Improving Music Mood Annotation Using Polygonal Circular Regression

by

Isabelle Dufour
B.Sc., University of Victoria, 2013

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Improving Music Mood Annotation Using Polygonal Circular Regression

by

Isabelle Dufour
B.Sc., University of Victoria, 2013

Supervisory Committee

Dr. George Tzanetakis, Co-Supervisor
(Department of Computer Science)

Dr. Yvonne Coady, Co-Supervisor
(Department of Computer Science)

Supervisory Committee

Dr. George Tzanetakis, Co-Supervisor
(Department of Computer Science)

Dr. Yvonne Coady, Co-Supervisor
(Department of Computer Science)

ABSTRACT

Music mood recognition by machine continues to attract attention from both academia and industry. This thesis explores the hypothesis that the music emotion problem is circular, and is a primary step in determining the efficacy of circular regression as a machine learning method for automatic music mood recognition. This hypothesis is tested through experiments conducted using instances of the two commonly accepted models of affect used in machine learning (categorical and two-dimensional), as well as on an original circular model proposed by the author. Polygonal approximations of circular regression are proposed as a practical way to investigate whether the circularity of the annotations can be exploited. An original dataset assembled and annotated for the models is also presented. Next, the architecture and implementation choices of all three models are given, with an emphasis on the new polygonal approximations of circular regression. Experiments with different polygons demonstrate consistent and in some cases significant improvements over the categorical model on a dataset containing ambiguous extracts (ones for which the human annotators did not fully agree upon). Through a comprehensive analysis of the results, errors and inconsistencies observed, evidence is provided that mood recognition can be improved if approached as a circular problem. Finally, a proposed multi-tagging strategy based on the circular predictions is put forward as a pragmatic method to automatically annotate music based on the circular model.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

Yvonne Coady and George Tzanetakis for mentoring, support, encouragement, and patience.

Peter van Bodegom, Rachel Dennison and Sondra Moyls for their work in the infancy of this project, including their contributions in building the dataset.

My parents, for encouraging my curiosity and creativity.

My friends, for long, true, and meaningful friendships, worth more than anything.

"There is geometry in the humming of the strings,
there is music in the spacing of the spheres."
Pythagoras

DEDICATION

To my father,
my mother,
and B.

Chapter 1

Introduction

Emotions are part of our daily life. Sometimes in the background, other times with overwhelming power, emotions influence our decisions and reactions, for better or worse. They can be physically observed occurring in the brain through both magnetic resonance imaging (MRI) and positron emission tomography (PET) scans. They can be quantified, analyzed and induced through different levels of neurotransmitters. They have been measured, modelled, analyzed, scrutinized and theorized by philosophers, psychologists, neuroscientists, endocrinologists, sociologists, marketers, historians, musicologists, biologists, criminologists, lawyers, and computer scientists. But emotions still retain some of their mystery, and with all the classical philosophy and modern research on emotion, few ideas have transitioned beyond theory to widely accepted principles.

To make matters even more complicated, emotional perception is to some degree subjective. Encountering a grizzly bear during a hike will probably induce fear in most of us, but looking at kittens playing doesn't necessarily provoke tender feelings in everyone. The emotional response individuals have to art is again, a step further in complexity. Why do colours and forms, or acoustic phenomena organized by humans provoke an emotional response? In considering music, what is the specific arrangement of sound waves that can make one happy, or nostalgic, or sad? Is there a way to understand and master the art of manipulating someone's emotions through sound?

Machine recognition of music emotion has received the attention of numerous researchers over the past fifteen years. Many applications and fields could benefit from efficient systems of mood detection with increases in the capacity of recommendation systems, better curation of immense music libraries, and potential advancements in psychology, neuroscience, and marketing to name a few. The task however is far

from trivial; robust systems require their designers to consider factors from many disciplines including signal processing, machine learning, music theory, psychology, statistics, and linguistics [39].

Applications

The digital era has made it much easier to collect music, and individuals can now gather massive music libraries without the need of an extra room to store it all. Media players offer their users a convenient way to play and organize music through typical database queries on metadata such as artist, album name, genre, tempo in beats per minute (BPM) etc. The ability to create playlists is also a basic feature, allowing the possibility to organize music in a more personal and meaningful way.

Most media players rely on the metadata encoded within the audio file to retrieve information about the song. Basic information such as the name of the artist, song title and album name are usually provided by the music distributor, or can be specified by the user. Research shows that the foremost functions of music are both social and psychological, that most music is created with the intention to convey emotion, and that music always triggers an emotional response [16, 34, 67, 75]. Unfortunately, personal media players do not yet offer the option to browse or organize music based on emotions or mood.

There exists a similar demand from industry to efficiently query their even larger libraries by mood and emotion, whether it is to provide meaningful recommendations to online users, or assist the curators of music libraries for film, advertising and retailers. To the best of my knowledge, the music libraries allowing such queries rely on expert annotators, crowd sourcing, or a mix of both; no system solely relies on the analysis of audio features.

The Problem

Music emotion recognition has been attracting attention from the psychological and Music Information Retrieval (MIR) communities for years. Different models have been put forward by psychologists, but the categorical and two-dimensional models have been favoured by computer scientists developing systems to automatically identify music emotions based on audio features. Both of these models have achieved good results, although they appear to have reached a *glass ceiling*, measured at 65% by Aucouturier and Pachet [53] in their tests to improve the performance of systems

relying on timbral features, over different algorithms, their variants and parameters.

This leads to the following questions: Have we really reached the limits in capabilities of these systems, or just not quite found the best emotional model yet? Providing an emotional model capable of better encompassing the human emotional response to music, could we push this ceiling further using a similar feature space? In this work, I make the following contributions:

- a demonstration of the potential of modelling the music emotion recognition problem as one that is circular
- an original dataset and its annotation process as a means to explore the human perception of emotion conveyed by music
- an exploration of the limits of the two mainly accepted models: the categorical and the two-dimensional
- an approximation to circular regression called *Polygonal Circular Regression*, as a practical way to investigate whether the circularity of the annotations can be exploited.

1.1 Terminology

Let me begin by defining terms that will be used throughout this thesis. In machine learning, classification is the class of problems attempting to correctly identify the category an unlabelled instance belongs to, following a training on a set of labelled examples for each defined category. Categories may be representing precise concepts (for example *Humans* and *Dogs*), or a group or cluster of concepts (for example *Animals* and *Vascular Plants*). Because of its name, the categories of a classification problem are often referred to as *classes*. Throughout this thesis the terms *category*, *cluster* and *class* are used interchangeably.

Music Information Retrieval (MIR) is an interdisciplinary science combining music, computer science, signal processing and cognitive science, with the aim of retrieving information from music, extending the understanding and usefulness of music data. MIR is a broad field of research that includes diverse tasks such as automatic chord recognition, beat detection, audio transcription, instrumentation, genre, composer and emotion recognition among others.

Emotions are said to be shorter lived and more extreme than moods, while moods are said to be less specific and less intense. However, throughout this thesis the terms *emotion* and *mood* are used interchangeably to follow the conventions established in existing literature on the music emotion recognition problem.

Last, it is also useful to clarify that Music Emotion Recognition (MER) systems can refer to any system who's intent is to automatically recognize the moods and emotions of music while Automatic Mood Classification (AMC) specifically refers to MER systems built following the categorical model architecture, treating the problem as a classification problem.

1.2 Thesis Organization

Chapter 1 introduces the problem, its application, and the terminology used throughout the thesis.

Chapter 2 begins with an overview of the different emotional models put forward in psychology, and reviews the state of the art music mood recognition systems.

Chapter 3 reports on the common methodologies chosen by the community when building a dataset, and details the construction and annotation of the dataset used in this work.

Chapter 4 defines the three different models built to perform the investigation, namely the categorical, polygonal circular and two-dimensional models.

Chapter 5 reports on the results of the different models used to conduct this investigation.

Chapter 6 analyzes the results, providing evidence of the circularity of the emotion recognition problem.

Chapter 7 discusses future work required to explore a full circular-linear regression model, in which a mean angular response is predicted from a set of linear variables.

Because part of the subject at hand is music, and to provide the reader with the possibility of auditory examples, two songs from the dataset will be used as case studies. They consist of two thirty second clips extracted from 0:45 to 1:15 of the following songs:

- *Life Round Here* from James Blake (feat. Chance The Rapper)
- *Pursuit of Happiness* from Kid Cudi (Steve Aoki Dance Remix)

They are introduced in Chapter 3, where they first illustrate how human annotators can perceive the moods of the same music differently, based on their background, lifestyle, and musical tastes. They are later used as examples of ground truth in the categorical, circular and two-dimensional annotations. In Chapter 5, their response to all three models is reported, and they are used in Chapter 6 as a basis for discussion.

There is no question about the necessity or demand for efficient music emotion recognition systems. Research in computer science has provided us with powerful computers and several machine learning algorithms. Research in electrical engineering and signal processing produced tools for measuring and analyzing multiple dimensions of acoustic phenomena. Research in psychology and neurology has given us a better understanding of human emotions. Music information retrieval scientists have proposed many models and approaches to the music emotion recognition problem utilizing these findings, but seem to have reached a barrier to expand the capabilities of their systems further.

This thesis presents the idea that human emotional response to music could be further improved by a using a continuous model, capable of better representing the nuances of emotional experience. I propose a continuous circular model, a novel approach to circular regression approximation called polygonal circular regression, and a pragmatic way to automatically annotate music utilizing this method. Comprehensive experiments have yielded strong evidence suggesting the circularity of the music emotion recognition problem, opening a new research path for music information retrieval scientists.

Chapter 2

Previous Work

Music emotion recognition (MER) is an interdisciplinary field with many challenges. Typical MER systems have several common elements, but despite continuous work by the research community over the last two decades, there is no strong consensus on the best choice for each of these elements. There is still no agreement on the best: emotional model to use, algorithm to train, audio features to employ or the best way to combine them. Human emotions have been scrutinized by psychologists, neuroscientists and philosophers, and despite all the theories and ideas put forward, there are still aspects that remain unresolved. The problem doesn't get any easier when music is added to the equation.

There is still no definitive agreement on the best way to approach the music emotion recognition problem. Although psychological literature provides several models of human emotion: discrete, continuous, circular, two and three-dimensional, and digital processing now makes it possible to extract complex audio features, we have yet to find which model best correlates this massive amount of information to the emotional response one has to acoustic phenomena. Despite numerous powerful machine learning algorithms now being readily available, the question remains, how do we teach our machines something we don't quite fully understand ourselves?

The MIR community is left with many possible combinations of models, algorithms and audio features to explore making the evaluation of each approach complex to analyze, and their comparison difficult. Nevertheless, this chapter presents some of the most relevant research on the music emotion recognition problem, beginning with an overview of the commonly accepted emotional models and terminology, followed by the strategies deployed by MER researchers to implement them.

2.1 Emotion Models and Terminology

The dominating methods for modelling emotions in music are categorical and dimensional, representing over 70% of the literature covering music and emotion between 1988 and 2008 according to the comprehensive review on music and emotion studies conducted by Eerola and Vuoskoski [10]. This section explores different examples of these models, their mood terminology and implementation.

2.1.1 Categorical Models

Categorical models follow the idea that human emotions can be grouped into discrete categories, or summarized by a finite number of universal primary emotions (typically including fear, anger, disgust, sadness, and happiness) from which all other emotions can be derived [11, 35, 37, 52, 58]. Unfortunately, authors disagree on which are the primary emotions and how many there actually are.

One of the most renowned categorical models of emotion in the context of music is the adjective checklist proposed by Kate Hevner in 1936 to reduce the burden of subjects asked to annotate music [29]. In this model, illustrated in Figure 2.1, the checklist of sixty-six adjectives used in a previous study [28] is re-organized into eight clusters and presented in a circular manner.

First, Hevner instructed several music annotators to organize a list of adjectives into groups such that all the adjectives of a group were closely related and compatible. Then they were asked to organize their groups of adjectives around an imaginary circle so that for any two adjacent groups, there should be some common characteristic to create a continuum, and opposite groups to be as different as possible.

Her model was later modified by others. First, Farnsworth [12, 13] attempted to improve the consistency within the clusters as well as across them by changing some of the adjectives and reorganizing some of the clusters. It resulted in the addition of a ninth cluster in 1954, then a tenth in 1958, but these modifications were made with disregard to the circularity. In 2003, Schubert [64] revisited the checklist, taking into account some of the proposed changes by Farnsworth, while trying to restore circularity. His proposition was forty-six adjectives, organized in nine clusters.

Hevner's model is categorical, but the organization of the categories shows her awareness of the dimensionality of the problem. One of the advantages of using this model according to Hevner herself, is that the more or less continuous scale accounted for small disagreements amongst annotators, as well as the effect of pre-

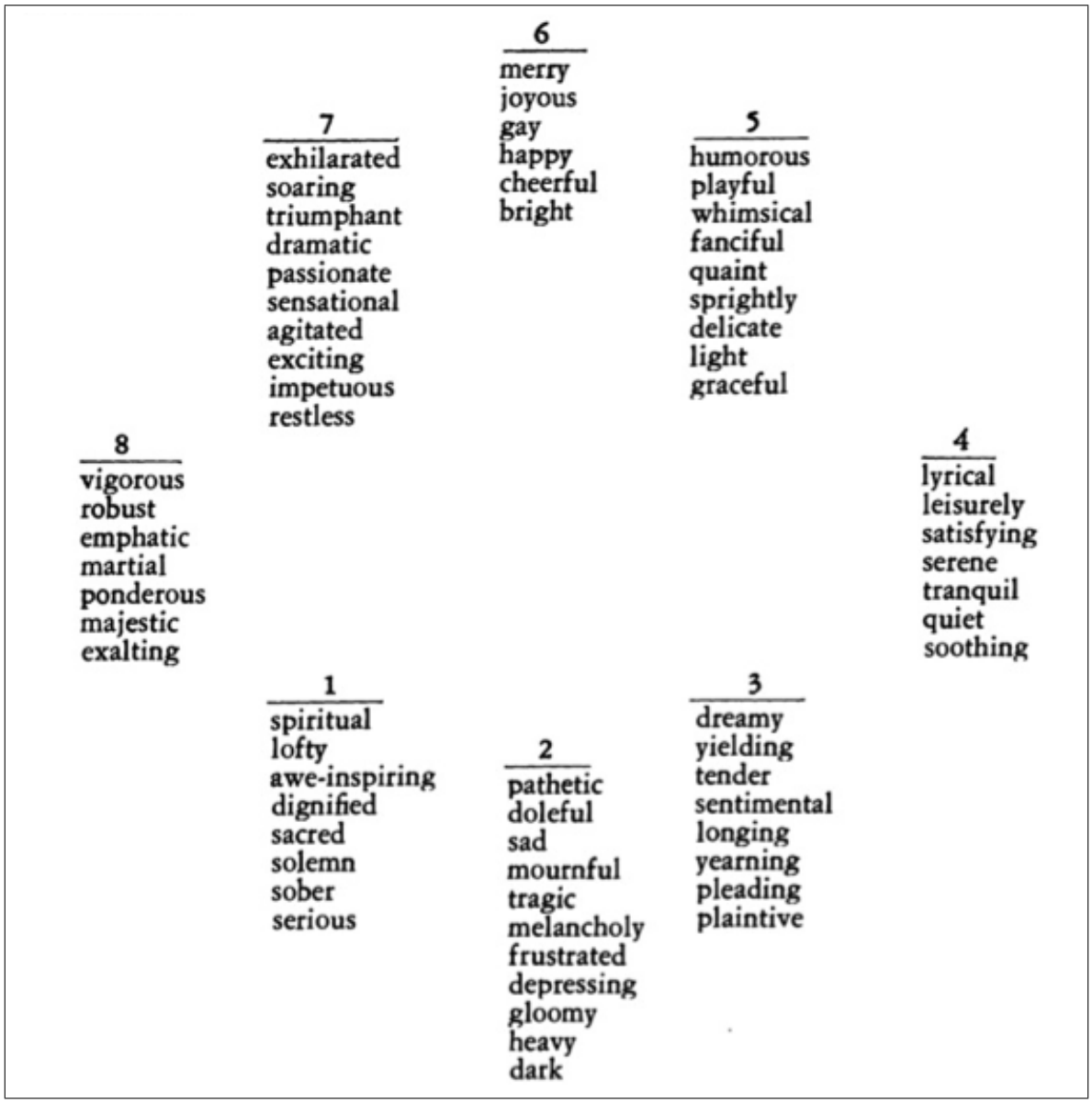

Figure 2.1: Hevner's adjective checklist circle [29].

existing moods or physiological conditions that could have affected the annotators' perceptions. Although Hevner's clusters are highly regarded, it has not been used in its original form by the MIR community.

To this day, there is no consensus on the number of categories to use, or their models [75] when it comes to designing MER systems. This makes comparing models and results difficult, if not nearly impossible. Nevertheless, the community-based framework for the formal evaluation of MIR systems and algorithms, the Music In-

formation Retrieval Evaluation eXchange (MIREX) [8], has an Audio Music Mood Classification (AMC) task regarded as the benchmark by the community since 2007 [33].

Five clusters of moods proposed by Hu and Downie [32] were created by means of statistical analysis of the music mood annotations over three metadata collections (`AllMusicGuide.com`, `epinions.com` and `last.fm`). The resulting clusters shown in Table 2.1 currently serve as categories for the task.

C1	C2	C3	C4	C5
Rousing Rowdy Boisterous Confident Passionate	Rollicking Amiable/ Good-natured Fun Cheerful Sweet	Autumnal Bittersweet Literal Wistful Poignant Brooding	Witty Humorous Whimsical Wry Campy Quirky Silly	Agressive Volatile Fiery Visceral Tense Anxious Intense

Table 2.1: MIREX Mood clusters used in AMC task

The AMC challenge attracts many MIR researchers each year, and several innovative approaches have been put forward. A variety of machine learning techniques have been selected to train classifiers, but most successful systems tend to rely on Support Vector Machines (SVM) [42, 55, 2].

Among the first publications on categorical models is the work of Li and Ogihara [46]. The problem was approached as a multi-label classification problem, where the music extracts are classified into multiple classes, as opposed to mutually exclusive classes. Their research came at a time where such problems were still in their infancy, and hardly any literature and algorithms were available. To achieve the multi-label classification, thirteen binary classifiers were trained on SVMs to determine if a song should receive or not, each of the thirteen labels based on the ten clusters proposed by Farnsworth in 1958 and an extra three clusters they added. The average accuracy of the thirteen classifiers is 67.9%, but the recall and precision measures are overall low.

The same year, Feng, Zhuang and Pan [14] experimented with a simple Back-Propagation (BP) Neural Network classifier, with ten hidden layers and four output nodes to perform a discrete classification. The three inputs of the system are audio features looking at relative tempo ($rTEP$), and both the mean and standard deviation of the Average Silence Ratio ($mASR$ and $vASR$) to model the articulation. The

output of the BP-Neural Network are scores given by the four output nodes associated with four basic moods: *Happiness, Sadness, Anger, Fear*. The investigation was conducted on 353 full length modern popular music pieces. The authors reported a precision of 67% and a recall of 66%. However, no accuracy results were provided, there is no information on the distribution of the dataset, and only 23 of the 353 pieces were used for testing (6.5%), while the remaining 330 was used for training (93.5%).

In 2007, Laurier et al. [42], reached an accuracy of 60.5% on 10-fold cross-validation at the MIREX AMC competition using SVM with the Radial Basis Function (RBF) kernel. To optimize the cost C and the γ parameters, an implementation of the grid search suggested by Hsu et al. [31] was used. This particular step has been incorporated in most of the subsequent MER work employing an RBF kernel on SVM classifiers. Another important contribution came from their error analysis; by reporting the semantical overlap of the MIREX clusters $C2$ and $C4$, as well as the acoustic similarities of $C1$ and $C5$, Laurier foresaw the limits of using the model as a benchmark.

In 2009, Laurier et al. [43] used a similar algorithm on a dataset of 110 fifteen second extracts of movie soundtracks to classify the music into five basic emotions (*Fear, Anger, Happiness, Sadness, Tenderness*), reaching a mean accuracy of 66% on ten runs of 10-fold cross-validation. One important contribution was their demonstration of the strong correlation between audio descriptors such as dissonance, mode, onset rate and loudness with the five clusters using regression models.

The same year, Wack et al. [74] achieved an accuracy of 62.8% at the MIREX AMC task also using SVM with an RBF kernel optimized by performing a grid search, while Cao and Ming reached 65.6% [6] combining an SVM with a Gaussian Super Vector (GSV-SVM), following the sequence kernel approach to speaker and language recognition proposed by Cambell et al. in 2006 [5].

In 2010, Laurier et al. [44] relied on SVM with the optimized RBF kernel, on four categories (*Angry, Happy, Relaxed, Sad*). In this case however, one binary model per category was trained (e.g. angry, not angry), resulting in four distinct models. The average accuracy of the four models is impressive, reaching 90.44%, but it is important to note that a binary class reaches 50% on random classification, and that efforts were made to only include music extracts that clearly belonged to their categories, eliminating any ambiguous extracts. Moreover, their dataset has 1000 thirty second extracts, but the songs were split into four datasets, one for each of the

four models. It results in having only 250 carefully selected extracts used by each model.

In 2012, Panda and Paiva also experimented with the idea of building five different models, but they followed the MIREX clusters and utilized Support Vector Regression (SVR). Using an original dataset of 903 thirty second extracts built to emulate the MIREX dataset, the extracts were then divided in five cluster datasets, each including all of the extracts belonging to the cluster labelled as 1, plus the same amount of extracts coming from other clusters labeled as 0. For example, dataset three included 215 songs belonging to cluster $C3$ labeled as 1, and an additional 215 songs belonging to clusters $C1$, $C2$, $C4$ and $C5$ labeled as 0. Regression was used to measure how much a test song related to each cluster model. The five outputs were combined and the highest regression score determined the final classification. No accuracy measures were provided, but the authors reported an F-measure of 68.9%. It is also interesting to note that the authors achieved the best score at the MIREX competition that year, with an accuracy of 67.8%.

The MIREX results since the beginning of the AMC tasks have slowly progressed from 61.5% obtained by Tzanetakis in 2007 [71] to the 69.5% obtained by Ren, Wu and Jang in 2011 [62]. The latter relied on the usual SVM algorithm, but their submission differed from previous works in utilizing long-term joint frequency features such as acoustic-modulation spectral contrast/valley (AMSC/AMSV), acoustic-modulation spectral flatness measure (AMSFM), and acoustic-modulation spectral crest measure (AMSCM), in addition to the typical audio features. To this day, no one has achieved better results at the MIREX AMC. Although less popular, other algorithms such as Gaussian mixture models [59, 47] have provided good results.

Unfortunately, the subjective nature of emotional perception makes the categorical models both difficult to define and evaluate [76]. Consensus among people is somewhat rare when it comes to the perception of emotion conveyed by music, and reaching agreement among the annotators building the datasets if often problematic [33]. It results in a number of songs and music being rejected from those datasets as it is impossible to assign them to a category, and they are thus ignored by the AMC systems. The lack of consensus on a precise categorical model can be seen both as a symptom and an explanation for its relative stagnation; if people can't agree on how to categorize emotions, how could computers? These weakness of categorical models continue to motivate researchers to find more representative approaches, and the most utilized alternatives are the dimensional models.

2.1.2 Dimensional Models

Dimensional models are based on the proposition that moods can be modelled by continuous descriptors, or multi-dimensional metrics. For the music emotion recognition problem, the dimensional models are typically used to evaluate the correlation of audio features and emotional response, or are translated into a classification problem to make predictions. The most commonly used dimensional model by the MIR community is the two-dimensional valence and arousal (VA) model proposed by Russell in 1980 [63] as the circumplex model, illustrated in Figure 2.2.

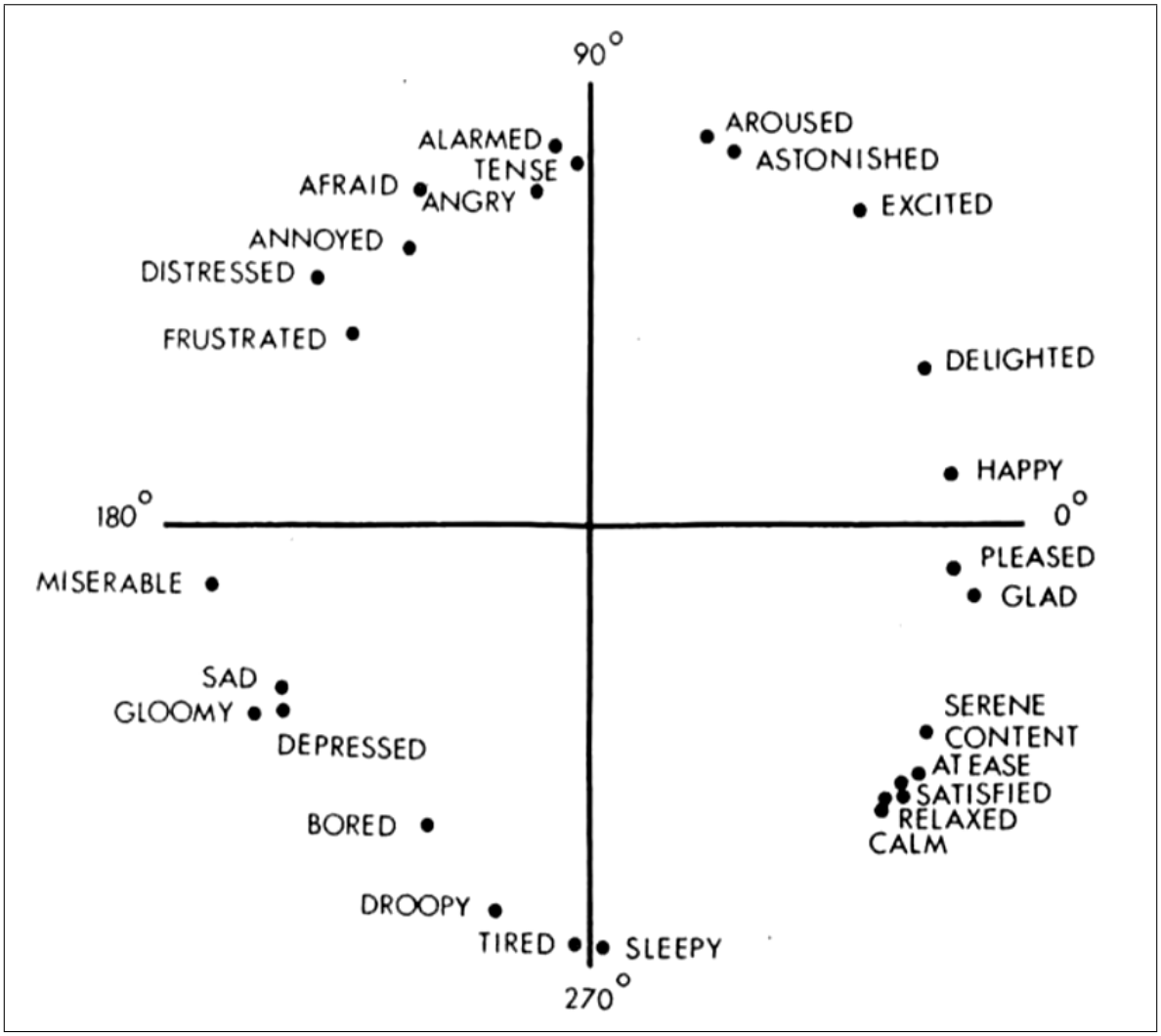

Figure 2.2: The circumplex model as proposed by Russell in 1980 [63].

The valence axis (x axis on figure 2.2) is used to represent the notion of negative vs. positive emotion, while the *Arousal* scale (y axis) measures the level of stimulation.

Systems based on this model typically build two regression models (regressors), one per dimension, and either label a song with the two values, attempt to position the song on the plane and perform clustering, or utilize the four quadrants of the two-dimensional model into categories, treating the MER problem as a categorical problem.

Another two-dimensional model based on similar axes and often used by the MIR community is Thayer's model [68], shown in Figure 2.3, where the axes are defined as *Stress* and *Energy*. This differs from Russell's model as both axes are looking at arousal, one as an energetic arousal, the other as a tense arousal. According to Thayer, valence can be expressed as a combination of energy and tension.

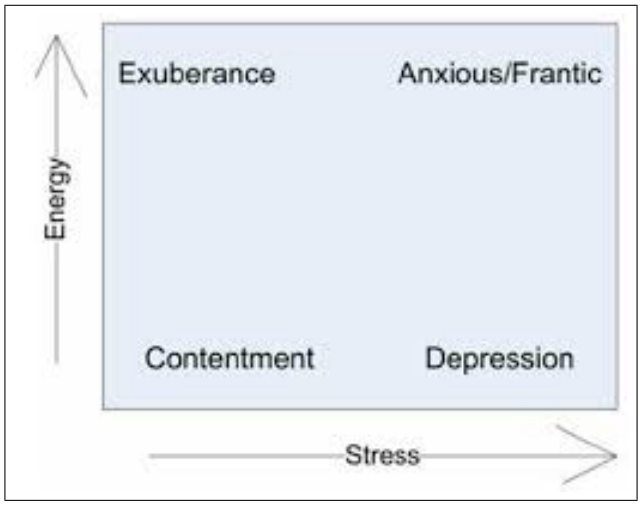

Figure 2.3: Thayer's mood model, as as illustrated by Trohidis et al. [69].

One of the first publications utilizing a two-dimensional model was the 2006 work of Lu, Lui and Zhang [47] where Thayer's model is used to define four categories, and the problem is approached as a classification one. They were the first to bring attention to the potential relevance of the dimensional models put forward in psychological research. Using 800 expertly annotated extracts from 250 classical and romantic pieces, a hierarchical framework of gaussian mixture models (GMM) was used to classify music into one of the four quadrants defined as *Contentment, Depression Exuberance, Anxious/Frantic*. A first classification is made using the intensity feature to separate clips into two groups. Next, timbre and rhythm are analyzed through their respective GMM and the outputs are combined to separate *Contentment* from *Depression* for group 1, and *Exuberance* from *Anxious/Frantic* for group 2. The accuracy reached was 86.3%, but it should be noted that several extracts are

used from the same songs to build the dataset, potentiality overfitting the system.

In 2007, MacDorman et al. [48] trained two regression models independently to predict the pleasure and arousal response to music. Eighty-five participants were asked to rate six second extracts taken from a hundred songs. Each extract was rated on eight different seven point scales representing pleasure (*happy-unhappy, pleased-annoyed, satisfied-unsatisfied, positive-negative*) and arousal (*stimulated-relaxed, excited-calm, fenzied-sluggish, active-passive*). Their study found that the standard deviation of the arousal dimension was much higher than for the pleasure dimension. They also found that the arousal regression model was better at representing the variation among the participants' ratings, and more highly correlated with music features (e.g. tempo and loudness) than the pleasure model.

A year later, Yang et al. [76] also trained an independent regression model for each of the valence and arousal dimensions, with the intention of providing potential library users with an interface to choose a point on the two-dimensional plane as a way to form a query to work around the terminology problem. Two-hundred and fifty-three volunteers were asked to rate subsets of their 195 twenty-five second extracts on two (valence and arousal) eleven point scales. The average of the annotators is used as the ground truth for support vector machines used as regressors. The R^2 statistics reached 58.3% for the arousal model, and 28.1% for the valence.

In 2009, Han et al. [25] also experimented with Support Vector Regression (SVR) with eleven categories placed over the four quadrants of the two-dimensional valence arousal (VA) plane, using the central point of each category on the plane as their ground truth. Two representations of the central point were used to create two versions of the ground truth: cartesian coordinates (valence, arousal), and polar coordinates (distance, angle). The dataset is built out of 165 songs (fifteen for each of the eleven categories) from the `allmusic.com` database. They obtained accuracies of 63.03% using their cartesian coordinates, and an impressive 94.55% utilizing the polar coordinates. The authors report testing on v-fold cross-validation with different values of v, but do not provide specific values. There is also no indication whether the results were combined for different values of v, or if they only presented the ones for which the best results were obtained.

In 2011, Panda and Paiva [55] proposed a system to track emotion over time in music using SVMs. For this work, the authors used the dataset built by Yang et al. [76] in 2008, selecting twenty-nine full songs for testing, based on the 189 twenty-five second extracts. The regression predictions on 1.5 second windows of a song are

used to classify it into one of the four quadrants of Thayer's emotional model. They obtained an accuracy 56.3%, measuring the matching ratio between predictions and annotations for full songs.

In 2013, Panda et al. [54] added melodic features to the standard audio features increasing the R^2 statistics of the valence dimension from 35.2% to 40.6%, and from 63.2% to 67.4% for the arousal dimension. The authors again chose to work with Yang's dataset. Ninety-eight melodic features derived from pitch and duration, vibrato and contour features served as melodic descriptors. They reported that melodic features alone gave lower results than the standard audio features, but the combination of the two gave the best results.

2.2 Audio Features

Empirical studies on emotions conveyed by music have been conducted for decades. The compilation and analysis of the notes taken by twenty-one people on their impressions of music played at a recital were published by Downey in 1897 [7] and are considered a pioneering work on the subject. How musical features specifically affected the emotional response became of interest a few years later. In 1932, Gundlach published one such work, looking at the traditional music of several indigenous North American tribes [23], and how pitch, range, speed, type of interval (minor and major 3^{rds}, intervals $< 3^{rds}$, and intervals $> 3^{rds}$), and type of rhythm relate to the emotions conveyed by the music. The study concluded that while rhythm and tempo impart the dynamic characteristics of mood, the other measurements did not provide simple correlations with emotion for this particular style of music, as they varied too greatly between the tribes. Hevner studied the effects of major and minor modes modes [27] as well as pitch and tempo [30] on emotion. In the subsequent years, there were several researchers continuing this work and conducting similar studies, exploring how different musical features correlate to perceived emotions and in 2008, Frieberg compiled the musical features that were found to be useful for music emotion recognition [18]:

- Timing - Tempo, tempo variation, duration contrast
- Dynamics: overall level, crescendo/decrescendo, accents
- Articulation: overall (staccato/legato), variability

- Timbre: Spectral richness, harmonic richness, onset velocity
- Pitch (high/low)
- Interval (small/large)
- Melody: range (small/large), direction (up/down)
- Harmony (consonant/complex-dissonant)
- Tonality (chromatic-atonal/key-oriented)
- Rhythm (regular-smooth/firm/flowing-fluent/irregular-rough)

Three more musical features reported by Meyers [51] are often added to the list [55, 56, 57]:

- Mode (major/minor)
- Loudness (high/low)
- Musical form (complexity, repletion, new ideas, disruption)

Unfortunately, not all of these musical features can be easily extracted using audio signal analysis. Moreover, no one knows precisely how they interact with each other. For example, one may hypothesize that an emotion such as *Aggressive* implies a fairly fast tempo, but there are several examples of aggressive music that are rather slow (think of the chorus of *I'm Afraid of Americans* by David Bowie, or *In Your Face* from Die Antwoord). This may explain why exploratory works on audio features in emotion recognition tends to confirm that a combination of different groups of features consistently gives better results than using only one [43, 48, 54]. On the other hand, using a large number of features makes for a high dimensional feature space, requiring large datasets and complex optimization.

Because we are still unsure of the best emotional model to define the music emotion recognition problem, the debate on the best audio features to use is still open. Nevertheless, some features have consistently provided good results for both categorical and dimensional models. These are referred to as *standard audio features* across the MER literature. These include many audio features (MFCC, centroid, flux, roll-off, tempo, loudness, chromes, tonality etc.), represented by different statistical moments. Some of the most recurring features and measures are briefly described next, but it is by no means an exhaustive list of the audio features used by MER systems.

2.2.1 Spectral Features

The Discrete Fourier Transform (DFT) provides a powerful tool to analyze the frequency components of a song. It provides a mathematical representation of a given time period of a sound by measuring the amplitudes (power) of each of the frequency bins (a range of frequency defined by the parameters of the DFT). Of course, for a DFT to have meaning, it has to be calculated over a short period of time (typically 10 to 20 ms.); taking the DFT of a whole song would report on the sum of all frequencies and amplitudes of the entire song. That is why multiple short-time Fourier Transforms (STFT) are often preferred. STFTs are performed at every s amount of samples, and their results are typically presented in a #of bins by $s/sampleRate$ matrix, and can be represented visually by a spectrogram. This gives us information on how the spectrum changes over time. Of course, using a series of STFTs to examine the frequency content over time is much more meaningful when analyzing music, but it requires a lot of memory without providing easily comparable representations from one song to another, making them poor choices as features. Fortunately, there are compact ways to represent and describe different aspects of the spectrum without having to use the entire matrix.

Mel Frequency Cepstral Coefficients (MFCC): the cepstrum is the Discrete Cosine Transform (DCT) of the logarithm of the spectrum, calculated on the mel band (linear below 1000 Hz, logarithmic above.). It is probably the most utilized audio feature as it is integral to speech recognition and many of the MIR tasks. DFTs are over linearly-spaced frequency, but human perception of frequencies is logarithmic above a certain frequency, therefore several scales have been put forward to represent the phenomena, the Mel-scale being one of them. The scale uses thirteen linearly-spaced filters and twenty-seven log-spaced filters, for a total of forty. This filtering reduces the spectrum's numerical representation by reducing the number of frequency bins to forty, mapping the powers of the spectrum onto the mel-scale and generating the mel-frequency spectrum. To get the coefficient of this spectrum, the logs of the powers at each mel-frequency are taken before a Discrete Cosine Transform (DCT) is performed to further reduce the dimensionality of the representation. The amplitudes of the resulting spectrum (called the cepstrum) are the MFCCs. Typically, thirteen or twenty coefficients are kept to represent the sound. The cepstrum allow us to measure the periodicity of the frequency response of the sound. Loosely

speaking, it is the spectrum of a spectrum, or a measure of the frequency of frequencies.

Spectral Centroid: Is best envisioned as the centre of gravity of the spectrum and is calculated by taking the mean of the weighted frequencies by their amplitude. It is also seen as the spectrum distribution and correlates with pitch and brightness of sound. The spectral centroid, along with the roll-off and flux, are the three spectral features attributed to the outcome of Grey's work on musical timbre [20, 21, 22].

Spectral Roll-off: The frequency below which 80 to 90% (depending on the implementation) of the signal energy is contained. Shows the frequency distribution between high and low frequencies.

Spectral Flux: Shows how the spectrum changes across time.

Spectral Spread: Defines how the spectrum spreads around its mean value. Can be seen as the variance of the centroid.

Spectral Skewness: Measures the asymmetry of a distribution around the mean (centroid).

Spectral Kurtosis: Measures the flatness/peakness of the spectrum distribution.

Spectral Decrease: Correlated to human perception, represents the amount of decrease of the spectral amplitude.

Pitch Histogram: It is possible to retrieve the pitch of the frequencies for which strong energy is present in the DFT. Direct frequency to pitch conversions can be made. Different frequency bins, mapping to the same pitch class (e.g. the C4 and C5 midi notes) can be combined in order to retain only the twelve pitches corresponding the chromatic scale over one octave.

Chroma: A vector representing the sum of energy at each of the frequencies associated to the twelve semi-tones of the chromatic scale.

Barkbands: Scale to approximate human auditory system. Can be used to calculate the spectral energy at each of the 27 Barkbands, and summed.

Temporal Summarization: Because sound and music happen over time, several numerical descriptors of the spectral feature are necessary for a meaningful representation. Considering that most Digital Signal Processing (DSP) is performed on short timeframes of sound (10-20 ms.), they are often summarized over a larger portion of time. Several methods are used, including statistical moments such as calculating the mean, standard deviation and kurtosis of these features over larger time scales (around 1-3 seconds). These longer segments of sounds have been termed texture windows [70].

2.2.2 Rhythmic Features

Beat Per Minute (BPM): Average tempo in terms of the number of beat per minute.

Zero-crossing rate: Number of times the signal goes from a positive to negative energy. Often used to measure the level of noise, since harmonic signals have lower zero-crossing values than noise.

Onset rate: The number of time a peak in the envelope is detected per second.

Beat Histograms: A representation of the rhythm over time, measuring the frequency of a tempo in a song. Good representation of the variability and strength of the tempo over time.

2.2.3 Dynamic Features

Root Mean Square (RMS) Energy: Measure the mean power or energy of a sound over a period of time.

2.2.4 Audio Frameworks

Most of the audio features used by the MER systems reviewed in this thesis were extracted with one, or a combination of the three main audio frameworks developed by and for the MIR community.

Marsyas: Marsyas stands for Music Analysis, Retrieval and Synthesis for Audio Signals. The open source audio framework was developed in C++ with the specific

goal to provide flexible and fast tools for audio analysis and synthesis for music information retrieval. Marsyas was originally designed and implemented by Tzanetakis [72], and later extended by many contributors since its first release.

MIRtoolbox: A Matlab library, the MIRtoolbox is a modular framework for the extraction of audio features that are musically-related, such as timbre, tonality, rhythm and form [41]. It offers a flexible architecture, breaking algorithms into blocks that can be organized to support the specific needs of its user. Contrary to Marsyas, the MIRtoolbox can't be used for real-time applications.

PsySound: PsySound, now in its third release (PsySound3) is another Matlab package, but it is also available as a compiled standalone version [4]. The software offers acoustical analysis methods such as Fourier and Hilbert transforms, cepstrum and auto-correlation. It also provides psychoacoustical models for dynamic loudness, sharpness, roughness, fluctuation, pitch height and strengths.

2.3 Summary

Much progress has been made since Downey's pioneering work in 1897 [7]. Emotional models have been proposed, musical features affecting the emotional response to music identified, signal processing tools to extract some of these features developed along with audio frameworks to easily extract them, and a multitude of powerful machine learning algorithms have been implemented. This progress and their combination are constantly being used to improve the capacity of MER systems. However, as is the case for any machine learning problem, building intelligent MER systems requires a solid ground truth for training and testing. The construction of datasets for MER systems is far from trivial, many key decisions need to be made. The next chapter briefly provides examples on how MIR researchers gather datasets, before detailing how the original dataset used for this thesis was assembled and annotated.

Chapter 3

Building and Annotating a Dataset

One of the challenges of the music mood recognition problem, is the difficulty in finding readily available datasets. Audio recordings are protected by copyright law, which prevents researchers in the field from sharing complete datasets; the mood annotations and features may be shared as data, but the audio files cannot. To assure consistency when using someone else's dataset, one would have to confirm that the artist, version, recording and format are identical to the ones listed. Moreover, because there is no clear consensus on mood emotion recognition research methodology, datasets utilizing the same music track may in fact look at different portions of the track, use a different model type (categorical vs. dimensional) and even different mood terminology.

These problems also exist within the same type of model. For example, the number of categories used in the categorical models can differ greatly; Laurier et al. [44], Lu, Liu and Zhang [47] as well as Feng, Zhuang and Pan [15] all use four categories, while Laurier et al. [43] uses five, Trohidis et al. [69] chose to use six, Skowronek et al. [65, 66] twelve, and Li and Ogihara [46] opted for thirteen (see Table 3.1). To complicate things further, there is no widely accepted annotation lexicon, and even in cases where the number of categories is the same, the mood terminology usually differs. For example, Laurier et al. [44], Lu, Liu and Zhang [47] , and Feng, Zhuang and Pan [14] may share the same number of categories but Laurier et al. defined theirs as *Angry, Happy, Relaxed, Sad*, Feng, Zhuang and Pan used *Anger, Happiness, Fear, Sadness*, while Lu, Liu and Zhang chose four basic emotions based on the two–dimensional model: *Contentment, Depression, Exuberance, Anxious/Frantic* and manually mapped multiple additional terms gathered from `AllMusic.com` to create clusters of mood terms.

Authors	**# of moods**	**# of songs**	**Genre**	**Annotators**	**Length**	**Portion used**
Feng et al. [15]	4	353	pop	N/A	full songs	full songs
Laurier et al. [44]	4	4x250	N/A	17 + Last.fm	30 sec	N/A
Lu et al. [47]	4	800/250	classical	3	20 sec.	multiple
Bischoff et al.[3]	4 & 5	1000	various	Allmusic.com Last.fm	30 sec	N/A
Hu et. al [33] MIREX dataset	5	600	various	3 (2 or 3/song)	30 sec	middle
Laurier et al.[43]	5	110	soundtrack	116	15 sec	N/A
Panda et al.[56, 57]	5	903	N/A	Allmusic.com	30 sec.	N/A
Trohidis et al.[69]	6	593	various	3	30 sec.	0:30 - 1:00
Han et al.[25]	11	165	pop	Allmusic.com	N/A	N/A
Skowronek et al.[65, 66]	12	1059	various	12 (6/song)	20 sec.	middle
Li et al.[46]	13	499	various	1	30 sec.	0:30 - 1:00
Korhonen et al.[40]	2D	6	classical	35	full songs	full songs
Yang et al.[76]	2D	195	pop	253 (10/song)	25 sec.	mostly chorus
Panda et al.[54] (from Yang[76])	2D	189	pop	253 (10/song)	25 sec.	mostly chorus
MacDorman et al.[48]	2D	100	various	85	6 sec.	1:30 - 2:00
Kim et al.[38]	2D	446	various	10	20 sec.	hand picked
Eerola et al.[9]	5 cat. & 3D	110	soundtrack	116	10 to 30 sec.	hand picked

Table 3.1: Literature examples of datasets design.

Another question that remains to be answered, is which portion of a song should be used in building a MER dataset. Although not as critical as in other MIR tasks such as chord recognition or beat detection, researchers have to be careful to consistently use the same portion of a music track, as the emotions conveyed by a piece of music can greatly vary over time. Think of the emotional journey of *Bohemian Rhapsody* from the British band Queen, unless the mood annotations are associated with a precise segment of the song, there is no way to assure consistency across datasets.

Most work on audio mood recognition has adopted the thirty-second segment format [69, 74, 42, 56, 33, 44, 3], but extracts of six seconds were used by MacDorman and Ho [48], a length of fifteen seconds was chosen by Laurier et al. [43], several researchers have opted for twenty seconds [65, 47, 38] while some have extracts of lengths varying from ten to thirty seconds [9]. Finally, there is still no convention

on which segments should be used, and authors do not always specify their choices. From those who do, we learn that Hu et al. [33] and Skowronek et al. [65, 66] chose to use the middle of their songs, Trohidis et al. [69] extracted their segments after the initial thirty seconds, and Yang et al. [76] preferred to use the chorus.

The most notable effort to create standards for the categorization models came with the construction of the dataset used in the benchmark AMC MIREX task. In 2007, Hu and Downie [32] first proposed a set of terms organized in five clusters based on the statistical analysis of music moods over three metadata collections (`AllMusicGuide.com`, `epinions.com` and `last.fm`). Their final categories can be seen in Table 3.2, and further details are given in Chapter 2. In 2008, Hu et al. [33] suggested guidelines for building the AMC dedicated dataset, including using thirty second extracts of diverse music genres, as well as asking annotators to ignore lyrics and providing them with exemplar tracks for each category. Unfortunately, the actual dataset remains secret, as it is used as a benchmark, so researchers are left with the choice of either trying to reproduce a comparable dataset, or building their own from scratch.

Considering that no dataset is readily available and widely adopted by my peers, I decided to create my own dataset to conduct this investigation. In doing so, I benefited from the insight of conducting the human annotation process. Additionally, having the latitude of designing several models (categorical, dimensional and circular) on the same dataset provided a meaningful comparison of the results across the models. Finally, creating my own dataset allowed ready access when analyzing the results and errors.

3.1 Data Acquisition

For the purposes of comparison to the MIREX results, a categorical model using five categories was first built. Considering that gathering roughly a hundred songs per category would provide enough data for both training and testing the algorithms on 10-fold cross-validation, nearly six-hundred songs were originally selected from the investigators' personal libraries. Building a dataset that included various music genres was also important to emulate the MIREX dataset, and music from all genres (pop, disco, soul, rock, jazz, electronic, hip-hop, classical, dance, heavy metal, contemporary, reggae, country, latino, traditional etc.) was included. Because genre classification is a problem in itself and the songs did not necessarily come with any

genre metadata, no specific statistics on genre distribution are presented in this work.

The dataset includes music both with and without lyrics. The majority of the songs with lyrics are in English, but a significant portion are in different languages including French, Spanish, Portuguese, Italian, Afrikaans, German, Bulgarian and Japanese.

Considering a song's mood can vary over time and following the recommendations in made by Hu et al. [33] for the AMC MIREX dataset, a thirty second segment was taken from each of the songs. In an attempt to capture a representative part of the full track and considering the tracks were originally of varying length, it was decided that the segment would be extracted from time 0:45 to 1:15. Each extract was verified to ensure they did not contained big musical, mood changes, or long silences and in the few cases where the originally selected thirty seconds was problematic, the extracts were replaced by more appropriate segments of the same song. The same thirty second extract was used in both the annotation of the ground truth and the signal analysis of each song.

3.2 Ground Truth Annotations

Originally, the dataset was annotated by individuals with the sole intention to build one categorical model. Later, when the hypothesis of circularity of the mood recognition problem was put forward, it became apparent that the annotation system would have to be redesigned. In addition to the necessity of a new set of annotations to accommodate the circular model, building a two-dimensional (valence/arousal) model to provide depth to the analysis of the results seemed reasonable, although this also required its own annotation system. I faced the challenge of re-annotating the entire dataset in a way that preserved the annotators original intentions, without having the time-consuming task or finding all of my initial annotators and convincing them to voluntarily annotate the dataset once again, this time with two completely different systems.

For clarity, all annotation systems and the methodology followed to create the alternate annotations are presented in this section. Precisely, the terminology and original annotations made by the volunteers for the categorical model is detailed in Section 3.2.1. The methodology followed to transform the categorical model into the circular model, along with the mathematical transformation of the original classification annotations into circular regression scores are explained in Section 3.2.2.

Finally, the transformation of the circular annotations into coordinates utilized by the two-dimensional model is explained in Section 3.2.3.

3.2.1 Categorical Annotation

A number of problems with the five mood clusters designed for the MIREX competition have been noted by researchers, including the semantic overlap between clusters two ($C2$) and four ($C4$) creating ambiguity, as well as the acoustic similarities between clusters one ($C1$) and five ($C5$) first reported by Laurier et al. [42]. Another observation was made: important mood terms and emotional dimensions are missing from these five clusters. It is interesting to note that these clusters were derived from the popular set (Top Songs, Top Albums); music expressing strong emotions, whether positive or negative, might be leaving a greater impression on the listener, and be more likely to be memorable, a key factor to popularity. This could explain why terms associated with low arousal and neutral valence, such as *Pensive* and *Tender*, are missing from those clusters. Based on these observations, the decision to modify the five MIREX mood clusters was made. For comparison, the MIREX mood clusters are presented in table 3.2.

C1	C2	C3	C4	C5
Rousing Rowdy Boisterous Confident Passionate	Rollicking Amiable/ Good-natured Fun Cheerful Sweet	Autumnal Bittersweet Literal Wistful Poignant Brooding	Witty Humorous Whimsical Wry Campy Quirky Silly	Agressive Volatile Fiery Visceral Tense Anxious Intense

Table 3.2: MIREX Mood clusters used in AMC task

First, some of the moods from $C4$ were incorporated in $C2$ to address the semantic overlap. Some of the mood terms were completely eliminated if an acceptable synonym was already in $C2$ in order to keep the number of terms at around eight per cluster. With $C4$ empty, it became possible to refine $C5$ as well as make space in $C3$ for important missing terms such as *Tender* and *Pensive*. The clusters were redesigned loosely following Hevner's continuity idea [29]. The final modified mood clusters used for the categorical model can be seen in Table 3.3.

Happy			Sad	Mad
C1	C2	C3	C4	C5
Rousing Rowdy Boisterous Thrilling Epic Exhilarated Exalted Ecstatic Spirited	Rollicking Fun Cheerful Sweet Sprightly Summery Playful Flirty	Autumnal Bittersweet Wistful Nostalgic Sentimental Tender Pensive Regretful	Poignant Brooding Melancholic Mournful Tragic Gloomy Dark Creepy Paranoid	Agressive Volatile Fiery Threatening Hostile Belligerent Arrogant Angry

Table 3.3: Mood Classes/Clusters used for the annotation of the ground truth for the categorical model

Human Annotation

Five volunteer annotators were asked to classify the entire dataset, and an additional seven annotators classified different subsets in order to get exactly eight votes for each thirty second clip. The twelve annotators were between the ages of twenty and forty, came from different cultures, backgrounds, and three were non-native english speakers (French, Slovenian and Spanish). Annotators were asked to ignore the meaning of the lyrics, and choose which of the five clusters best represented the intended emotion or mood of the music. In other words, we did not want to know what emotion they felt listening to the extract, but rather, what they perceived to be the intention of the musician. To clarify our intention, we asked them to see this annotation task as the curation of a soundtrack library for movies, and exemplary extracts were given. When all the extracts had exactly eight votes, the annotations were combined by assigning the class for which the majority of the annotators agreed upon. A few examples of the annotation results and final classification decisions are shown in Table 3.4.

Although the annotators mostly agreed on the majority of the clips, careful examination was given to each clip by the human compiler to assure an accurate final classification. Twenty-seven of the songs were classified with such disagreement that it was impossible to reconcile the votes to one class, as illustrated by the case of *Beat It - Michael Jackson* shown in Table 3.4. These clips were eliminated from the dataset, leaving a total of 564 clips.

Unanimity among the annotations was reached for only 127 clips (22.5%), and an agreement among six or more annotators was reached for 417 clips (73.9%), leav-

Audio Clip	**C1**	**C2**	**C3**	**C4**	**C5**	**GT**
Don't Come Home A Drinkin - Loretta Lynn	0	7	1	0	0	2
Bambino - Plastic Bertrand	6	2	0	0	0	1
Get Lucky - Daft Punk feat. Pharrell Williams	0	8	0	0	0	2
Life Round Here (feat. Chance The Rapper) - James Blake	0	0	5	3	0	3
Motion Picture Soundtrack - Radiohead	0	0	0	8	0	4
Rite of Spring - Glorification of the Chosen One - Igor Stravinsky	0	0	0	2	6	5
Beat It - Michael Jackson	3	2	0	1	2	X

Table 3.4: Example annotations and resulting ground truth classes (GT) based on eight annotators.

.

ing 147 clips (26.1%) as ambiguous (see Table 3.5). Because eliminating ambiguous extracts for categorical models is an established practice throughout the literature, the decision to create two datasets was made, so the results of this categorical model could be compared to the results of others. These two datasets are referred to as:

- the *full dataset* including all of the 564 clips
- the *unambiguous dataset*, only including the 417 clips for which six or more annotators agreed upon.

Further investigation revealed that half or more of the annotators disagreed for 94 clips (16.7%) (excluding the initial discarded ones); in 58 of those cases (10.3%), the annotations were equally split between two neighbouring classes. In these instances, the compiler made an executive decision in favour of one of the two classes.

Annotators' Agreement	***#clips*/564**	**% of clips**
Unanimity 8/8	127	22.5%
Strong: $\geqslant 6/8$	417	73.9%
Weak: $\leqslant 4/8$	94	16.7%
Equally split between two neighbouring classes	58	10.3%

Table 3.5: Agreement statistics of eight annotators on the full dataset.

3.2.2 Circular Annotation

To transpose the classification annotations to a circular system, two steps were required. First, the mood clusters were wrapped around the circle following the circumplex model [63], and the mood terms were ordered in a way they could be regarded as a gradual continuum. This reorganization of the mood terms was performed by the five annotators who had previously labelled the entire dataset for the categorical model. The choice of mood terms and their distribution around the circle was based on synonym proximity using both `Thesaurus.com` [1] and the `AllMusic.com` mood similarity overview.

To wrap the clusters around the circle, each cluster was first flattened onto a line of range $[((c-1)+0.626),(c+0.5)]$, where c is the label number of the cluster from the categorical model. Eight terms of each cluster were ordered on the line to create a continuum from the previous neighbouring cluster, to the next neighbouring cluster, with the central cluster mood term at value c. Each of the terms was given a value following equation 3.1

$$termValue_i = termValue_1 + 0.125(i-1); \quad for \quad i = \{1, 2, ..., 8\} \tag{3.1}$$

where $termValue_1 = (c-1)+0.625$. For example, the eight terms of $C3$ were flattened on a line of range $[2.625, 3.5]$. The five lines were then appended and wrapped around a circle. The final circular model is shown in Figure 3.1.

The human annotations gathered for the categorical model (see Section 3.2.1) were translated to match the circular model. Each music clip was given a value in the range of $[0.626, 5.5]$ representing the mean classification of our annotators, and served as the ground truth dependant variables for the regression model. This allowed for an annotation that represented all of the annotators inputs, as opposed to only considering the opinion of the majority as was done in the categorical model. It also had the advantage of disambiguating extracts where annotators were equally split, by allowing an annotation that sits exactly at the boundary of two clusters. Examples of circular annotation on our two case studies are shown in Table 3.6, and illustrated in Figure 3.1.

Audio Clip	C1	C2	C3	C4	C5	GT	Reg
Life Round Here (feat. Chance The Rapper) - James Blake	0	0	5	3	0	3	3.375
Pursuit of Happiness (Steve Aoki Dance Remix) - Kid Cudi	4	0	0	0	4	1	5.5

Table 3.6: Circular regression annotation on the two case studies

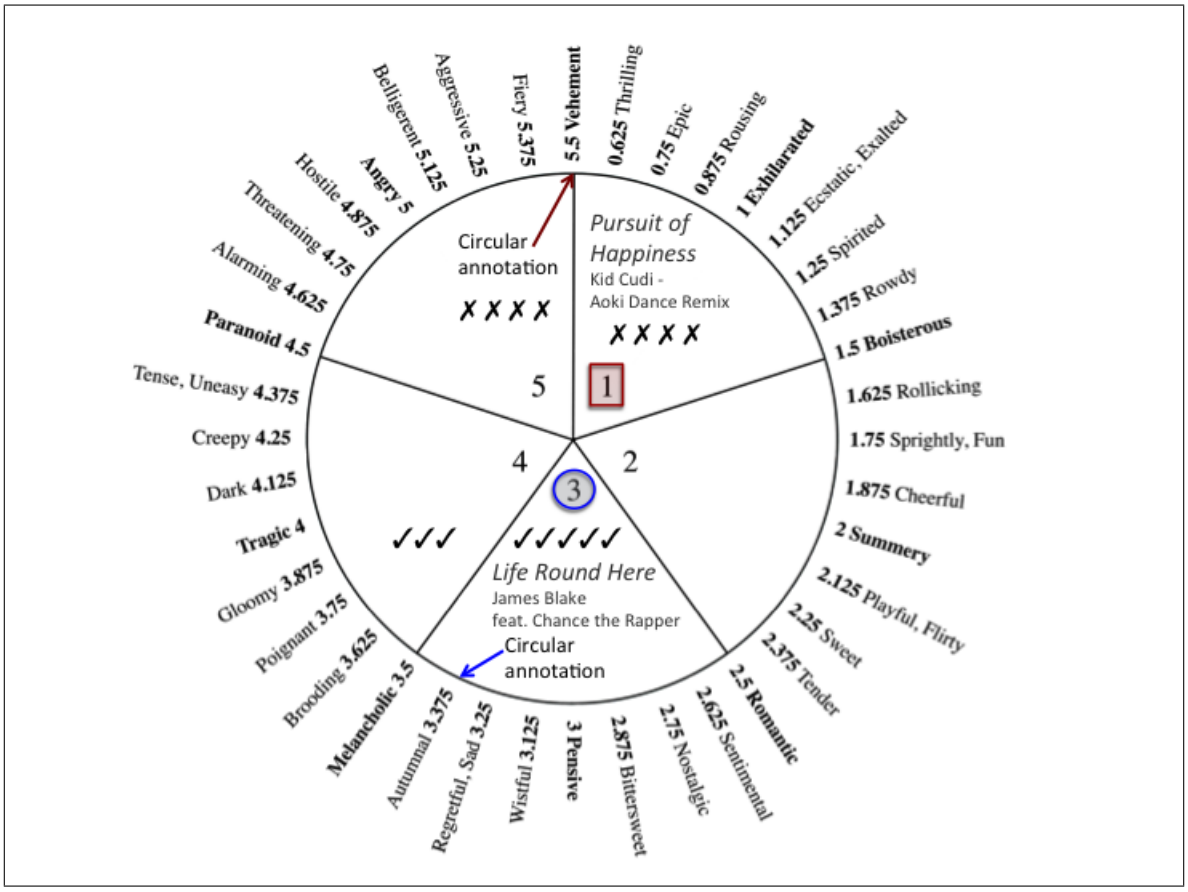

Figure 3.1: Wrapped circular mood model illustrating categorical and circular annotations of the case studies.

3.2.3 Dimensional Annotation

Finally, to allow comparison with a third type of model, the circular annotations were transformed to accommodate the creation of the commonly accepted two-dimensional valence/arousal model. To achieve this, the circular regression annotations were converted to cartesian coordinates by defining the arousal axis as the diameter going from the regression value 5.5:*Vehement* to value 3:*Pensive*, and the valence axis the one from 4.25:*Creepy*, to 1.75:*Sprightly/Fun*, both on ranges $[-1.25, 1.25]$. See Table 3.7

for examples of two-dimensional VA annotations, and Figure 3.2 for an illustration of the three annotation models.

Audio Clip	**GT**	**Reg**	**Valence**	**Arousal**
Don't Come Home A Drinkin - Loretta Lynn	2	2.125	0.875	-0.375
Bambino - Plastic Bertrand	1	0.75	0.25	1
Get Lucky - Daft Punk feat. Pharrell Williams	2	2	1	-0.25
Life Round Here (feat. Chance The Rapper) - James Blake	3	3.375	-0.375	-0.875
Motion Picture Soundtrack - Radiohead	4	4	-1	-0.25
Rite of Spring - Glorification of the Chosen One - Igor Stravinsky	5	4.75	-0.75	0.75
Pursuit of Happiness (Steve Aoki Dance Remix) - Kid Cudi	5	5.5	0	1.25

Table 3.7: Examples of Valence and Arousal annotations.

.

3.3 Feature Extractions

A total of 126 features were extracted using the Marsyas (Music Analysis, Retrieval, and Synthesis for Audio Signals) framework [70], spanning the typical types (intensity, timbre, register, rhythm and articulation) for mood and genre classification [42, 43, 47, 38, 69, 39]. Of these, 97 have been retained, including statistical moments (mean, standard deviation) of spectral centroid, flux and rolloff, zero-crossings, 13 coefficient MFCCs, chromas and tempo (BPM). Further pruning of the features proved to be over-fitting the system to the given data (increasing performance on given stratified subsets of the data, while decreasing it on others).

This set of audio features was selected because their performance on similar datasets and categorical models is known. Since this work's focus was to investigate the validity of a continuous circular emotional model as opposed to the commonly accepted categorical and two-dimensional models, no thorough investigations of the best features, their combination and the frameworks used to extract them were conducted.

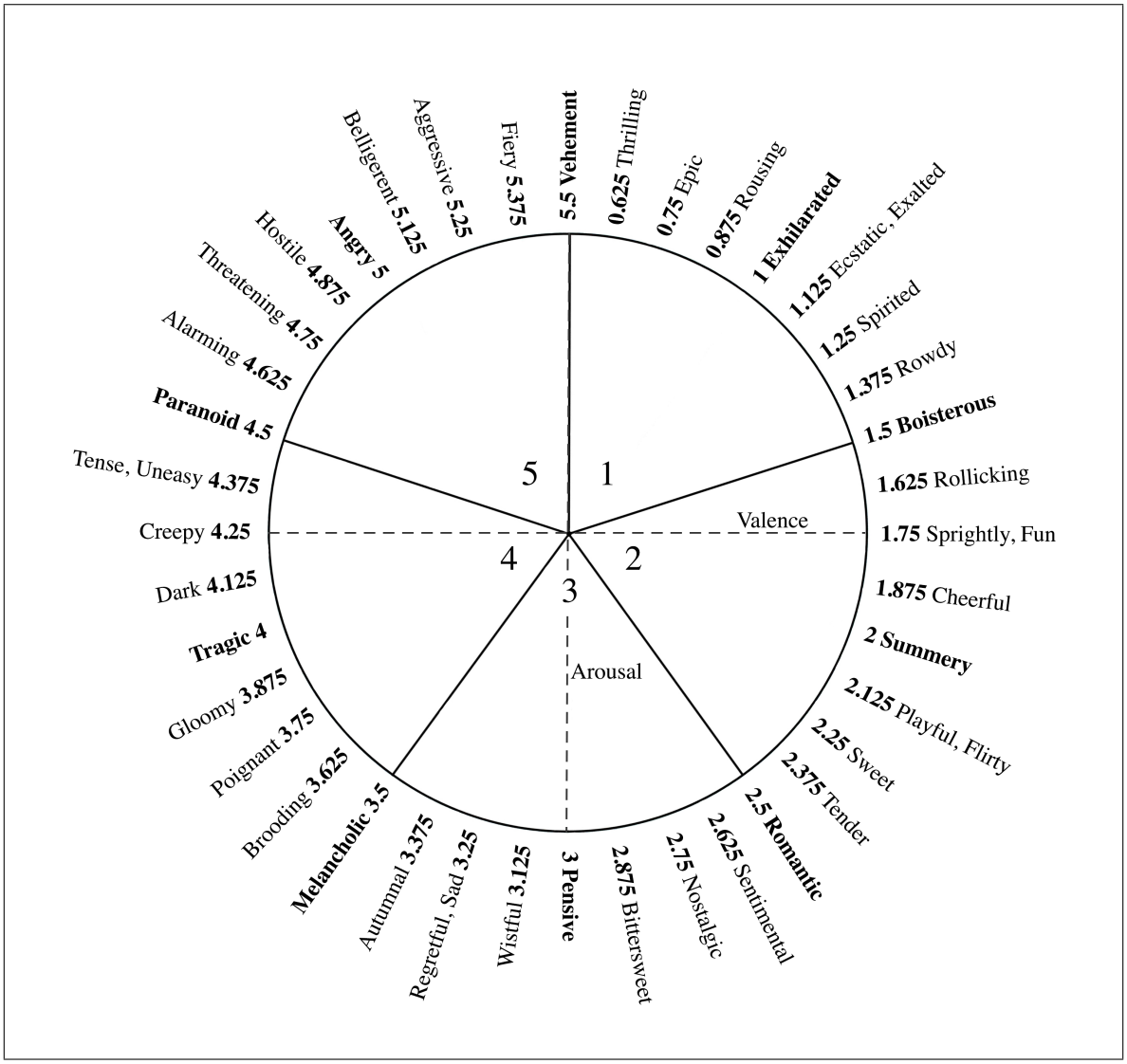

Figure 3.2: Wrapped circular mood model for annotations. The circular annotation model is shown around the circle, categorical clusters are represented by the pie chart, and the Valence and Arousal axes as dashed lines.

3.4 Summary

Building datasets to adequately test and train MER systems is another task involving numerous decisions. Which model should the annotation follow, how many categories or dimensions should be used, and which terminology applied? Once these questions are answered, the designers of MER systems have to decide how many songs to gather, if they should choose a specific genre or not, use whole songs or extracts, and which length and section of a song would give the best results. And it doesn't stop here, how many annotators is reasonable? Who should they be? And most importantly, how will the annotations be compiled considering the variation within? The literature presents many examples of choices made by MIR researchers, but no clear and consistent guidelines or methodology stands out.

Going through the process of annotating the dataset and observing first hand the different perceptions people have of the same song is highly instructive. The inability to establish a ground truth on several musical extracts, in both this work and the literature, systematically excludes important subsets of music. To work around this problem, the MIR community has been experimenting with dimensional models in the hopes to better account for the variability in the emotional perception of music. The most popular, the two-dimensional VA model has exhibited a moderate correlation with audio features on one of its dimensions. This leaves the question: could there be a better emotional model to fully encompass the MER problem? Considering that this variability seems to be confined to neighbouring emotions, and that all music emotions have at least two neighbours, could this model be circular?

Building one dataset with a set of three different annotations permits a full investigation on all three models. The architecture and implementation of the models used to conduct this investigation, as well as the introduction of an approximation to circular regression are the subjects of the next chapter.

Chapter 4

Building Models

This chapter details the implementation, training and testing methods utilized to research the three models chosen for this work: the categorical, circular and two-dimensional Valence-Arousal (VA) models. Additionally, the procedures followed to transform the circular and two-dimensional VA model as classifiers are given. The investigation of the categorical model was conducted on both the full dataset and the unambiguous dataset to create a baseline for comparison. The unambiguous dataset was used specifically to follow procedures reported in the literature on categorical models, providing another baseline comparison for this study. Only the full dataset was used with the circular and two-dimensional models, as they were designed to account for the annotators' disagreements.

4.1 Categorical Model

A number of Weka's [24] implementations of machine learning algorithms were tested, including Radial Basis Function Network (RBFNetwork), Random Forest, Naive Bayes and Simple KMeans. The best results were obtained for both datasets (*full* and *unambiguous*) by training a classifier using Support Vector Machines.

Support Vector Machines (SVMs) were invented by Vladimir Vapnik in 1979 [73]. Loosely speaking, SVMs are used to find the hyperplane that separates two categories of data points with the maximum margin. The hyperplane is found using the training data, and is then used to classify new data points. Because many hyperplanes may be valid separators, it comes down to finding the one that maintains the greatest distance from all the points. This maximization problem is a very large quadratic

programming (QP) optimization problem.

When using SVMs with multiple categories or classes, as in our audio mood classification problem, a number of binary classifiers are built. In *one-versus-all* architecture, a binary classifier is built for each category against all the others, while the *one-versus-one* trains a binary classifier for each coupling of the categories, none of these simplify the QP problem. In 1998, Platt [60] proposed the Sequential Minimal Optimization (SMO) algorithm, that breaks large QP problems into the smallest ones possible, then proceeds to solve them analytically. The Weka implementation of the SMO algorithm provided the best results for the audio music mood problem at hand, and was therefore chosen for the categorical model. The multiclass training was made building the logistic models using pairwise coupling [26], where the probabilistic output obtained building the multiple one-versus-one models are combined to a set of posterior probabilities.

4.2 Polygonal Circular Regression Models

One common application of linear regression is to provide a way to fit a predictive model based on a set of dependant and explanatory variables. Let Y be a linear dependant variable, X be a linear variable, x an explanatory variable, and the general linear regression model:

$$E(Y|X = x) = a + bx \tag{4.1}$$

where $E(Y|X = x)$ should read as the mean value of Y given $X = x$. These models are fitted utilizing linear algebra and are written in vector form as:

$$\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{\xi} \tag{4.2}$$

where ξ is the vector of errors terms. Loosely speaking, linear regression attempts to find the equation for the line best describing all of the given observed variables, that is, the regression line minimizing the distance (or error ξ) between all observations and itself. In the case of observations distributed in a circular manner (hours around a clock for example), the distance is at its smallest if the regression line is any diameter (e.g. the straight line going through 9:00 and 3:00). Unfortunately, it implies that points situated above or below the mid-point of the diameter (in our example points near 12:00 and 6:00), will both be mapped to the same point on the regression line.

To verify the circularity of the music emotion recognition problem, I propose a

circular regression approximation which I term *polygonal circular regression*, fitting several linear models and combining them in ways that account for the circular distortion. Three such models performing linear regression on different sets of arcs of the wrapped circular mood annotations (Figure 3.2) were constructed utilizing the regression annotation.

4.2.1 Full Pentagon Model

The first model, referred to as the *full pentagon* ($F - poly$) model consists of a combination of five linear regression submodels such that, for each submodel m_i, the circle is cut at point i and lain on a flat line, for a full rotation around the 5 classes, following:

$$\mathbf{Y_i} = \boldsymbol{X\beta} + \boldsymbol{\xi}; \quad y_{i\_j} = \begin{cases} y_j, & \text{if } y_j \geq i \\ y_j + 5, & \text{otherwise} \end{cases}; \quad \forall i \in \{1, 2, 3, 4, 5\} \quad (4.3)$$

where $y_j \in \mathbf{Y}$ is the regression annotation of the song j, and $y_{i\_j} \in \mathbf{Y_i}$ is the regression value used by model m_i for the song j. Models are fitted and tested on 10-fold cross-validation with consistent folds across all five models. Circularity was confirmed by the poor results obtained for the classes at each extremity of the line (for example $C1$ and $C5$ for model m_1) on any rotation of the model with a mean error of 1.529, while the classes in the middle (for example $C2$, $C3$ and $C4$ for model m_1) had a mean error of 0.56. To account for this circular distortion, the five models are combined for a final prediction by taking the average of the predictions on the three central classes of each model.

4.2.2 Reduced Pentagon Model

A *reduced pentagon* (RP-poly) model was next built. Again a total of five models were constructed, but this time the circular distortion was minimized by using subsets of the dataset for fitting models over smaller arcs of the circle. A song j contributes to a model m_i iff $i \leq y_{i\_j} < i + 3$ following:

$$\mathbf{Y_i} = \boldsymbol{X\beta} + \boldsymbol{\xi}; \quad y_{i\_j} = \begin{cases} y_j, & \text{if } y_j \geq i \\ y_j + 5, & \text{otherwise} \end{cases}; \quad \forall i \in \{1, 2, 3, 4, 5\} \quad (4.4)$$

Dashed lines are used to illustrate the partition of the values for the regression submodels in Figure 4.1. The reduced pentagonal model implies that each clip contributes to three submodels m, but the submodels are independent. Again, the results of the prediction of each model using 10-fold cross-validation is combined by taking the mean of the predictions, but this time the entire set of predictions of each model is used.

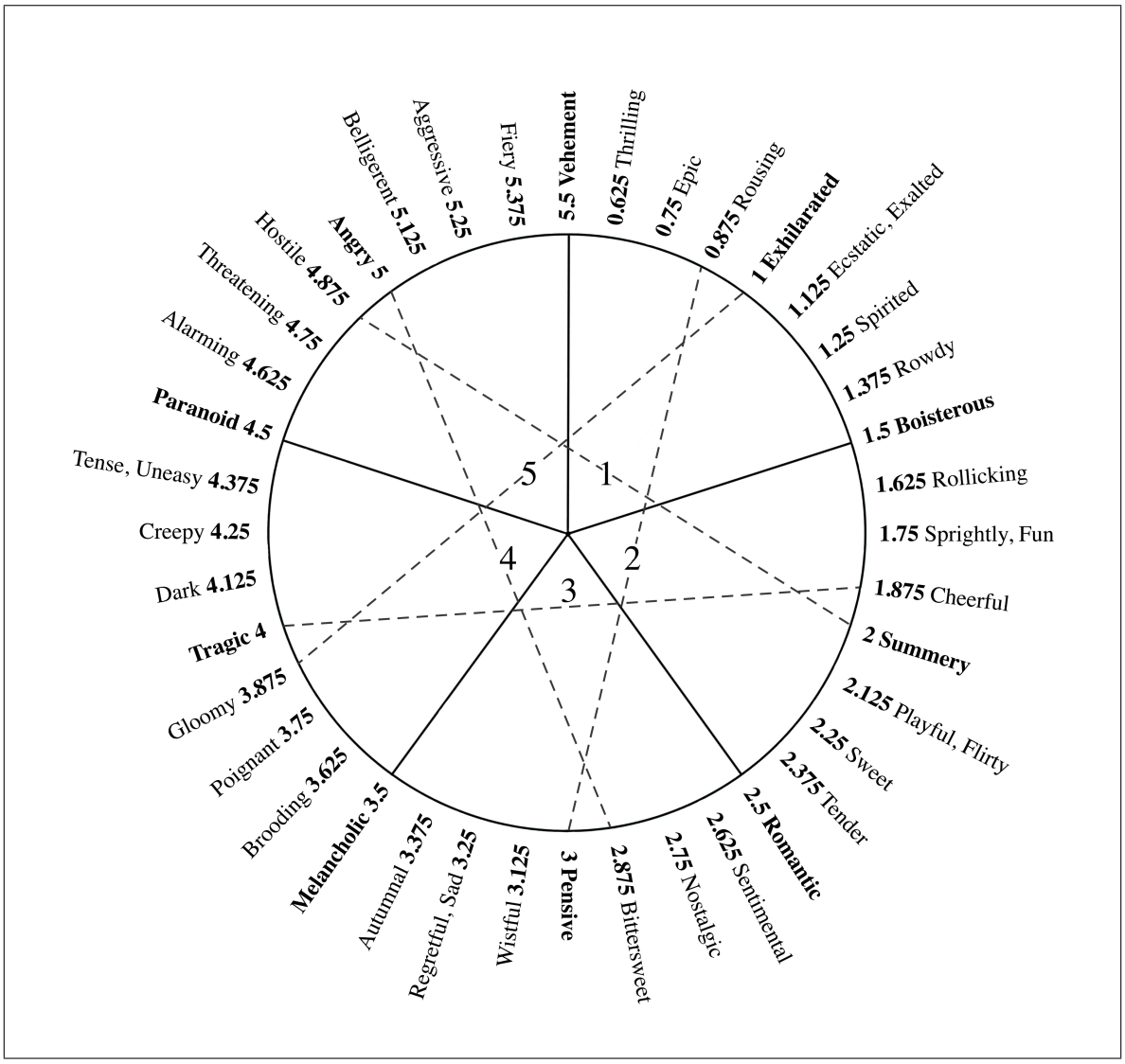

Figure 4.1: The five partitions of the submodels for the reduced pentagon model, indicated by dashed lines.

4.2.3 Decagon Model

Finally, a *decagon* (D-poly) model was built, using smaller arcs. This time, ten submodels were built as followed: a song j contributes to a model m_i iff $i \leq y_{i\_j} < i+2$ following:

$$\mathbf{Y_i} = \boldsymbol{X\beta} + \boldsymbol{\xi}; \quad y_{i\_j} = \begin{cases} y_j, & \text{if } y_j \geq i \\ y_j + 5, & \text{otherwise} \end{cases};$$
$$\forall i \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5\} \quad (4.5)$$

Each clip contributes to four submodels, and in this case also, the submodels are independent. The predictions of the models on 10-fold cross-validation are combined in the same manner as for the *reduced pentagon* model.

4.3 Dimensional Models

To add depth to the analysis, three siblings of the polygonal circular models (full, reduced pentagon, and decagon) were constructed utilizing the two-dimensional (V,A) coordinates as given in Section 3.2.3. To be able to compare the results of the polygonal circular models, the predictions of the two-dimensional models are projected on the unit circle, measuring the angular response θ following:

$$\theta = \arctan\left(\frac{Valence_{predition}}{Arousal_{prediction}}\right) \quad (4.6)$$

For the *full dimensional* (F-dim) model, one linear regression model is fitted on each of the two dimensions (valence and arousal) utilizing the entire dataset, and the angular response is obtained following equation 5.1. The other two-dimensional sibling models, *dimensional reduced pentagon* (RP-dim) and *dimensional decagon* (D-dim), are fitted using the same subsets of the dataset, and evaluated on the same folds as the 10-fold cross-validation used for their polygonal circular siblings.

Polygonal circular and Dimensional Models as Classifiers

A mentioned in Chapter 2, dimensional models are often utilized for classification. Therefore, to compare the results of the polygonal circular and two-dimensional models with the results of the categorical model, the output of each model has been used

to classify the test data into the original clusters of the categorical model. To achieve this, a simple mapping from the angular response to classes is performed, where each class represents a specific arc of the circle.

4.4 Summary

This chapter explains the choices made in building the models used in this thesis. One categorical model built on SVMs was constructed to be trained and tested with both datasets (full and unambiguous). This model was used to evaluate the dataset in comparison to the MIREX dataset, and establish a baseline comparison for the other two models. The polygonal circular regression model was built with three variations: the full pentagon, reduced pentagon and decagon variants. These variants were built to be compared to each other, as well as with their three two-dimensional siblings. The valence/arousal models are trained and tested with the same subsets of the dataset used for their respective polygonal siblings and are built to provide a comparison for the pentagonal circular models. Finally, both polygonal and two-dimensional models are transformed into classifiers, to measure their ability to perform the classification. The results obtained by each model are presented in the next chapter, with their analysis, evaluation and comparison in Chapter 6.

Chapter 5

Experimental Results

In this chapter, the results of the three models are presented, however their analysis will be detailed in the next chapter. The results are organized by model, with the results obtained on both datasets from the categorical model in Section 5.1, followed by the results of the polygonal circular models in Section 5.2, and the results of the two-dimensional models in Section 5.3.

5.1 Categorical Results

The categorical model was evaluated with both the full and *unambiguous* datasets.

Results on the Full Dataset

The full dataset (564 music extracts) yielded an accuracy of 59% on 10-fold cross-validation. The confusion matrix is shown in Table 5.1.

Truth/Predicted	C1	C2	C3	C4	C5
C1	**56.0**	19.0	0.0	7.0	18.0
C2	11.3	**58.9**	13.7	4.0	12.1
C3	0 .0	13.2	**68.6**	18.2	0.0
C4	4.9	11.7	22.5	**54.2**	6.7
C5	18.5	12.2	1.0	12.2	**56.1**

Table 5.1: Confusion Matrix of the full dataset

Examining the confusion matrix reveals that 74.01% of the classification errors happened within neighbouring classes. For example, 84.09% of the misclassifications of the music extracts labelled $C1$ were erroneously classified as either $C2$ or $C5$ by

the SVM. The proportion of the misclassification errors occurring within neighbouring classes are detailed in Table 5.2.

% per classes	C1	C2	C3	C4	C5	Total
Neighbouring classes error	84.09	60.78	100.00	63.64	69.77	74.01

Table 5.2: Percentage of misclassifications by the SMO algorithm observed within the neighbouring classes on the full dataset

.

Results on the Unambiguous Dataset

The accuracy obtained on the unambiguous dataset (417 music extracts) was 61.7% on 10-fold cross-validation. The unambiguous dataset was used to create a baseline comparison for the polygonal circular and two-dimensional models, but also to evaluate the categorical model against the works reported in the literature review. The complete confusion matrix for the unambiguous dataset can be seen in Table 5.3.

Truth/Predicted	C1	C2	C3	C4	C5
C1	**52.6**	17.1	0.0	9.2	21.1
C2	12.0	**65.2**	7.6	4.3	10.9
C3	1.1	9.9	**73.6**	14.3	1.1
C4	3.6	10.8	16.9	**61.5**	7.2
C5	21.9	19.2	0.0	6.8	**52.1**

Table 5.3: Confusion Matrix of the unambiguous dataset

Eliminating some of the ambiguous extracts decreased the proportion of the misclassification errors occurring within neighbouring classes to 68.55%, a decrease of 7.38%. The proportion of neighbouring classes errors are reported in Table 5.4.

% per classes	C1	C2	C3	C4	C5	Total
Neighbouring classes error	77.78	56.25	91.67	62.50	60.00	68.55

Table 5.4: Percentage of errors observed within the neighbouring classes on the unambiguous dataset

.

5.2 Polygonal Circular Regression Results

In order to have a system capable of automatically tagging music extracts with actual moods as opposed to numbers, the system is designed to annotate a music extract with the set T of tags t, made from the n closest tags to the output of the circular prediction. Following the categorical approach where a music extract gets a cluster of eight mood tags, the decision to retrieve seven tags ($n = 7$) seemed well-founded. More precisely, the tag corresponding to the numerical output, referred as the *target tag*, is the first term in the set, and the three tags on both sides of it on the circle are then added.

Using our case studies as examples, the song *Life Round Here* from James Blake (feat. Chance The Rapper), with a value of 3.375 would have *Autumnal* as its target tag, and would be annotated with the set of moods:

- 3.375 = [Autumnal, Melancholic, Sad/Regretful, Brooding, Wistful, Pensive, Poignant]

while the song *Pursuit of Happiness* from Kid Cudi (Steve Aoki Dance Remix) with a value of 5.5 would have *Vehement* as its target tag, and be annotated with the set of moods:

- 5.5 = [Vehement, Fiery, Thrilling, Epic, Aggressive, Rousing, Belligerent]

Because this approach is unique, it is difficult to compare the results with existing models. In order to evaluate it, statistics were drawn between the intended output and actual prediction by calculating the accuracy in terms of tag distance. For simplicity, the intended output corresponding to the regression annotation of the ground truth is the *target tag*, and the tag distance is formalized as:

$$TagDistance = \frac{|Prediction - TargetTag_{value}|}{0.125} \tag{5.1}$$

such that a prediction at a tag distance of zero matches the regression annotation, and a tag distance of one would substitute one tag in the set for another, as illustrated in Figure 5.1.

The accuracy represents the percentage of predictions falling within a given tag distance. Measures of up to a distance of four tags are calculated, as it seemed a reasonable threshold distance considering half of the intended tags would still be

retrieved. Moreover, a misclassification in a neighbouring class from the categorical model would have a general tag distance of up to eight.

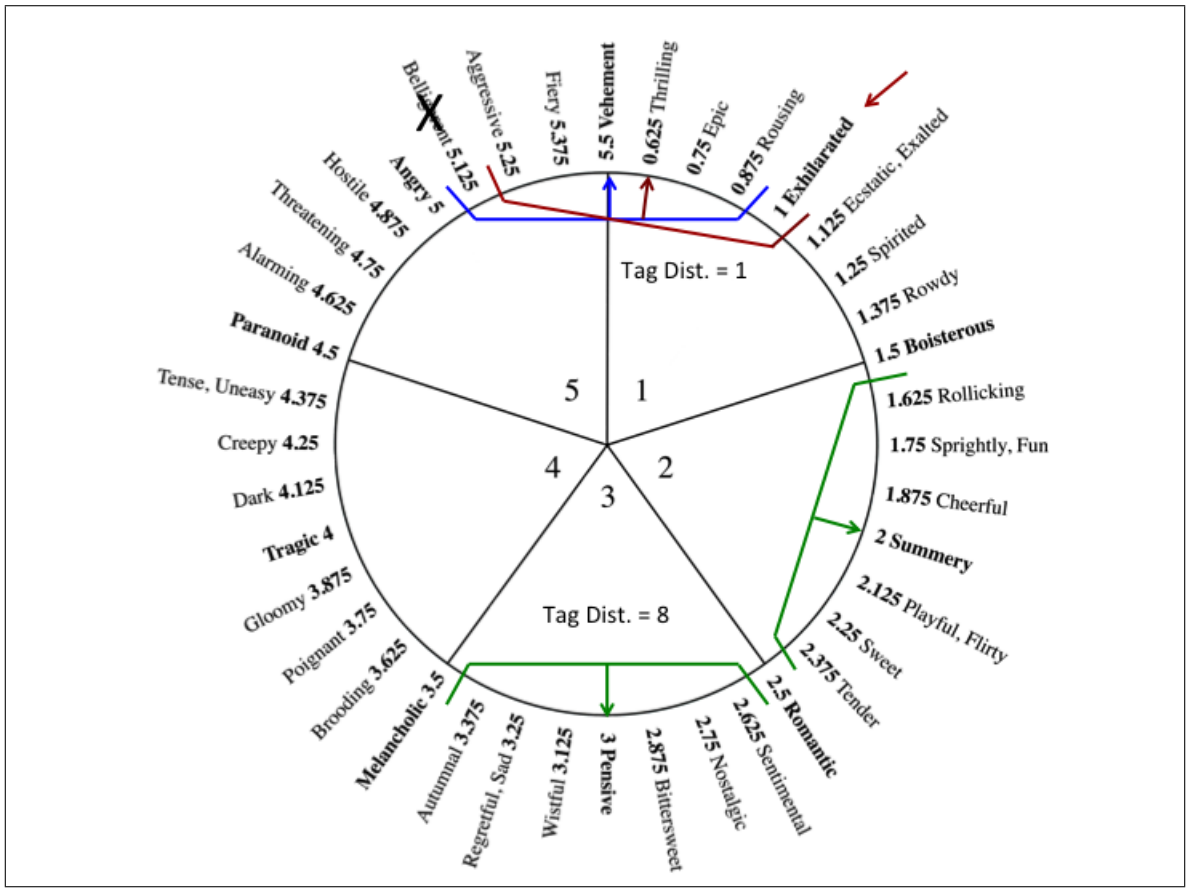

Figure 5.1: Examples of tag distance. The top example shows a tag distance of 1, and the bottom illustrates a misclassification in a neighbouring class, a tag distance of 8.

The accuracies obtained on 10-fold cross-validation on the full dataset are shown in Table 6.5 for the three models utilizing the circular annotations.

Tag Distance	Full Pentagon	Reduced Pentagon	Decagon
0	10.66%	12.97%	17.94%
1	32.14%	36.23%	48.67%
2	51.87%	57.19%	71.76%
3	68.56%	74.60%	84.01%
4	79.75%	86.15%	91.83%

Table 5.5: Accuracy in terms of distance to target tag for the three polygonal models.

Polygonal Circular Classification Results

The three polygonal circular models were used as classifiers to compare their results to the ones obtained by the categorical models. The confusion matrices of the three models on the full dataset are shown in Table 5.6.

Full Pentagon Model	Total accuracy : **60.57%**				
Truth/Predicted	C1	C2	C3	C4	C5
C1	**65.0**	11.0	12.0	0.0	12.0
C2	14.5	**62.1**	22.6	0.8	0.0
C3	0 .0	13.2	**68.6**	18.2	0.0
C4	0.8	0.0	20.8	**54.2**	24.2
C5	38.8	1.0	0.0	8.2	**52.0**

Reduced Pentagon Model	Total accuracy : **63.41%**				
Truth/Predicted	C1	C2	C3	C4	C5
C1	**67.0**	16.0	4.0	0.0	13.0
C2	21.8	**55.6**	22.6	0.0	0.0
C3	0 .0	9.1	**77.7**	13.2	0.0
C4	0.0	0.0	23.3	**55.0**	21.6
C5	27.5	0.0	0.0	10.2	**62.3**

Decagon Model	Total accuracy : **77.98%**				
Truth/Predicted	C1	C2	C3	C4	C5
C1	**69.0**	19.0	4.0	0.0	8.0
C2	10.5	**80.6**	8.9	0.0	0.0
C3	0 .0	5.0	**78.5**	16.5	0.0
C4	0.0	0.0	8.3	**75.9**	15.8
C5	8.2	2.0	0.0	4.1	**85.7**

Table 5.6: Confusion matrices of the full dataset for the polygonal circular models.

A total accuracy of 63.41% was achieved for the full dataset with the reduced pentagon and 77.98% with the decagon model, compared to 58.9% in the strict categorical model, an increase of 4.51% and 19.01% respectively. The percentage of misclassification errors occurring within neighbouring classes have significantly increased in all three models, as shown in Table 5.7.

% of neigh. classes error/Class	C1	C2	C3	C4	C5	Total
Full Pentagon	65.71	97.87	100	98.18	97.87	93.24
Reduced Pentagon	87.88	100	100	100	100	98.06
Decagon	87.10	100	100	100	85.71	91.47

Table 5.7: Percentage of errors observed within the neighbouring classes on the full dataset.

5.3 Two-Dimensional Models

The performance of the two-dimensional models was first measured by tag distance to allow for comparison with the performance of the polygonal circular models. As was the case with their polygonal siblings, all two-dimensional models were evaluated on 10-fold cross-validation. The results are presented in Table 5.8

Tag Distance	Full Dimensional	RP - Dimensional	D - Dimensional
0	7.28%	12.79%	17.23%
1	22.74%	35.17%	47.96%
2	39.08%	54.35%	72.82%
3	51.33%	66.61%	87.39%
4	59.86%	77.44%	94.49%

Table 5.8: Accuracy in terms of distance to target tag for the three two-dimensional models (RP: Reduced Pentagon, D: Decagon).

Two-Dimensional Classification Results

As with the polygonal circular siblings, the performance of the two-dimensional models as classifiers was measured. The confusion matrices of the three variations are given in Table 5.9.

Finally, similar statistics to the ones given regarding the misclassifications errors within neighbouring classes for the polygonal circular models are reported in 5.10.

Full Pentagon Model	Total accuracy : **51.50%**				
Truth/Predicted	C1	C2	C3	C4	C5
C1	**61.0**	9.0	3.0	7.0	20.0
C2	23.4	**46.0**	18.5	4.0	8.1
C3	2.5	14.0	**74.4**	9.1	0.0
C4	7.5	7.5	35.8	**39.2**	10.0
C5	34.7	16.3	1.0	12.2	**35.7**

Reduced Pentagon Model	Total accuracy : **60.92%**				
Truth/Predicted	C1	C2	C3	C4	C5
C1	**67.0**	10.0	3.0	0.0	20.0
C2	25.8	**62.1**	12.1	0.0	0.0
C3	0 .0	11.6	**81.0**	7.4	0.0
C4	3.3	2.5	29.2	**51.7**	13.3
C5	40.8	3.1	2.0	14.3	**39.8**

Decagon Model	Total accuracy : **78.50%**				
Truth/Predicted	C1	C2	C3	C4	C5
C1	**77.0**	13.0	4.0	0.0	10.0
C2	10.5	**81.5**	8.1	0.0	0.0
C3	0.0	7.5	**85.1**	7.4	0.0
C4	0.0	0.0	14.2	**71.6**	14.2
C5	14.3	0.0	0.0	9.2	**76.5**

Table 5.9: Confusion matrices of the full dataset for the dimensional models.

% of neigh. classes error/Class	C1	C2	C3	C4	C5	Total
Full Dimensional	74.36	77.61	90.32	75.34	73.02	76.92
RP - Dimensional	90.90	100	100	87.93	91.53	93.18
D - Dimensional	100	100	100	100	100	100

Table 5.10: Percentage of errors observed within the neighbouring classes on the full dataset. Reduced Pentagon (RP), Decagon (D).

Chapter 6

Evaluation, Analysis and Comparisons

The initial goal of this investigation was to explore the music emotion recognition problem, with hopes to modestly contribute to the existing models. In building an original dataset and going through the annotation process, I observed first hand that the categorical model had the limitation of not well representing the varied emotional response to music of the different annotators. This proved fundamental to the direction this research took. In this Chapter, I will examine the results presented in Chapter 5 for each model, but the discussion will start by reviewing the observations made while conducting the human annotation process, and how it led to the circular hypothesis.

6.1 Ground Truth Discussion

Let's begin by recalling some of the statistics presented in Chapter 3. A total of twelve human annotators were asked to classify either different subsets or the entire original dataset of 591 thirty second extracts into five clusters (see Table 6.1) of moods inspired by the MIREX AMC task. The annotation process was conducted such that each audio clip was evaluated by eight of the twelve annotators, and their annotations were combined by means of majority (examples of which are shown in Table 6.2). Of the original 591, twenty-seven were classified with such disagreement that it was impossible to reconcile the votes to one class, and they were thus removed, leaving a *full dataset* of 564 thirty second extracts from diverse music genre.

Happy			Sad	Mad
C1	C2	C3	C4	C5
Rousing Rowdy Boisterous Thrilling Epic Exhilarated Exalted Ecstatic Spirited	Rollicking Fun Cheerful Sweet Sprightly Summery Playful Flirty	Autumnal Bittersweet Wistful Nostalgic Sentimental Tender Pensive Regretful	Poignant Brooding Melancholic Mournful Tragic Gloomy Dark Creepy Paranoid	Agressive Volatile Fiery Threatening Hostile Belligerent Arrogant Angry

Table 6.1: Mood Classes/Clusters used for the annotation of the ground truth for the categorical model

Unanimity among the annotations was reached for only 127 clips (22.5%), and an agreement among six or more annotators was reached for 417 clips (73.9%). These observations are important, as similar disparities among the emotional perception of the annotators have been observed in every related work, regardless of the model chosen, and this in both psychological and MER studies. Unanimity is somewhat rare (22.5% in this case), however strong agreement is more commonly observed (six out of eight annotators reached agreement in 73.9% of the cases) supporting the theory that human emotion can be efficiently modelled.

A closer look at the eliminated 147 clips (26.1% of the full dataset) removed from the full dataset to build the unambiguous one, revealed that more than half of the annotators disagreed for 94 clips (16.7% of the dataset, 63.9% of the 147 disagreeing cases); in 58 of those cases (10.3% of the dataset, 39.5% of the disagreeing cases), the annotations were equally split between two neighbouring classes. Other notable observations came from the tendency for native english speakers to perceive music of different languages as somewhat *aggressive* or *threatening*. Of course this is a generalization, but the hypothesis that given the predominance of english lyrics in the music industry, non-english speakers were more likely to have listened and enjoyed music at an earlier stage in their life without understanding the lyrics.

Examining the conflicting cases, it was noticed that not only were the disagreements mostly observed between neighbouring classes but that a similar split was occurring between class one ($C1$) and class five ($C5$). It became apparent that the acoustic similarity reported by Laurier et al. [42], paired with the culture, background and lifestyle of the annotators was playing a major role in explaining these

Audio Clip	**C1**	**C2**	**C3**	**C4**	**C5**	**GT**
Don't Come Home A Drinkin - Loretta Lynn	0	7	1	0	0	2
Bambino - Plastic Bertrand	6	2	0	0	0	1
Get Lucky - Daft Punk feat. Pharrell Williams	0	8	0	0	0	2
Life Round Here (feat. Chance The Rapper) - James Blake	0	0	5	3	0	3
Motion Picture Soundtrack - Radiohead	0	0	0	8	0	4
Rite of Spring - Glorification of the Chosen One - Stravinsky	0	0	0	2	6	5
Beat It - Michael Jackson	3	2	0	1	2	X

Table 6.2: Example annotations and resulting ground truth classes (GT) based on eight annotators

Annotators' Agreement	***#clips*/564**	**% of clips**
Unanimity 8/8	127	22.5%
Strong: $\geqslant 6/8$	417	73.9%
Weak: $\leqslant 4/8$	94	16.7%
Equally split between two neighbouring classes	58	10.3%

Table 6.3: Agreements statistics of eight annotators on the full dataset.

disagreements. As an example, electronic dance music was often perceived as *thrilling* and *rousing* by ravers, while perceived as *aggressive* and *fiery* by opera and musical enthusiasts. The observation that they were both legitimate, and non-exclusive gave rise to the idea that the emotional response to music should be approached using a continuous model of emotion.

Kate Hevner had made similar observations in 1936 [29] while conducting her studies on human's perception of the moods conveyed by music. As mentioned in Chapter 2, Hevner's model was categorical, but the organization of the categories shows her awareness of the dimensionality and continuous nature of the problem. One of the advantages of using this model according to Hevner herself, is that the more or less continuous scale accounted for small disagreements amongst annotators, as well as the effect of pre-existing moods or physiological conditions that could have influenced the annotators' perceptions. Although Hevner's clusters are highly regarded, they have not been used in their original form by the MIR community. One of the reasons for this might come from the dated choice of adjectives. Important moods, mostly on the aggressive and dark side of the human emotional spectrum

are missing. Hevner conducted her study exclusively on Western classical music at a time where recordings were rare and of poor audio quality, and in a society generally disapproving of the expression of less than noble emotions.

One solution to better represent the diverse emotional response to music could come in allowing multiple categories, but the evidence points towards having only a subset of the tags for each of the neighbouring categories being compatible without starting to lose their consistency. For example, allowing a song to get the moods of both $C1$ and $C5$ would annotate it with sixteen mood terms, a third of all possible terms (following the categories put forward in this work). The idea of increasing the number of classes in order to define them better was explored, but the exercise became recursive, and was subsequently abandoned.

6.2 Categorical Results Analysis

The results obtained with the categorical model were satisfying with a 58.9% accuracy on the full dataset, and 61.7% for the unambiguous one, using 10-fold cross-validation. These results demonstrate the validity of the categorical annotation ground truth and datasets, as they compare to the average accuracies obtained at the AMC MIREX over the last eight years, on a comparable dataset (thirty second extracts of varied music genres) and five classes (or clusters). More precisely, the 2014 AMC benchmark competition[1] saw accuracy results ranging from 46.3% to 66.3%. Moreover, the full dataset still includes extracts for which the annotators were evenly split, making the dataset noisier than the one used for the MIREX AMC.

The most interesting aspect of the results came from the analysis of the classification errors. The confusion matrix of the full dataset showed that 74% of the misclassification errors were made within neighbouring classes when extending the neighbouring concept to $C1$ and $C5$. This reflected the observations made when compiling the human annotations on how personal music preference, lifestyle, cultural background etc. influenced the annotators' perception (Section 6.1). This also reflected the recurring misclassification errors in the AMC MIREX competition for $C1$ and $C5$, first reported by Laurier et al.[42]. In our case 84.09% of the $C1$ extracts that were misclassified were put in either $C2$ or $C5$.

This comparison between the divided opinions of the annotators constructing the

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

ground truth and the errors of the systems showed some consistency within the inconsistencies; the system did indeed seem to represent some degree of truth, just not necessarily the average over eight annotators' truths.

Numerous examples were found going back to the original annotations. Lets examine our case study *Life Round here* from James Blake (feat. Chance The Rapper) shown in Table 6.4. Five annotators considered the extract to fall within $C3$ (Autumnal, Bittersweet, Wistful, Nostalgic, Sentimental, Tender, Pensive and Regretful) while the remaining three put it in the $C4$ (Poignant, Brooding, Melancholic, Mournful, Tragic, Gloomy, Dark, Creepy, Paranoid); the final class for the ground truth annotation was therefore $C3$. The trained algorithm misclassified the extract by putting it in $C4$, agreeing with 37.5% of our annotators.

Lets now give a closer look at our other case study, *Pursuit of Happiness* from Kid Cudi (Steve Aoki Dance Remix). This time, the annotators were evenly divided between $C1$ (Rousing, Rowdy, Boisterous, Thrilling, Epic, Exhilarated, Exalted, Ecstatic and Spirited) and $C5$ (Agressive, Volatile, Fiery, Threatening, Hostile, Belligerent, Arrogant and Angry). The human compiler made the executive decision to classify this extract as $C1$. The system on the other hand agreed with the other four annotators, and classified it in $C5$.

Audio Clip	C1	C2	C3	C4	C5	GT	SMO
Life Round Here (feat. Chance The Rapper) - James Blake	0	0	5	3	0	3	4
Pursuit of Happiness (Steve Aoki Dance Remix) - Kid Cudi	4	0	0	0	4	1	5

Table 6.4: Example of annotations, resulting class (GT), and final classification by the SMO

These are two examples among many, and are presented here to illustrate how the disagreements among the annotators between neighbouring clusters were later observed as misclassifications by the system. In fact, the confusion matrices of the algorithms submitted at the 2014 AMC MIREX competition[2] all show similar misclassification patterns: most of the errors are occurring in semantically neighbouring clusters (MIREX cluster 2: Rollicking, Amiable/Good-natured, Fun, Cheerful and Sweet, and cluster 4: Witty, Humorous, Whimsical, Wry, Campy, Quirky and Silly are semantic neighbours, but not numerical neighbours). Observing the MIREX mood

[2] `http://www.music-ir.org/nema_out/mirex2014/results/act/mood_report/overallconfusion.html`

clusters, one can find that the semantic boundaries between clusters are questionable. For example, *Rollicking* ($C2$) is a synonym to both *Playful* ($C2$) and *Boisterous* ($C1$), in turn respectively synonyms of *Whimsical* ($C4$) and *Intense* ($C5$).

One can easily argue that having more annotators would have made the ground truth converge differently, but the argument that it might create more cases where opinions are spread across the five classes is just as valid.

The question of the validity of the categorical model as one that fully encompasses the MER problem has to be asked. It captures some of its dimensions, but seems to lack finesse by attempting to put human emotions and their artistic creations into boxes. Nonetheless, it provides a vocabulary users understand and can easily relate to. The dimensional model has the advantage of better representing both the continuity and circularity of the problem, but predictions are in most cases reduced to a single label, quadrant, or values on an axis that are not intuitive from a listeners' point of view.

6.3 Polygonal Circular and Two-Dimensional Results Analysis

Overall, the polygonal approach had better accuracy than its dimensional siblings, except for the dimensional decagon model (see table 6.5). These higher results are by no means unexpected and can be explained by the fact that the size of the arcs used to fit the models considerably reduces one of the two dimensions; for instance, only a third of the valence axis is considered for the model m_1.

Dist.	F-dim	FP-poly	RP-dim	RP-poly	D-dim	D-poly
0	7.28%	10.66%	12.79%	12.97%	17.23%	17.94%
1	22.74%	32.14%	35.17%	36.23%	47.96%	48.67%
2	38.08%	51.87%	54.35%	57.19%	72.82%	71.76%
3	51.33%	68.56%	66.61%	74.60%	87.39%	84.01%
4	59.86%	79.75%	74.44%	86.15%	94.49%	91.83%

Table 6.5: Accuracy in terms of distance to target tag for the dimensional (-dim) and polygonal (-poly) versions of the models: F: Full , RP: Reduced Pentagon and D: Decagon

Correlation coefficient values vary from $[-1, 1]$, where 1 denotes a perfect positive correlation, 0 no correlation, and -1 a perfect negative correlation. Typically, if the

absolute value of a correlation coefficient is greater or equal to 0.7, there is a strong linear relationship. If the absolute value is smaller then 0.3, the linear relationship is said to be weak. Similar to the results reported in the literature (Chapter 2), the arousal regressor of the full dimensional model has a stronger linear correlation with a correlation coefficient value at 0.74, compared to the valence regressor with a correlation coefficient value at 0.48.

The annotations were not collected for the two-dimensional model, therefore it would be presumptuous to use the observed results on those models to draw strong conclusions. However, its consistency with the results found in the literature hints that the calculated annotations have some validity. As an example, let's examine the correlation coefficient of a smaller model, for the valence regressor m_1 of the reduced pentagon model, we can see that the linear correlation is weak at 0.36, and has a relative absolute error of 103.98%. The polygonal sibling on the other hand has a correlation coefficient of 0.72 and a relative absolute error of 62.13%, showing a better fit to the problem on the same data.

This reinforces the hypothesis that the music emotion problem is dimensional, and the choice of valence as one of the measured dimensions does not seem to be a strongly correlated descriptor, neither in these results, nor the ones reported in the literature. The strong results from the dimensional decagon model might just come from overfitting the submodels and the distortion created by projecting onto the unit circle when calculating the angular response.

As for the polygonal models, the same overfitting concern can legitimately be raised, but in this case, even the full pentagon model using the entire *full dataset* in each of the five submodels, offers superior accuracy in its annotation (68.56%), and this with a tag distance as little as 3. Most importantly, the accuracy of five models out of six, at a tag distance of only 3 outperformed the classification model by at least 10%, and up to 20%. Using our James Blake example to illustrate how a tag distance of 3 would affect the automatic annotation, such an error could replace the desired set of tags:

- from [Autumnal, Melancholic, Sad/Regretful, Brooding, Wistful, Pensive, Poignant]
- with [Autumnal, Melancholic, Brooding, Poignant, Gloomy, Tragic, Dark]

preserving four of the original tags of the set, a much less cost-full error than say, a classification error.

In taking a final look at our two case studies misclassified by the SVM system, we can observe that the reduced pentagon (RP) regression model perfectly represented *Life Round Here*, with a prediction error of 0.003 on the mean prediction as reported in Table 6.6.

Audio Clip	Anno.	RPr	TPr	ePr	GT	RC
Life Round Here (feat. Chance The Rapper) - James Blake	3.375	3.375	3.372	0.003	3	3
Pursuit of Happiness (Steve Aoki Dance Remix) - Kid Cudi	5.5	5.375	5.407	0.093	1	5

Table 6.6: Summary of the reduced pentagon regression predictions for two clips showing the annotation (Anno), rounded prediction (RPr), true prediction (TPr), prediction error (ePr), original classification ground truth (GT) and classification by regression (RC).

Dimensional models have the advantage of better encompassing the variability within the human emotional response to music. However the most commonly used dimensional model to this day, the valence/arousal two-dimensional model, has shown some weakness in at least one of its dimensions (valence), and this in both the experiments reported in this work, and throughout the literature on the subject. The polygonal approximations to circular regression on the other hand, provided good results, significantly outperforming the categorical model in all of its polygonal variants. Moreover, this circular approach provides a flexible method to automatically annotate music without having to sacrifice a diverse lexicon users can better relate too, as opposed to the scale values, or an emotional space reduced to four quadrants.

6.3.1 Regression Models as Classifiers

When the regression models were used to classify the music extracts into the five original clusters of the categorical model, a noticeable improvement was observed with the reduced pentagon circular model, while both the polygonal and two-dimensional decagon siblings offered significant improvements (see Table 6.7). It is interesting to note that the full dimensional model gave the worst results of all models for classification.

The polygonal circular models correctly classified *Life Round Here* (feat. Chance the Rapper) from James Blake. As for *Pursuit of Happiness (Steve Aoki Dance*

SMO	F-dim	FP-poly	RP-dim	RP-poly	D-dim	D-poly
58.9%	**51.50%**	60.57%	60.92%	**63.41%**	**78.50%**	**77.98%**

Table 6.7: Classification accuracy compared to original SMO model.

Remix), although the prediction error is small 0.093 and the target tag distance is one, the classification error remains, and the clip still sits at the boundary of class five.

Using the dimensional models for classification was done to see if the dimensionality, whether two-dimensional or circular, would help reduce the classification errors of the SVMs. However, it should be noted that the automatic mood classification task will always have the disadvantage of trying to put human emotion into boxes, forcing boundaries were there might be none. Also, forcing the dimensional models back into categorical ones will always results in a loss of finesse, as illustrated by the classification error on our second case study in Table 6.6.

The circularity of human emotional response has been proposed in scientific literature as early as 1935 [28], and later refined by several researchers [63, 64, 61, 68]. Most of these studies agree on the dimensionality of human emotion, and the dimensionality of the emotional response to music in particular. The two-dimensional arousal/valence regression approach was proposed in an attempt to represent the circularity of the problem, however the resulting mapping of the emotions by the regressors are lain over a plane, the annotations therefore typically rely on clustering tags in subareas of the plane, or translating the results into a classification problem, with the four quadrants being used as four classes. This has the effect of distorting the original assumptions of the circumplex model of emotion as proposed by Russell in [63] by not taking into account the full relationship between the two dimensions (for example the highest arousal point implies a neutral valence), and the circularity is therefore somewhat lost.

A continuous circular model of emotion offers both a flexible representative model, as well as a simple way to retrieve meaningful mood tag annotations that users can relate too.

Chapter 7

Conclusions

Despite the attention given to the music emotion recognition problem, there is still no system capable of annotating moods to music with great accuracy, solely based on audio features. Researchers in the field have explored different types of emotional models from psychological research, concentrating their efforts on categorical and two-dimensional models. Both have provided results satisfactory enough to show that the emotional response to music can be modelled, but none have reached the accuracies obtained in similar disciplines such as speech recognition. Several algorithms and combinations of audio features have been explored, but the results seem to have reached a ceiling just below 70%.

Categorical models have the advantage of presenting their annotators and system users with a terminology that they can relate to, but the task of drawing clear boundaries for each category either implies reducing the emotional space to a few basic terms, or eliminating ambiguous songs from datasets as they can't be confidently labelled for the ground truth. The exercise could be compared to creating categories to classify continuous shades of grey, where any demarcation will be arbitrary.

Dimensional models take into account the variability of the emotional response to music by the annotators and allow for a more flexible model encompassing the continuity of the emotional spectrum, but fail to provide a clear terminology and often reduce the predictions to four basic emotions. Moreover, the most commonly used dimensional model has one of its dimensions recurrently causing problems. Effectively, if the arousal dimension proves to be a good descriptor, one on which the annotators and users mostly agreed upon, the valence dimension turns out to be more subjective, and the regressor models fitted on such measures show weaker correlation.

This thesis explores the hypothesis that the use of a continuous circular model to

solve the emotion recognition problem might help push the accuracy ceiling higher. Such models have the advantages of providing: (1) a measure that takes into account the variability of the human experience, (2) their annotators with terms they relate to, and (3) a practical way to annotate multiple emotions and mood terms at once, without the burden of defining clusters and boundaries.

In summary, all evidence in this exploratory study points towards the circularity of the music emotion recognition problem and the limits of the current approaches to fully take advantage of this circularity. By building a dataset, analyzing the annotation process and experimenting with different types of models on the same dataset, I obtained results supporting the hypothesis that:

- the human response to emotion in music is better represented when modelled in a continuous matter
- this continuum is circular, as each emotion has more than one similar neighbour
- music emotion recognition systems provide better results when modelled in a continuous circular manner

Because there is no readily available implementation of algorithms to fit circular problems of such high dimensionality among the usual machine learning tools, I propose an approximative approach to circular regression. Polygonal circular regression provides a first step in the investigation of such problems, while the need of a full flexible implementation is yet to be fulfilled.

7.1 Future Work

This thesis presented evidence that the emotion recognition problem is better represented as a continuous circular model. However, many questions still require thorough investigation before the circular model can be defined. The first aspect in need of further research is how the emotions are distributed around the circle. During this investigation, the naive assumption that the clusters were normally distributed around the circle was made, and terms were placed following the circumplex model, but the circular model used for this work is by no means definitive, as many may argue that important mood tags are missing, while redundant ones have been used. Several methods should be considered to revise this distribution including:

- redistributing the classes asymmetrically around the circle (e.g. distributing the terms of cluster one over half the circle and the other four classes on the other half) to see how different distributions influences the results
- using graph algorithms to explore semantic similarity
- conducting circular statistical analysis on both the emotional and feature space.

The possibility that the data might lie on a torus is also worth investigating, as its added dimensionality over the circle might further refine the model. Such a distribution could be envisioned in other MIR tasks where similar concepts need to be delimited, as in genre classification for example, where a genre like blues is closely related to country, rock, pop and jazz. Of course, a thorough re-evaluation of the audio features in a circular context could also provide a better fit.

After drawing more informed conclusions on the distribution, another crucial step would be re-annotating the dataset. For example, the extract from *Motion Picture Soundtrack* Radiohead (Chapter 3.3) taken from 0:45 to 1:15 was labelled as class four by all eight annotators. It is reasonable to think that the mood tags motivating their decision were *Melancholic* and *Brooding* more than *Creepy* or *Paranoid*. So it is thus reasonable to think that the actual value of this clip would be closer to 3.625 if the same annotators were presented with the circular model used for this investigation, rather than the 4.0 calculated by taking the mean of the annotations. Also of interest is to find out how strongly annotators feel about their annotations. For example, it is reasonable to think that all eight annotators hesitated between class three and four before unanimously opting for four in the Radiohead song example.

The next step will be to implement a linear-circular regression model to fully measure the circularity of the problem and offer a full model instead of the multiple component models used during this investigation. A number of circular regression models have been proposed since the seventies: the *barbers pole* model in which the circular response, a curve spiralling up the surface of an infinite cylinder as the explanatory variable increases [49, 19, 45, 50], and a simpler model in which the response completes a single rotation as the explanatory variable increases [36, 17]. Finding the way to meaningfully fit such models while considering the high dimensionality of the feature space is an exciting challenge. As for the polygonal approximation of circular regression, a work on a formal definition and mathematical proof should be conducted to clearly establish its bounds and evaluate other possible applications.

Bibliography

[1] Roget's 21st Century Thesaurus, Third Edition. April 2014.

[2] Seung-Ryoel Baek and Moo Young Kim. Music genre/mood/ composer classification: Mirex 2014 submissions. 2014.

[3] Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. Music mood and theme classification-a hybrid approach. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 657–662, 2009.

[4] Densil Cabrera, Sam Ferguson, and Emery Schubert. 'psysound3': Software for acoustical and psychoacoustical analysis of sound recordings. 2007.

[5] William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2):210–229, 2006.

[6] Chuan Cao and Ming Li. Thinkit's submissions for mirex2009 audio music classification and similarity tasks. In *MIREX abstracts, International Conference on Music Information Retrieval*. Citeseer, 2009.

[7] June E. Downey. A musical experiment. *The American Journal of Psychology*, 9(1):pp. 63–69, 1897.

[8] J. Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[9] Tuomas Eerola and Jonna K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[10] Tuomas Eerola and Jonna K Vuoskoski. A review of music and emotion studies: approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, 30(3):307–340, 2013.

[11] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[12] Paul R Farnsworth. A study of the hevner adjective list. *Journal of Aesthetics and Art Criticism*, pages 97–103, 1954.

[13] Paul R Farnsworth. *The Social Psychology of Music.* Dryden, Oxford, England, 1958.

[14] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Music information retrieval by detecting mood via computational media aesthetics. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pages 235–241. IEEE, 2003.

[15] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Music information retrieval by detecting mood via computational media aesthetics. In *Web Intelligence, 2003. Proc. IEEE/WIC International Conference on*, pages 235–241. IEEE, 2003.

[16] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Popular music retrieval by detecting mood. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–376. ACM, 2003.

[17] Nicholas I Fisher and Alan J Lee. Regression models for an angular response. *Biometrics*, pages 665–677, 1992.

[18] Anders Friberg. Digital audio emotions: An overview of computer analysis and synthesis of emotions in music. In *11th International Conference on Digital Audio Effects, DAFx 2008, Espoo, Finland, 1-4 September 2008*, pages 1–6, 2008.

[19] A Lawrence Gould. A regression technique for angular variates. *Biometrics*, pages 683–700, 1969.

[20] John M Grey. *An Exploration of Musical Timbre.* Number 2. Dept. of Music, Stanford University, 1975.

[21] John M Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[22] John M Grey and James A Moorer. Perceptual evaluations of synthesized musical instrument tones. *The Journal of the Acoustical Society of America*, 62(2):454–462, 1977.

[23] Ralph H. Gundlach. A quantitative analysis of indian music. *The American Journal of Psychology*, 44(1):pp. 133–145, 1932.

[24] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[25] Byeong-Jun Han, Seungmin Ho, Roger B Dannenberg, and Eenjun Hwang. SMERS: Music emotion recognition using support vector regression. In *Proc. of the Intl. Symposium on Music Information Retrieval*, 2009.

[26] Trevor Hastie, Robert Tibshirani, et al. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

[27] Kate Hevner. The affective character of the major and minor modes in music. *The American Journal of Psychology*, pages 103–118, 1935.

[28] Kate Hevner. Expression in music: a discussion of experimental studies and theories. *Psychological Review*, 42(2):186, 1935.

[29] Kate Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, pages 246–268, 1936.

[30] Kate Hevner. The affective value of pitch and tempo in music. *The American Journal of Psychology*, pages 621–630, 1937.

[31] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[32] Xiao Hu and J Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proc. of the Intl. Symposium on Music Information Retrieval*, pages 67–72, 2007.

[33] Xiao Hu, J Stephen Downie, Cyril Laurier, Mert Bay, and Andreas F Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 462–467, 2008.

[34] David Huron. Perceptual and cognitive applications in music information retrieval. *Perception*, 10(1):83–92, 2000.

[35] Carroll E Izard. Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, 99(3):561–565, 1992.

[36] Richard A Johnson and Thomas E Wehrly. Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, 73(363):602–606, 1978.

[37] Philip N Johnson-Laird and Keith Oatley. Basic emotions, rationality, and folk theory. *Cognition & Emotion*, 6(3-4):201–223, 1992.

[38] JungHyun Kim, Seungjae Lee, SungMin Kim, and Won Young Yoo. Music mood classification model based on arousal-valence values. In *Advanced Communication Technology (ICACT), 13th Intl. Conf. on*, pages 292–295. IEEE, 2011.

[39] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. of the Intl. Symposium on Music Information Retrieval*, pages 255–266, 2010.

[40] Mark D Korhonen, David Clausi, M Jernigan, et al. Modeling emotional content of music using system identification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(3):588–599, 2005.

[41] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[42] Cyril Laurier, Perfecto Herrera, M Mandel, and D Ellis. Audio music mood classification using support vector machine. *MIREX task on Audio Mood Classification*, pages 2–4, 2007.

[43] Cyril Laurier, Olivier Lartillot, Tuomas Eerola, and Petri Toiviainen. Exploring relationships between audio features and emotion in music. In *ESCOM: Conf. of European Society for the Cognitive Sciences of Music*, 2009.

[44] Cyril Laurier, Owen Meyers, Joan Serrà, Martin Blech, Perfecto Herrera, and Xavier Serra. Indexing music by mood: Design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, 48(1):161–184, 2010.

[45] P J Laycock. Optimal design: regression models for directions. *Biometrika*, 62(2):305–311, 1975.

[46] Tao Li and Mitsunori Ogihara. Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, volume 3, pages 239–240, 2003.

[47] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):5–18, Jan 2006.

[48] Karl F MacDorman, Stuart Ough, and Chin-Chang Ho. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, 36(4):281–299, 2007.

[49] J. K. Mackenzie. The estimation of an orientation relationship. *Acta Crystallographica*, 10(1):61–62, Jan 1957.

[50] Kanti V Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 349–393, 1975.

[51] Owen Craigie Meyers. *A Mood-based Music Classification and Exploration System*. PhD thesis, Massachusetts Institute of Technology, 2007.

[52] Keith Oatley. *Best Laid Schemes: The Psychology of the Emotions*. Cambridge University Press, 1992.

[53] Francois Pachet and Jean-Julien Aucouturier. Improving timbre similarity: How high is the sky. *Journal of Negative Results in Speech and Audio Sciences*, 1(1):1–13, 2004.

[54] R Panda, B Rocha, and RP Paiva. Dimensional music emotion recognition: combining standard and melodic audio features. In *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 583–593, 2013.

[55] Renato Panda and Rui Pedro Paiva. Using support vector machines for automatic mood tracking in audio music. In *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[56] Renato Panda and Rui Pedro Paiva. Music emotion classification: Dataset acquisition and comparative analysis. In *15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.

[57] Renato Panda, Bruno Rocha, and Rui Pedro Paiva. Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence*, 29(4):313–334, 2015.

[58] Jaak Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford university press, 1998.

[59] Geoffroy Peeters. A generic training and classification system for mirex08 classification tasks: Audio music mood, audio genre, audio artist and audio tag. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 08)*, 2008.

[60] John Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods: Support Vector Learning*, 3, 1999.

[61] Robert Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harpercollins College Division, New York, USA, 1980.

[62] J. Ren, M. Wu, and J. Jang. Automatic music mood classification based on timbre and modulation features. *Affective Computing, IEEE Transactions on*, PP(99):1–1, 2015.

[63] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[64] Emery Schubert. Update of the Hevner adjective checklist. *Perceptual and Motor Skills*, 96(3c):1117–1122, 2003.

[65] Janto Skowronek, Martin F McKinney, and Steven Van De Par. Ground truth for automatic music mood classification. In *Proc. of the Intl. Symposium on Music Information Retrieval*, pages 395–396, 2006.

[66] Janto Skowronek, Martin F McKinney, and Steven Van De Par. A demonstrator for automatic music mood estimation. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 345–346, 2007.

[67] John A Sloboda and Patrik N Juslin. Psychological perspectives on music and emotion. In Patrik N Juslin and John A Sloboda, editors, *Music and Emotion: Theory and Research.*, pages 71–104. Oxford University Press, 2001.

[68] Robert E Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, Oxford, England, 1989.

[69] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *Proc. of the Intl. Symposium on Music Information Retrieval*, pages 325–330, 2008.

[70] George Tzanetakis. Marsyas-0.2: a case study in implementing music information retrieval systems. *Intelligent Music Information Systems. IGI Global*, 14, 2007.

[71] George Tzanetakis. Marsyas submissions to mirex 2009. *Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

[72] George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organised Sound*, 4(03):169–175, 2000.

[73] Vladimir Naumovich Vapnik and Samuel Kotz. *Estimation of Dependences based on Empirical Data*, volume 40. Springer-verlag New York, 1982.

[74] Nicolas Wack, Enric Guaus, C Laurier, O Meyers, R Marxer, D Bogdanov, Joan Serra, and P Herrera. Music type groupers (mtg): generic music classification algorithms. *Music Information Retrieval Evaluation eX-change (MIREX) extended abstract*, 2009.

[75] Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):40, 2012.

[76] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. A regression approach to music emotion recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):448–457, 2008.