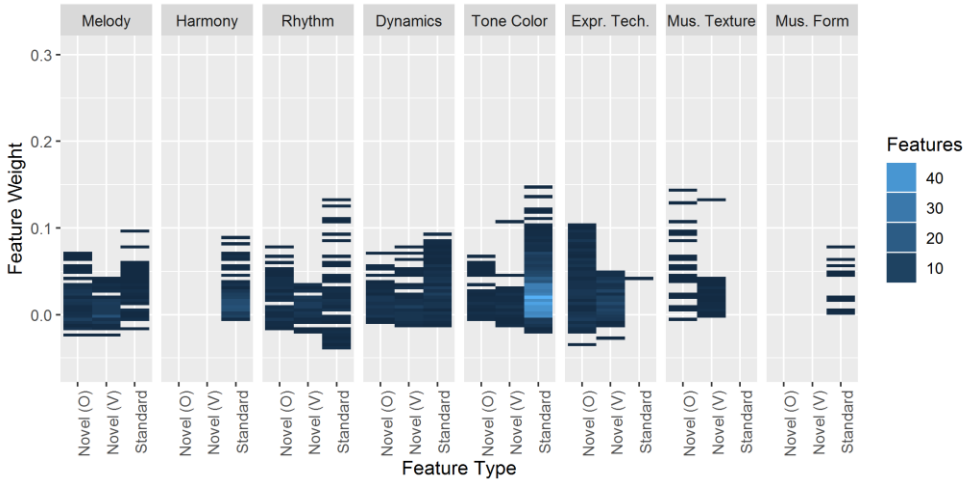


Figure C.7. Feature weights related to quadrant 1 vs. other quadrants classification, divided by musical dimension.

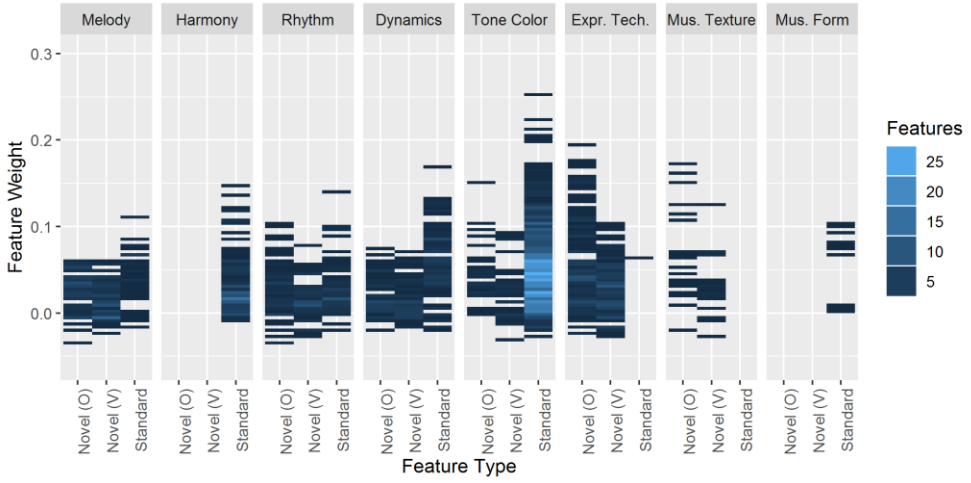


Figure C.8. Feature weights related to quadrant 2 vs. other quadrants classification, divided by musical dimension.

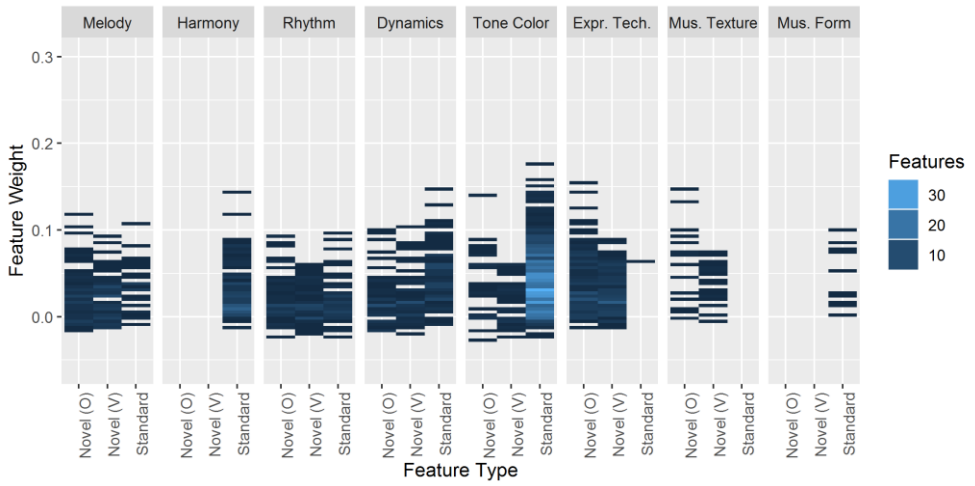


Figure C.9. Feature weights related to quadrant 3 vs. other quadrants classification, divided by musical dimension.

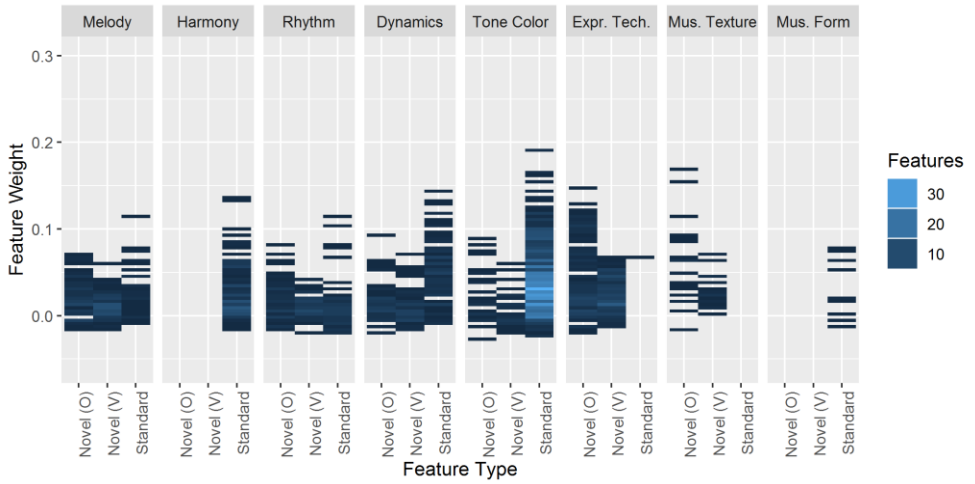


Figure C.10. Feature weights related to quadrant 4 vs. other quadrants classification, divided by musical dimension.

C.2. Best Features for each Emotion Problem

The group of 10 figures from Figure C.11 to Figure C.20 complement the feature analysis presented in Section 4.6. These show the distribution of the top 10, 20, 30, 50 and 100 features per musical dimension for each of the 11 emotion classification problems.

A. Distribution of the Top 10 Features

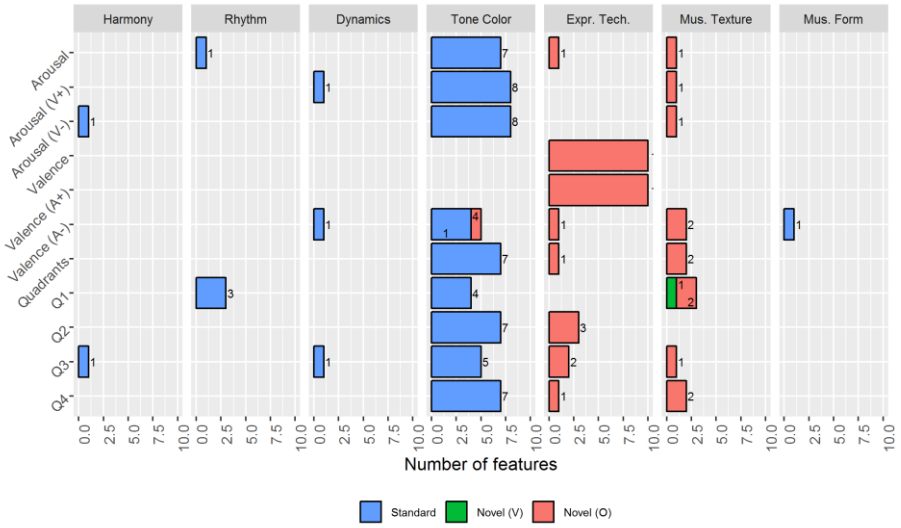


Figure C.11. Distribution of the top10 features per musical dimension (horizontal).

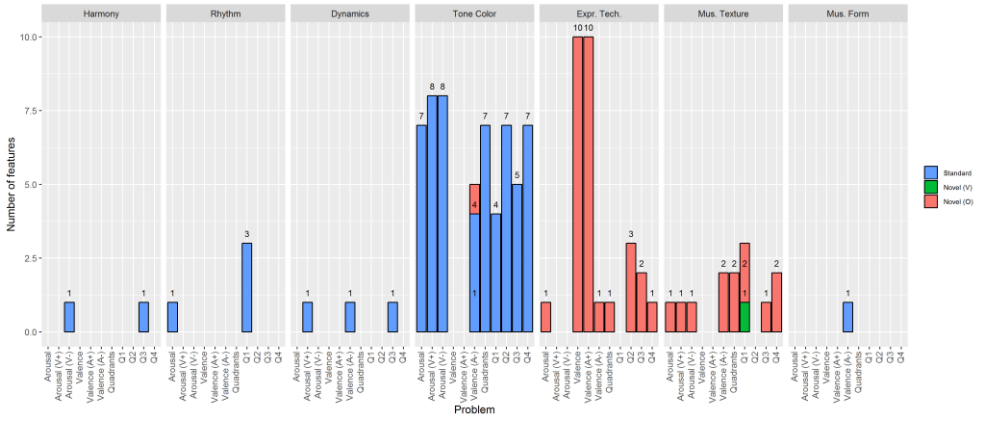


Figure C.12. Distribution of the top10 features per musical dimension (vertical).

B. Distribution of the Top 20 Features

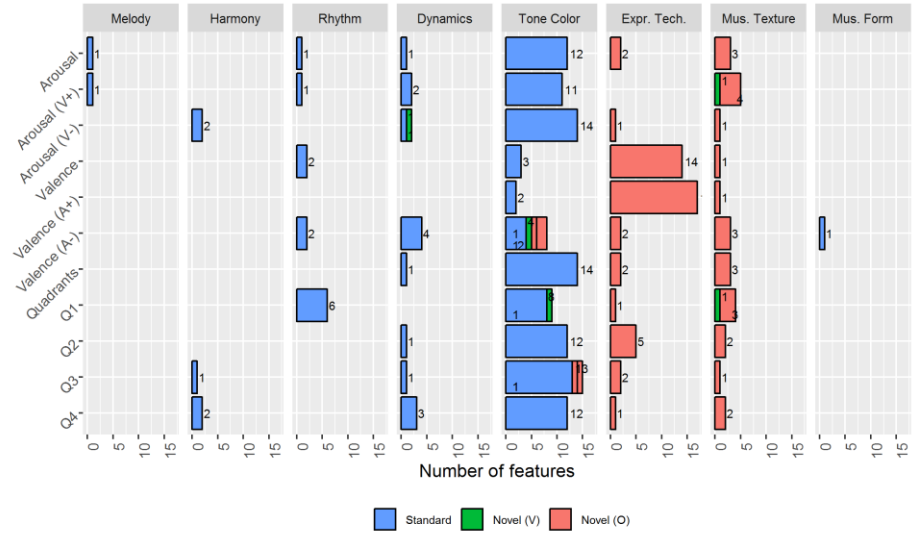


Figure C.13. Distribution of the top20 features per musical dimension (horizontal).

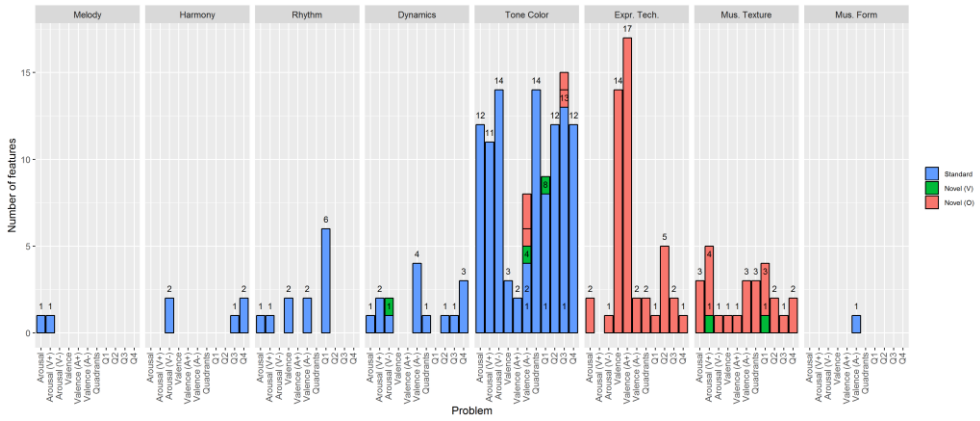


Figure C.14. Distribution of the top20 features per musical dimension (vertical).

C. Distribution of the Top 30 Features

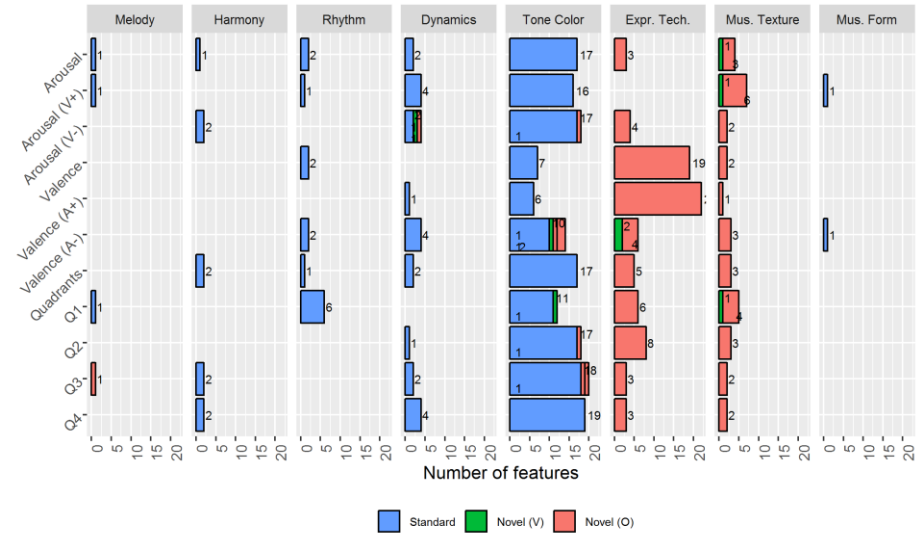


Figure C.15. Distribution of the top30 features per musical dimension (horizontal).

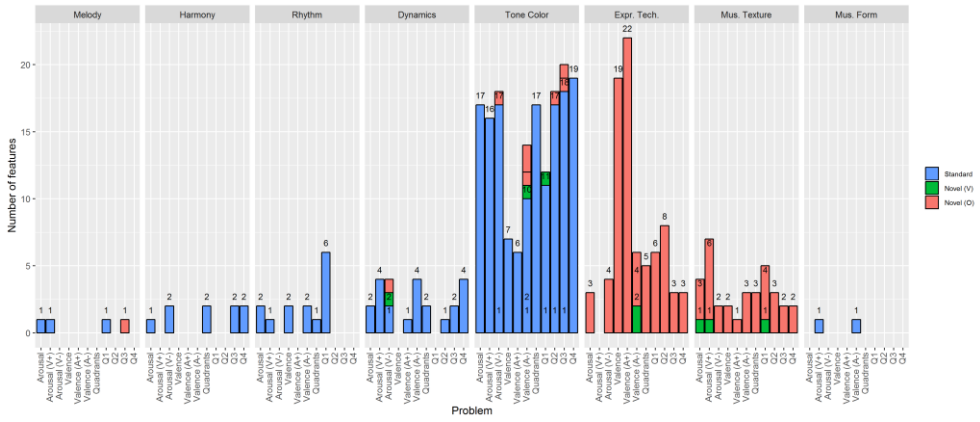


Figure C.16. Distribution of the top30 features per musical dimension (vertical).

D. Distribution of the Top 50 Features

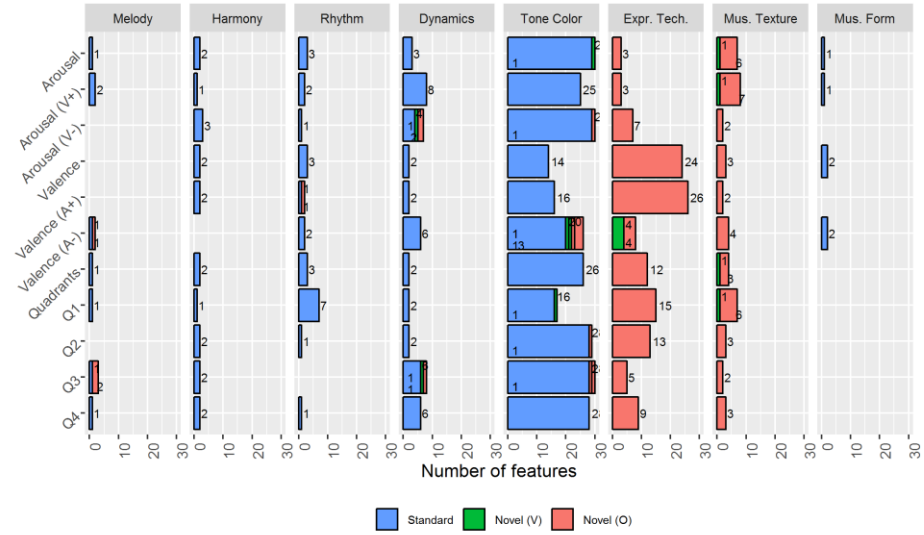


Figure C.17. Distribution of the top50 features per musical dimension (horizontal).

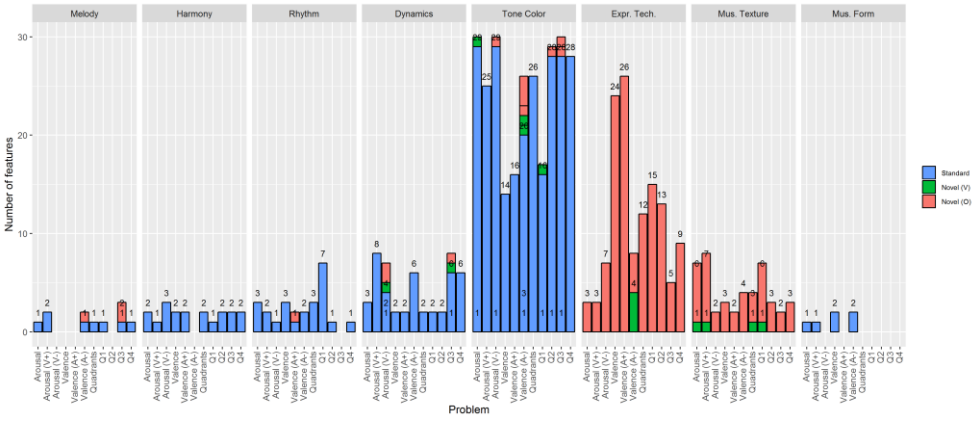


Figure C.18. Distribution of the top50 features per musical dimension (vertical).

E. Distribution of the Top 100 Features

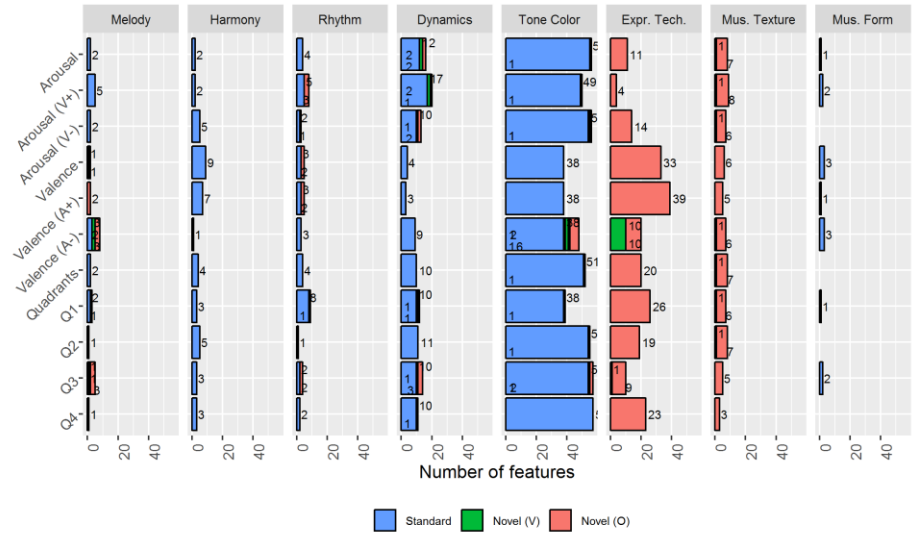


Figure C.19. Distribution of the top100 features per musical dimension (horizontal).

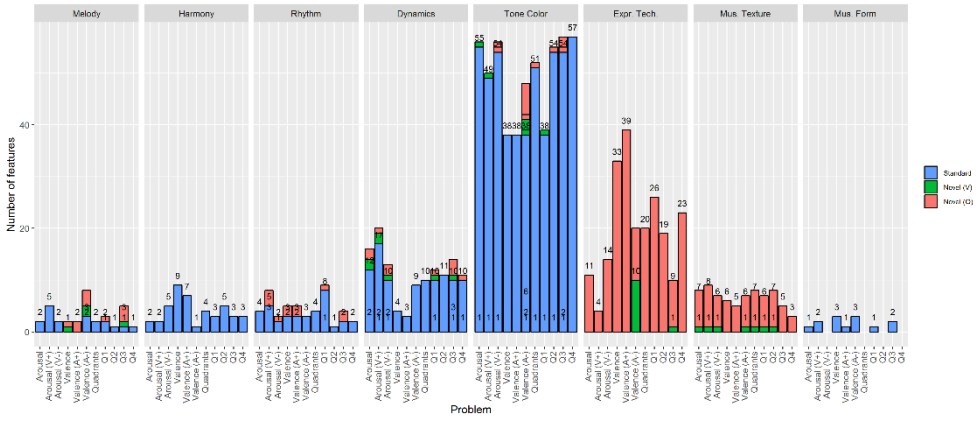


Figure C.20. Distribution of the top100 features per musical dimension (vertical).