

# **Modeling of Emotions expressed in Music using Audio features**

Jens Madsen

Kongens Lyngby 2011  
MSC-EA-2011-1

Technical University of Denmark

Informatics and Mathematical Modelling  
Building 321, DK-2800 Kongens Lyngby, Denmark  
Phone +45 45253351, Fax +45 45882673  
[reception@imm.dtu.dk](mailto:reception@imm.dtu.dk)  
[www.imm.dtu.dk](http://www.imm.dtu.dk)

DTU Electrical Engineering  
Acoustical Technology  
Ørsteds Plads, building 352, DK-2800 Kongens Lyngby, Denmark  
[www.elektro.dtu.dk](http://www.elektro.dtu.dk)

MSC-EA: ISSN 0909-3192

# Abstract

---

This thesis presents an alternative method of organizing and rating music, using the emotions expressed in music. This measure can serve as a standalone parameter for searching for new music or in combination with already established methods e.g. *happy jazz* or *sad rock*. The approach is to create a mathematical model that automatically can predict labels of emotional expression in music based on audio content. The audio content is quantified using audio features using spectral, cepstral, temporal, musical and perceptual features computed from 7 different feature packs. To measure the emotions expressed in music a listening experiment is developed using experimental design. Participants rate excerpt of 15 seconds on two 9-point iconic scale (*SAM*) representing the dimensions of valence and arousal. All ratings are modeled using fitted beta distributions, where outliers are removed appropriately based on empirical measures. A thorough investigation into the consequences of the design and the resulting ratings is made. Furthermore the influence of participants' musical experience, their mood before starting the test and understanding of the test are investigated if there is a connection to their emotional ratings.

Using audio features and emotional ratings a mathematical model is designed where the best performing is a stepwise regression model trained on features selected by a Sequential feature selection method using Least Squares and Root Mean Squared Error. The most suitable features are found to model emotions in music that include MFCC, Pulse Clarity, Main Loudness, Pulse Clarity, Spectral Flatness per. band, Inharmonicity and *CENS*. Compared to a formulated baseline error measure the model performs 15 % and 47 % better for valence and arousal respectively. Resulting in an average error of 0.727 ratings on the arousal scale and valence of 0.887 ratings given that participants rated on a 9 point scale. The model can be used to predict emotional labels for greater datasets for future testing or to predict ratings on a shorter time scale to group musical excerpt based on the dynamic emotional structure in music.



# Resumé

---

Denne opgave præsenterer en alternativ metode til at organisere og ordne musik ved hjælp af det emotionelle indhold udtrykt i musik. Denne målemetode kan bruges som et selvstændigt parameter til at søge efter ny musik eller i kombination med allerede etablerede metoder som f.eks. ”glad jazz” eller ”sørgelig rock”. Fremgangsmåden er at lave en matematisk model, der automatisk kan forudsige værdier for følelser udtrykt i musik baseret på lydindholdet. Lydindholdet er kvantificeret ved hjælp af audio features (spektral, cepstral, temporal, musikalske og perceptuelle), der er beregnet ud fra 7 forskellige pakker af implementeringer. Deltagere blev bedt om at bedømme musikklip af 15 sekunders varighed, på en 9 punkts ikonisk skala, der repræsenterer de emotionelle dimensioner ophidselse (opstemt-ikke opstemt) og valens (positiv-negativ). Alle bedømmelser er modeleret ved brug af beta fordelinger, hvor afvigere er fjernet ved hjælp af empiriske metoder. En grundig undersøgelse af konsekvenserne af det udviklede design og de heraf følgende bedømmelser er udført. Derudover bliver deltagernes musiske baggrund, deres humør inden testens start og forståelse af eksperimentet undersøgt og set, om der er en forbindelse til deres emotionelle bedømmelser. Ved brug af audio features og de emotionelle vurderinger er der designet en matematisk model, hvor den model, der klarer sig bedst, er en trinvis regression-model, der er trænet på audio features udvalgt ved hjælp af Sequential feature selection metode, der bruger Least Squares og Root Mean Squared Error. De bedst egnede audio features til at modellere det udtrykte emotionelle indhold i musik er MFCC, pulse clarity, main Loudness, pulse clarity, spectral flatness per. band, inharmonicity and *CENS*, blandt mange andre. Sammenlignet med et formuleret standard fejlmål resulterer den udviklede model i, at den klarer sig henholdsvis 15 % og 47 % bedre for valens og ophidselse. Dette resulterer i en gennemsnitlig fejl på 0.727 trin for ophidselse og 0.887 trin på valens skalaen, forudsat at deltagere bedømmer på en 9-punkt skala. Den udviklede model kan bruges til at forudsige emotionelle værdier for ny musik. Dette kan bruges til

at udvælge egnede musikklip til at bedømme i lytteeksperimenter i fremtiden. Den kan også bruges til at forudsige emotionelle værdier på en kortere tidsskala, der kan bruges til at gruppere musikklip baseret på den dynamiske emotionelle struktur i musik.

# Preface

---

This thesis was prepared at Cognitive Systems group at DTU Informatics in collaboration with department of Acoustical technology at the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Master of Engineering Acoustic degree. The work was carried out from December 2010 - June 2011, having a workload of 30 ECTS credits.

The thesis deals with different aspects of mathematical modeling of systems using data and partial knowledge about the structure of the systems. The main focus is on the extraction of audio features in music and using these to model emotions expressed in music. Obtained by a listening experiment developed using experimental design.

Lyngby, June 2011

Jens Madsen





# Acknowledgments

---

I would like to thank the employees of the Cognitive Systems group, DTU Informatics, for letting me make my master thesis there. The entire group has been very patient and helpful in all aspects of the project and has assisted with many productive discussions.

I would like to thank Professor Lars Kai Hansen, DTU Informatics for his invaluable support in the period up to the start of the project. Additionally i would like to extend my great appreciation to Professor Henrik Spliid for his invaluable help within statistical measures and approaches in this thesis. I would like to thank Associate Professor Jan Larsen for overall guidance and directions within the project. I would also like to thank Associate Professor Michael Kai Petersen for helpful discussion in the field of music psychology. I would like to extend my appreciation to all the students and employers that participated in the listening experiments that founded the basis for this project. Lastly i would like to thank my family and friends who have stood by me through some difficult times.

Thank you all.



# Contents

---

<b>Abstract</b>	<b>i</b>
<b>Resumé</b>	<b>iii</b>
<b>Preface</b>	<b>v</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Previous work . . . . .	2
1.3 Modeling of emotions in music . . . . .	5
1.4 Problem statement . . . . .	7
1.5 Approach . . . . .	7
1.6 Thesis layout . . . . .	7
<b>2 Analysis</b>	<b>9</b>
2.1 Music . . . . .	9
2.1.1 Musical Structure . . . . .	10
2.1.2 Musical descriptors . . . . .	10
2.1.3 Music and Lyrics . . . . .	12
2.2 Emotion . . . . .	12
2.2.1 Definition of Emotion . . . . .	12
2.2.2 Models of emotions . . . . .	13
2.2.3 Quantification of emotions . . . . .	15
2.3 Emotions and Music . . . . .	18
2.3.1 Emotions in Music . . . . .	18
2.3.2 Perception vs. Induction of emotion . . . . .	19
2.3.3 Models of emotions in music . . . . .	22
2.3.4 Acoustical cues and emotional response . . . . .	23

2.3.5	Temporal emotional dimension . . . . .	25
2.4	Problem specification . . . . .	26
<b>3</b>	<b>Audio features</b>	<b>31</b>
3.1	Initial considerations . . . . .	31
3.2	Separation of audio data . . . . .	33
3.2.1	Temporal separation . . . . .	33
3.2.2	Spectral separation . . . . .	34
3.3	Spectral features . . . . .	35
3.4	Temporal features . . . . .	38
3.5	Cepstral features . . . . .	40
3.6	Perceptual features . . . . .	41
3.7	Musical Features . . . . .	43
3.8	Misc. features . . . . .	45
3.9	Post-processing . . . . .	45
3.9.1	Alignment of features . . . . .	46
3.9.2	Effect of Lossy-compression . . . . .	47
3.9.3	Output corrections . . . . .	47
3.10	Conclusion . . . . .	47
<b>4</b>	<b>Listening experiment</b>	<b>49</b>
4.1	Experimental considerations . . . . .	49
4.2	Design of listening experiments . . . . .	52
4.3	Quantification of emotion . . . . .	53
4.3.1	Response attribute . . . . .	53
4.3.2	Response format . . . . .	53
4.4	Common experimental variables . . . . .	54
4.4.1	User interface . . . . .	54
4.4.2	Instructions . . . . .	55
4.4.3	Reproduction system . . . . .	55
4.4.4	Listening room . . . . .	55
4.4.5	Calibration . . . . .	55
4.4.6	Order of presentation . . . . .	56
4.5	Pilot 1 . . . . .	56
4.5.1	Test paradigm . . . . .	56
4.5.2	Subjects . . . . .	57
4.5.3	Stimuli . . . . .	57
4.5.4	Order of stimuli . . . . .	57
4.5.5	Data processing . . . . .	58
4.5.6	Meta data . . . . .	58
4.5.7	Results . . . . .	59
4.5.8	Discussion . . . . .	62
4.6	Pilot 2 . . . . .	63
4.6.1	Test paradigm . . . . .	63

4.6.2	Subjects . . . . .	64
4.6.3	Stimuli . . . . .	64
4.6.4	Order of stimuli . . . . .	65
4.6.5	Instructions . . . . .	65
4.6.6	Data processing . . . . .	65
4.6.7	Results . . . . .	69
4.6.8	Discussion . . . . .	75
4.7	Conclusion . . . . .	77
<b>5</b>	<b>Mathematical model</b>	<b>79</b>
5.1	Initial considerations . . . . .	79
5.2	Pre-analysis of data . . . . .	80
5.2.1	Audio features . . . . .	80
5.2.2	Target labels . . . . .	81
5.3	Labels and features . . . . .	81
5.4	Selection of Model . . . . .	83
5.5	Linear regression . . . . .	83
5.5.1	$L_2$ -regularized regression . . . . .	84
5.5.2	$L_1$ -regularized regression . . . . .	85
5.5.3	Stepwise regression . . . . .	87
5.5.4	Sequential Feature selection . . . . .	88
5.6	Error measures . . . . .	88
5.6.1	Root Mean Squared Error . . . . .	89
5.6.2	Euclidean distance between means . . . . .	89
5.6.3	Kullback Leibler divergence . . . . .	90
5.6.4	Baseline error . . . . .	90
5.7	Cross-validation . . . . .	91
5.7.1	K-fold . . . . .	91
5.7.2	Temporal issues . . . . .	92
5.8	Results . . . . .	93
5.8.1	Baseline error . . . . .	93
5.8.2	Sequential feature selection . . . . .	93
5.8.3	Predictions and errors . . . . .	94
5.8.4	$LR$ - Beta mean . . . . .	95
5.8.5	$LR$ - Beta distributions . . . . .	99
5.9	Post data analysis . . . . .	105
5.9.1	Emotional ratings and audio features . . . . .	105
5.9.2	Temporal modeling of emotions . . . . .	105
5.10	Discussion . . . . .	106
5.11	Conclusion . . . . .	108
<b>6</b>	<b>Conclusion</b>	<b>109</b>
6.1	Discussion . . . . .	109
6.2	Summary . . . . .	112

<b>A</b>	<b>Analysis</b>	<b>113</b>
A.1	Musical Descriptors . . . . .	113
<b>B</b>	<b>Audio features</b>	<b>117</b>
B.1	Overview of features . . . . .	117
B.2	Effect of MP3 encoding on audio feature extraction . . . . .	122
B.2.1	Results . . . . .	123
B.2.2	Discussion . . . . .	133
B.3	Effect of resampling on audio features . . . . .	134
B.3.1	Results . . . . .	136
B.3.2	Discussion . . . . .	139
B.4	Effect of <i>NaN</i> painting . . . . .	140
B.4.1	Results . . . . .	140
B.4.2	Discussion . . . . .	141
<b>C</b>	<b>Listening experiment</b>	<b>143</b>
C.1	Pilot1 - Graphical interface . . . . .	143
C.1.1	Instructions . . . . .	143
C.1.2	Meta data . . . . .	146
C.1.3	Prior mood . . . . .	148
C.1.4	Primary interface . . . . .	148
C.2	Pilot1 - Meta data analysis . . . . .	149
C.2.1	Temporal analysis . . . . .	149
C.2.2	Analysis of scales . . . . .	150
C.2.3	Analysis of excerpt length . . . . .	150
C.3	Pilot2 - User interface . . . . .	152
C.3.1	Meta data . . . . .	152
C.4	Pilot2 - Bitrates of musical data . . . . .	152
C.5	Pilot2 - All ratings . . . . .	154
C.5.1	Arousal . . . . .	154
C.5.2	Valence . . . . .	156
C.6	Pilot2 - Pre-emotional ratings . . . . .	158
C.6.1	Arousal . . . . .	158
C.6.2	Valence . . . . .	159
C.7	Pilot2 - Analysis of pre-emotional ratings . . . . .	160
C.8	Pilot2 - <i>OC1</i> data foundation . . . . .	160
C.9	Pilot2 - Outlier removal . . . . .	162
C.10	Pilot2 - Distribution of beta mean . . . . .	163
C.11	Pilot2 - Meta data analysis . . . . .	164
C.11.1	Temporal analysis . . . . .	164
C.11.2	Rating of scales . . . . .	165
C.11.3	Musical background . . . . .	166
C.12	Pilot2 - Meta data influence on ratings . . . . .	169

---

<b>D</b>	<b>Mathematical modeling</b>	<b>171</b>
D.1	Features selected by <i>SFS</i>	171
D.2	Features selected by <i>LARS</i>	173
D.3	Features selected by <i>stepwise</i>	173
D.4	Emotional ratings and audio features	175
D.4.1	Results	176
D.4.2	Discussion	177
D.5	Temporal emotional modeling	178
D.5.1	Results	178
D.5.2	Discussion	181





# Introduction

---

## 1.1 Background

Music has for many years been a means of entertainment for people around the world. People listen to music when they need to cheer up, when they are happy and want to dance, when they want to remember a loved one and feeling sad. There are numerous occasions where music fits in naturally in our everyday life. What exactly in music makes us prefer one track over another and what makes us think that it is a e.g. happy or a sad song. These questions have puzzled researchers and music producers for many years. One might say that music is a form of communication or a way of communicating a specific feeling or emotion, by means of expression. Often a person listens to a musical track and can instantly tell you if the person likes it or not. If they have heard it before, know someone that likes it, or everyone else likes it, etc. also influences if the person prefers this track or not.

Finding new music has changed throughout history from going to the local record store and ask for what the new music was and discovering what is new through advertising in different media, such as radio, television, newspapers including reviews of music and mouth to mouth through people you know.

In modern time the Internet has revolutionized the means of communicating and advertising for different music. Now the availability is not limited due to the music the local distributor or shops buy, but through the Internet vast amounts of music have become available to the general public. As the world becomes more global and people are exposed to many different styles and genres of music, musicians are inspired and genres merge. Before genres as rock, jazz

and pop were perhaps clearly defined or separated, but now genres such as pop-rock, pop-jazz and jazz-rock are emerging. Similarly artists are not only defined by one genre but are often described by a great number of different genres. This makes the recommendation and music search increasingly difficult.

Numerous search engines such as [www.itunes.com](http://www.itunes.com) and [www.amazon.com](http://www.amazon.com) etc. are emerging in order to fill holes in the market so that music becomes fast and easily available to people. New approaches to finding music is the fast arising user communities such as [www.last.fm](http://www.last.fm) where users indicate what genre, mood or impression they got from an artist or song by tagging. They further use programs such as [www.audioscrobbler.com](http://www.audioscrobbler.com) to track peoples listening habits in order to make more specific recommendations to the user. For communities such as *Last.fm* the sheer number of people voting/tagging a track or artist makes it more likely it will be recommended to you. This is an inherent weakness in finding relatively unknown music, since a great number of tags forming a *tag – cloud* are needed before they will be recommended. Instead of users rating or categorizing music experts also have their websites where they can recommend music such as [www.pandora.com](http://www.pandora.com) and [www.allmusic.com](http://www.allmusic.com) (AMG) where genres, moods, styles and themes are rated. Common for these methods of finding music you like is the dependence of other people have listened to it before you.

As previously mentioned genres merge and become ever more difficult to distinguish such as the approximately 840 genres represented in *AMG*. For this reason an alternative method of classifying music tracks is proposed in the form of emotional content. This information could be used cross genres as a supplement to the already existing descriptives such as artist, track name, album, year, genre, etc. That could be happy jazz music from the 80's or all sad Beatles music.

## 1.2 Previous work

The paradigm is that there is an emotional reaction in humans due to music, this reaction can be measured using an experimental setup. This emotional reaction can be modeled using structural information about music. Many different approaches have been made both in measuring emotional reaction elicited by music, further to describe music in a descriptive language and last the use of many different mathematical models. Almost all work that has been done in the past differs in the way they represent emotions. There are two major fractions within this modeling and that is the categorical modeling that is based on mutually exclusive emotions (e.g. happy, sad, angry, aggressive, etc.) that is often modeled using classification. The other is the dimensional model where emotions are placed in a plane spanned by dimensions such as Valence and Arousal. These types are often modeled using regression techniques. Indirectly the way you measure or obtain emotional data then chooses the mathematical model

chosen to model this. The work done within recent years in acquiring emotional data, what mathematical models chosen and structural musical information is then the main choices.

### Emotional data

Work has been dominated by two different methods of obtaining descriptions of humans emotional reaction to music. One being self-report methods using a group of test participants to rate music stimuli and the other being the gathering of so called *tags* from social websites. In [Hu and Downie, 2007] they create a dataset of metadata from *AMG*, *www.epinions.com* and *Last.fm*. The *AMG* has 183 different mood labels that are said to describe the song, album or “over all body of work”, these include happy, sad, druggy, nocturnal, rollicking, wry etc. By webmining *AMG* they obtain a data set and divide the tags into five mood clusters. They combine genre and mood tags, usage-statistics and mood tags and last artist and mood tags to form data sets. They show that for artist and genre a dataset can be constructed verified by using corroborative data analysis from *Last.fm* showing that their *AMG* data set is stable enough to be constructed.

In [Laurier et al., 2009c] they create a semantic mood space using *Last.fm* tags searching for 120 semantic mood terms among 7 mil. tags from over 0.5 mil. songs. Out of the 120 adjectives a subset of 80 is used frequently, and out of the 0.5 mil. tracks found with tags a subset of 60 thousand tracks has multiple mood tags. Thus showing that these tags are rarely used by users tagging music. Given so many sources of emotional data from web sources a possibility is to combine them as was proposed in [Turnbull et al., 2007] they combine information from social tags, webmining, surveys, autotags and listening games using Rank based interleaving. Similar approach is done in [Laurier et al., 2009b].

For these simple types of metadata it is expected that people have some sort of agreement. A problem is that the interface can inherently cause bias. Initially a blank cloud is present when a new song or tracks is present on e.g. *Last.fm*. Users then choose themselves what to tag, the system itself then shows to other users before they tag what other people have tagged before. Making it much easier to reach an agreement with other users, creating a bias. Another issue is the tags themselves, what are they a measure of when using them for e.g. as emotional rating. No real instructions are given on these pages, to what is meant, so the information is vague.

Self-report methods using listening experiments have also been popular and in recent years some work has been done in the *MIR* community. The *Audio Mood Classification* (AMC) in the *MIREX* competition that is a part of the *ISMIR* conferences has obtained a dataset of mood labels which is reviewed in [Hu et al., 2008]. They use 5 clusters of emotional semantic descriptors including Rowdy, Sweet, Literate, Witty, Volatile, etc. to describe those 5 clusters. They use the developed *Evaluatron* 6000, a web-based device to annotate the musical excerpts by 1250 candidates. Each 30-seconds excerpt is rated using

a single label for the whole track, which is chosen in the middle of each song to reduce change in emotional content along the duration of the track. Their criteria is 3 or 2 agreements between participants of mood cluster, where in average they achieve 67.8% agreements and 32.2% disagreement in a total of 864 votes. In [Schmidt and Kim, 2010] they use *Moodswings*, a self-developed interface for self-report rating within a self developed 2-dimensional emotional model. Using 150.000 ratings of 1000 songs from the *USPOP2002* data set they create a model. This model is used to sample the entire dataset to create a subset of 240 15 second excerpts chosen to approximate an even distribution across the four primary quadrants of the A-V space. These excerpts are then subjected to intense focus, to provide a significantly higher amount of ratings. They argue that emotions change through time and therefore create a model that models this emotional change using post-ratings.

### Mathematical models

Different mathematical models and approaches have been used to model emotions in music. In [Yang et al., 2008] they reach the best performance using a *Support Vector Machines* (SVM) as the regressors where features are selected using *RReliefF* and correlation between Arousal and Valence is reduced using *Principal Component Analysis* (PCA). The best performance using  $R^2$  statistics reaches 58.3% for arousal and 28.1% for valence.

In [Eerola et al., 2009] they compare 3 data reduction algorithms the *Stepwise Regression* (SR), *PCA* and *Partial Least Squares* (PLS) regression. Their best linear model is the *PLS* model using Box-Cox transformed variables, that account for 72% for Valence, 85% for Activity and 79% for tension only using soundtrack music.

In [Hu et al., 2008] they review 9 different approaches and they find that *SVM* is the approach that reaches the highest performance of classification with different implementations. *LibSVM* reaches 61% average accuracy, *WekaSMO* 58% and *DAG - SVM* 57%, using a broad range of genres. Other approaches to model the time varying emotional content in music is done by [Schmidt and Kim, 2010] where they compare *Multivariate Linear Regression* (MLR), *PLS*, *Support Vector Regression* (SVR) approaches first to predict the distribution for 15-second clips and subsequently shorten the analysis window to follow 1-second clips within the A-V space. Each musical excerpt is modeled with a single 2D-Gaussian on the A-V space, where each point is a 1-second projection. They find that using a combination of *MLR* in multiple stages provide the best performance.

The issue is that all use a different emotional model and furthermore rating scales, so comparison between them is very difficult. The emotional measuring methods also differ so ratings or labels are obtained in many different ways, also making it difficult to compare.

### Structural musical information

Within the *MIR* community both audio, MIDI and lyrical features have been used to structurally describe music. Audio and acoustical features have been the dominating source since the data is always present regardless of there is singing. In [Kim et al., 2010] they review the most recent work within the modeling of emotions in music. They show that the most frequent used features include, RMS energy, MFCCs, spectral shape, spectral contrast, roughness, harmonic change, key clarity, majoriness, chromagram, chroma centroid and deviation, rhythm strength, regularity, tempo, beat histograms, event density, attack slope and attack time. In previous years a search of the main contributing features has been to course in the modeling emotions in music, but this approach has not shown any dominating factor. In recent years the course has been to use multiple features subsequently employing dimensionality reduction techniques or feature fusion. Regardless of feature fusion or dimensionality reduction methods, the most successful systems combine multiple audio feature types.

The other major contributor to describing music structure has been lyrics. They can be used as the only feature foundation or as a multi-modal modeling which was investigated in [Lu et al., 2010] 26 audio features resulting in 79 dimensions are used from *jAudio*, 102 dimension of MIDI and lyrical features using a *Bag-Of-Words* (BOW) approach constructing both uni- and bi-gram feature set using *TD-IDF*. The best performance was found using only audio features with 59.8%, MIDI features only with 58.6% and lyrical features only with 49.1% of classification error. When combined all three in combination scores highest by a margin resulting in 72.4% audio and lyrics 72%, MIDI and lyrics 71.2% and MIDI and audio 61.2% suggesting that MIDI gives the least amount of contribution to classification accuracy. In [Hu and Downie, 2010] they also compare lyrics and audio features performance in classification. Using 63 spectral audio features from *MARSYAS* for the audio and *BOW* approach in the lyrical feature set using content words (CW) that was constructed from different N-Grams (Uni, Bi and Tri). They show that for their 18 mood categories, seven of the lyrical feature types significantly outperform audio only categories. Where only one audio feature outperformed all lyric-based features. They argue that their work was limited due to only using audio spectral features, where other audio dimension should be considered in the future.

## 1.3 Modeling of emotions in music

The approach to model the emotional reaction by human to music is taken in the direction of using emotional data that is acquired by self-designed experiments. The sheer amount of data in the different social webs seem promising in the *MIR* community. When modeling emotional content in music, that potentially is changing through time, by using “1-tag fits all” data such as *Last.fm* or others could be utilized by a mean effect. The problem is the bias which is inherent in these types of websites, furthermore no verification or supervision

can be made on data or experimental variables.

Choices of emotional models, structural musical data or mathematical models to make this model cannot be made at this point. This is due to the great variation in different error measures, measuring techniques, emotional models, etc. used. Therefore this has to be analyzed further.

## 1.4 Problem statement

The main problem of this project is

- The development of a method of extracting emotional information from music.
- Finding a suitable emotional model to represent emotions in music.
- The extraction of structural data from musical tracks.
- To find specific structural information that describes emotions in music.
- To develop a mathematical model to use structural data from music to predict emotional content in music.
- To verify the model performance using appropriate baseline error measure.

## 1.5 Approach

Since this project is very multidisciplinary the approach taken is to break down the project into subproblems. Each subproblem will be handled in separate chapters where results, discussion and part conclusions will be made for each topic. Each topic can therefore be seen as independent chapters and can be read as such. Concerns and thoughts raised along the research into the modeling of emotions expressed in music, will be dealt with by introducing sections labeled e.g. *initial considerations* in each chapter.

## 1.6 Thesis layout

- Chapter 2 contains an analysis of the aspect necessary to formulate a problem specification based on the given problem statement. The aspects of music, emotions and the cross field of emotions and music. The models that describe them, the effect that music has on emotions, both expressed and induced and how to quantify these will be presented.
- Chapter 3 as a result of the analysis audio and acoustical features are chosen to describe music. In this section spectral, temporal, cepstral, musical, perceptual features etc. will be presented. Furthermore an investigation into what the effect of lossy audio compression, temporal alignment and output corrections have on audio features is presented.
- Chapter 4 presents all the aspects of designing, executing and lastly the results of a listening experiment, that has the purpose of collecting emotional ratings of music. The influence of participants mood, personal data and understanding of the test will be investigated.

- Chapter 5 introduces different mathematical models, error measures, cross validation schemes and presents the results of the modeling. Features appropriate for modeling emotions in music are investigated, and clustering of musical excerpts is attempted based on emotional dynamics.

Throughout the entire report the use of the appendix will be made in tight connection with the main report. Cross references will be made often so for convenience the appendix should be read in parallel to the main report to ease the reading.

The structure of the appendix is that each chapter, e.g. 2, 3, 4 and 5 will have an appendix for each i.e. A, B, C and D.

- Appendix A contains additional information regarding musical descriptors.
- Appendix B contains an overview of features, an investigation into the effect of lossy-compression, resampling of features and output corrections have on audio features.
- Appendix C contains the GUI used for both listening experiments, meta data analysis and ratings by participants.
- Appendix D contains a list of the audio features chosen by 3 different feature selection methods, an investigation into the correlation between audio features and emotional ratings. Furthermore a clustering of musical excerpts based on emotional dynamics.



# Analysis

In order to find a method to solve the problem statement an analysis of relevant areas within music and psychology will be presented. In figure 2.1 a simplified schematic is shown that represents the knowledge prior to the analysis. A musical source communicates to subjects through some channel, e.g. a concert, headphones in the train, etc. The music reaches the subject which, using all senses, perceives this music and based on this perception results in some emotional reaction.

Therefore an investigation of music, how it's structured and what describes

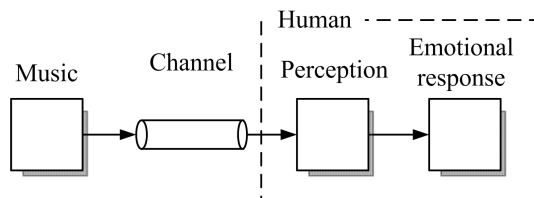


Figure 2.1: Simplified schematic of the structure that is the aim to model in this thesis.

it should be made. The influence of the channel, e.g. performance, acoustical properties of playback, etc. on the resulting emotional response should be investigated. The human perception of sound and the resulting emotional response should also be analyzed. Further a definition of emotions and the emotional model that is needed in order to quantify them, should be analyzed. In this section these elements will be discussed.

## 2.1 Music

Music as a term has been broadened through the evolution of humans, even in the past few hundred years. From *Wolfgang Amadeus Mozart* in the 18th

century producing classical music, to the the percussion group Stomp<sup>1</sup> that plays on oil barrels, plastic bags, metal bars, etc. Defining music is not trivial due to its many varieties, even on a subjective level people enjoying *Mozart* would not enjoy *Thrash metal* artists such as *Biohazard* or the music genre noted as *noise*. The distinction between music and noise lies not in the sound, but in the way human beings make use of it [Hallam et al., 2009]. From an acoustical point of view that is not pragmatic when applied, since the judgment of an acoustical signal being music, then becomes subjective. This approach is a two sided matter in the music industry, how to judge what music is. Traditionally a bottom-up approach has been used to determine what genre music is. Bottom-up in the sense that the judgment of whether it was rock or pop was determined by the musical instruments being played in the track, and the way they were played. A more modern means of determining genre and music classification is using social aspects about an artist, where these metadata regarding an artist and the society they are represented in determines the genre. This applies to all genres and types of music, so artists can still apply noise or pure voice and still be classified as music. What we think of it is another matter and one just has to remember that music is that art form which medium is sound, whatever that is.

### 2.1.1 Musical Structure

In general a structure exists in music, often adapted from the classical music. Western tonal music has developed a notation that represents pitch and duration information fairly explicit, but intensity and tone quality only approximately. Other relationships, such as group boundaries, metrical levels higher than the measure, and patterns of motion, tension and relaxation are unspecified [Palmer, 1997]. This view might not apply to all types of music but leaves the room for personal expression of music, and many layers of coding messages to the listener. In modern rock, pop, jazz, etc sheet music might not be the dominant way of going about making music. “Jamming” and playing what sounds right is an approach that is often used. This is very genre and culture specific. Whereas modern pop music is rarely played live by musicians and is very much “produced” and manipulated leaving small room for acoustical expression. Even though modern music has changed a lot, structure with intro, bridge, chorus and verse etc. is still very much used, both musically and lyrically.

### 2.1.2 Musical descriptors

Music performance is often seen as a means of communicating an idea or expression from a composer. A piece of music can be noted in some systematic fashion and decoded and re-encoded by the artist playing the piece of music. The audience then decodes the acoustical signal produced by the musicians into some ideas [Palmer, 1997]. This acoustical signal consist of the music played

---

<sup>1</sup><http://www.stomponline.com>

by instruments, the lyrics of the song and the expression of the vocal. Often different descriptive words are associated with music such as Pitch, Amplitude, Register, Harmonics, Harmony, Tonality, Brightness, Timbre, Loudness, Roughness, Tone attack/voice onset, Tempo/Speech rate, Articulation/pauses, Rhythm/meter/mode, Jitter/vibrato which are described in greater detail in section A.1. These are used throughout music psychology and by musicians. In [Leman et al., 2005] a descriptive hierarchy is adopted to distinguish between the level of descriptors.

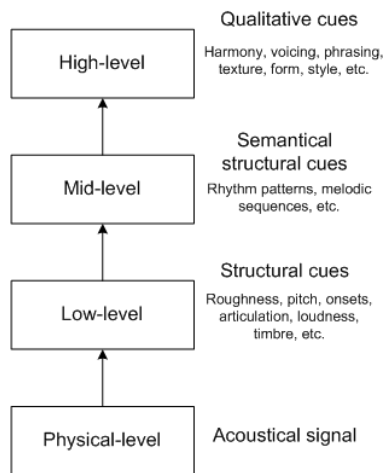


Figure 2.2: Illustration of 3 levels of musical descriptors.

The lowest level is the physical acoustical signal, which is represented as a waveform. The low-level descriptors consist of structural cues such as roughness, onset, pitch prominence, articulation and so on. In [Schubert, 1999a] he further suggests descriptors such as loudness, timbre and duration, where duration is also related to articulation. Mid-level descriptors as semantical structural cues such as rhythm patterns, melodic sequences and so on [Leman et al., 2005]. The high-level descriptors are of quality descriptive nature, such as emotional adjectives [Leman et al., 2005].

In [Schubert, 1999a] he suggests high level descriptors as harmony, voicing, phrasing, texture, form and style. Although definitions and levels of the descriptors are not always clear and the two levels low- and high level are extremes, where rhythm, contour, envelope, and articulation lie in between.

It is clear that definitions are many and connections between low- and high level features are not trivial.

### 2.1.3 Music and Lyrics

The descriptors and structure discussed so far has primarily been the acoustical aspect of music. An equal conveyor of information lies in the lyrical content of a song, in the case of non-instrumental music. Often musical producers create tracks by entwining lyrics, melody and different instrumentation to create an emotional expression. The role that the lyrics play on the overall expression of music is a subjective matter. Whether or not people listen to the words or the melody, or they perceive it is a whole. In [Bonnell et al., 2001] the relationship between lyrics and the music was investigated in French operatic music. They show that song and music are not perceived as one percept but rather that they are processed independently. Cognitively this implies that if melody and lyrics are present we are able to distinguish the two sources of information separately. Even though we can process them independently they are highly dependent on each other as investigated in [Nichols et al., 2009]. Here the relationship between the lyrics and melody in western popular music was investigated by statistical analysis. They investigate western tonal sheet music, using a database of melody, lyrics and chords which includes timing information, words boundaries, syllables and key information. They show that the level of syllabic stress is highly correlated with the strength of the metric position. The stopwords (at, or, the, of, etc.) are much less likely to coincide with melodic peaks than non-stopwords, and strongly correlated with weak metric positions. This shows that the composer of music synchronizes the structure of the lyrics to match the expression of the melody and instrumentation.

## 2.2 Emotion

To make a model of the emotional content in music it is essential to define what emotions are and further how to measure and relate them to each other. In this section the terms behind emotion will be defined following a presentation of different models that relate different adjectives that describe distinct emotions from each other.

### 2.2.1 Definition of Emotion

Emotions are often defined as distinct feeling states such as happy or sad, although different layers of feeling states exist, different affective terms are often used to describe a specific state a person is in e.g. emotion, mood, affect, feeling, arousal and appraisal. Affect is viewed as a more instinctive reaction to a certain stimuli, this process is manifested in individuals before any more complex emotional state can form and thus is underlying for many other term. It is viewed as an umbrella term that covers evaluative or valenced states such as emotion, mood and preference [Juslin and Västfäll, 2008]. The affective term mood is used as the underlying emotional state, which often has lower intensity and last longer for several hours to days [Juslin and Västfäll, 2008]. One can be in a good mood but still suddenly feel angry. The term emotion or an emotional

state is influenced by a number of emotion components such as the underlying mood, physiological arousal, expression etc. It is of relative high intensity and can last from minutes to hours [Juslin and Västfäll, 2008]. The term arousal is used to describe the physiological arousal or the excitement involved in an emotional state as valence is used to describe the positive or negative nature of the emotion. Lastly the term feelings is used as the subjective evaluation of an or all emotions, and thus they are often measured using self-report methods.

### 2.2.2 Models of emotions

Intuitively one could think of emotions as a result of some stimuli or situation, the emotion then inflicts different bodily responses. For example if one would be in danger of being killed, the person would feel fear and it would result in increased heart beat. The opposite could also be true that we experience an emotion such as fear due to the fact that we perceive our bodily functions in response to an event. One could also argue that emotions exist as to signal to the consciousness to reevaluate a situation. Equally the body can inhibit emotional response to different events such as being less startled when anticipating a slamming door. Ruling out any or the inclusion of all is not a trivial matter, instead it is interesting to look at how these emotions are described. Adjectives like sad and happy are often used to describe emotions and affect although one could find a great number of synonyms for happy such as cheerful, contented, content, glad, elated, euphoric, felicitous, joyful and joyous. How do these differ, how are they related and by how much? Models of emotions have been attempted for many years in order to describe the relationship between such adjectives. There exists two main directions in modeling emotions, one is the categorical model the other the dimensional model

**Categorical model** assumes a few innate emotions that carry different meaning such as happiness, sadness, fear, anger, disgust, etc. which are distinct and independent. All other emotions can be derived from these basic emotions. The amount or names of these basic emotions are often debated and no clear consensus is known. Assessing music and the emotional content and expression were done by [Hevner, 1936] who suggested that a long list of adjectives (67) could be divided into eight subgroups. This would enable listeners easily and accurately report his or her interpretation of the music. These include serious, melancholy, sentimental, quiet, humorous, merry, sensational, vigorous etc.

**Dimensional model** assumes that all emotions are interrelated and can be described in some emotional space spanned by a number of independent dimensions. The distinction between independence of dimensions and the issue of bipolarity is important when posing a research question. The question of the bipolarity of any dimension is not equivalent to the question of how many independent dimensions are required to describe affect [Russell and Carroll, 1999]. The issue of unipolarity or bipolarity will be discussed further in section 2.2.3. Within psychology the number and choice of independent dimension to describe

the emotional dimensional model is greatly discussed. Recently there seems to be consensus on a model: a two-dimensional space where each dimension shows independence [Russell and Barret, 1998]. In [Russell, 1980] he proposes a circumplex model of affect with two bipolar dimensions, valence and activation, where as seen on figure 2.3(a). Later a model was also proposed as seen on figure 2.3(b) where a list of adjectives are placed along the circumference. Many others assume this polar relation between the adjectives in the multidimensional models. Where exactly the semantic descriptors lie on the circle circumference is very different depending on measurements.

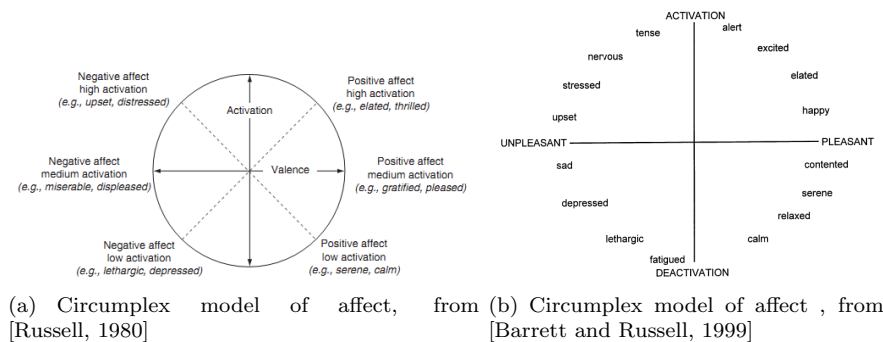


Figure 2.3

By circumplex is meant that each adjective is spaced equidistant to the center of a circle that is spanned by the two emotional dimensions, but does not imply that they are evenly spaced around the circle [Russell and Barret, 1999]. The model offers a structure for psychologists to represent the cognitive structure behind affect. It was supported by evidence of scaling 28 adjectives in 4 different ways. The two-dimensional space is further explored in [Russell and Barret, 1998] where the independence of the dimensions is tested. They show that the valence dimension was independent of activation, positive affect the bipolar opposite of negative affect, and deactivation the bipolar opposite of activation. Although one can argue there to be many other dimensions, the two dimensions are shown to account for most of the variance but not all, in affect rating [Russell and Barret, 1998]. Another approach to the definition of a model is made in [Russell and Barret, 1999] where more complex prototypical emotions are modeled as several core affects. These core affects are reflected in a two-dimensional bipolar model of pleasantness and activation. In [Barrett and Bliss-Moreau, 2009] a more exploratory approach is made to the structure of the circumplex model of affect. Although they do not read the evidence of which descriptors are best suited for anchoring the circumplex model, when looking at the brain structure, nonetheless descriptors can be scientifically useful. In [Posner et al., 2009] they measure the correlation between ratings of valence and arousal when partici-

pants are presented with emotion-denoting words and BOLD (Blood Oxygen Level Dependent) signals measured using *fMRI*. They show that there exist two underlying neural networks that subserve the affective dimensions of valence and arousal.

Although research indicates that the two dimensions of valence, meaning positive or negative felt quality in all emotions and arousal or activation meaning preparedness for action, are the main dimensions within modeling affect, more dimensions are used by some. In [Bradley and Lang, 1994] they use a three dimensional model to measure affect consisting of, affective valence (pleasant-unpleasant), arousal (calm-excited) and a dimension called dominance or control, where they show that the last dimension is not strongly-related.

Other three-dimensional models can be found by combining the circumplex model of Russell and Thayer's variant. The resulting model have the dimensions of valence, energy arousal and tension arousal, which is shown not to be able to be reduced to a two-dimensional model [Zentner and Eerola, 2010].

Both the categorical and the dimensional models, although seemingly are opposites, can coexist by postulating that the core affect and the underlying mechanisms are best described by a dimensional model. Where the conscious interpretation of these are categorical and influenced by the conceptual categories people have for emotions [Zentner and Eerola, 2010]. This Conceptual Act Model proposed by Barrett then consists of two layers: the Interpretive layer and the Affect layer, each described by the two models described.

### 2.2.3 Quantification of emotions

In order to model emotions it is essential to find a model that describes emotions and within this model find a method to quantify them. These two topics are highly interrelated where one quantification method or response format indirectly implies the use of a certain set of models. It is further suggested that in order to make a mathematical model of emotion in music it is necessary to convert some measure of affect or emotions into numerical values. Five different methods of measuring emotions are evaluated here.

**Psychological behavioral**, where experimenters and researchers observe test subjects during an exposure to certain events, that could be music playing. In this way the behavior or expressiveness can be interpreted into some emotional space.

**Psychophysiological** is another method that would require medical or biophysical equipment in order to observe changes. Often used methods are Electrocardiogram (ECG) for heart and pulse rate, different Biochemical responses such as saliva, blood and urine samples, skin conductance, respiration, blood pressure, muscular tension, temperature, chills, etc. In recent years methods

such as Electroencephalography (EEG) have also been used. Common issue is what and where something is measured, since emotions are not isolated and often there is a mixture of different emotions at one time, and also evolving over time. So is measured is not a trivial task.

**Functional Neuroimaging** *Functional Magnetic Resonance Imaging* (fMRI), *Positron Emission Tomography* (PET) and other *Event Related Potentials* (ERP) have been used. Common for the use of most physiological brain imaging methods is the issue of noise and timing. Depending of the strength of the magnet the machine itself creates overwhelming noise of up to above 130 dB SPL, which makes it difficult to listen to music. Different measures such as sparse temporal sampling or *Interleaved Silent Steady State : Exp* (ISSS) imaging can be used so that images are taken in interval leaving a quieter environment when listening to music.

**Self-report** method which in principle is a very broad group, which includes Likert rating scaling, adjective checklist, visual analogue scale, continuous report instruments, ranking and matching. Self-report methods can be implemented numerous ways, through PC interface, verbal report, through games, etc. A number of likert scales was discussed in [Russell, 1980] where the issue of ambiguous scales was discussed. By presenting a user with two 7-point unipolar scale from 1 to 7 or a bipolar scale from -7 to 7 makes a great difference in the way people rate, it gives an assumption of the way the given emotional dimension is modeled. So which scales and the underlying assumption of the emotional model is tightly connected.

In many self-report methods different types of scaling of emotions is required, either unipolar or bipolar. It is therefore important to know the implications of the use of either of them.

**Bipolar** is seen as two descriptors from a semantic vocabulary that are antonyms such as happy and sad or unhappy. These are positioned opposite on an axis, where a neutral term is centered between the two. These can either be measured using unipolar or bipolar scales, which further results in problems in the proof of bipolarity of affect. This is discussed in [Russell and Carroll, 1999] where they investigated the bipolarity of positive and negative affect, by using a unipolar format in order not to impose the assumption of bipolarity upon their data. They conclude that their model and data indeed suggest that bipolarity exist between positive and negative affect. The issue of response format will not be discussed further here, but will be discussed in greater detail in 4.3.2.



**Unipolar** measuring scales can be used to quantify multiple dimensional models, which in turn is a way to rate numerous adjectives on a scale from neutral to e.g. happy. The evaluation of unipolar multidimensional structures is entwined with the issue of bipolarity. Many factors can mask bipolarity, and as these artifacts are eliminated, unidimensional affect dimensions can be shown to be part of a bipolar space [Russell and Barret, 1999]. An argument for the unipolar format is the issue of people feeling both happy and unhappy at the same time. It could be that they are depressed but happy to hear some music. In [Russell and Carroll, 1999] and [Barrett and Bliss-Moreau, 2009] they explain this with the argument of time, suggesting that people switch between two or more different emotions such as happy and unhappy at some speed, thus indicating a unipolar connection when it is still in fact bipolar.

Which scales and response formats to use is highly interrelated with the amount of bias or which kind of bias one wishes to eliminate or choose to accept for the experiment. Likewise what assumptions are made about the emotional model as discussed previously.

**Bias in self-report**, some often encountered bias were discussed in [Zentner and Eerola, 2010] and include.

- **Demand characteristics**, where the participant figures out what is being tested and so conveying the hypothesis resulting in hypothesis-consistent behavior. Much like the issue of bipolar scales or unipolar scales, but could also be many others.
- **Self-presentation bias**, is when the participant is asked to do or rate something they feel is socially undesirable. Could be the liking of music that “no one else” likes and therefore would be hesitant to rate that they like it.
- **Limitation of the awareness of ones emotions**, can seriously limit the rating of emotions in music. If the participant is unclear of how the person actually feels regarding a song. The level of detail also plays in here, the greater the detail of a model and thereby factors to measure the more “fine tuned” the participant has to be of their emotions.
- **Communication** is a barrier if the person does not know what is being meant by the scale, anchors, labels, etc. If there is any doubt as to what is being asked the participant to do in the test, it would lead to inconsistent data.

The validity of using self-report methods in measuring emotions in music is discussed in [Zentner and Eerola, 2010]. They argue that this method is valid given the bias discussed, due to the fact that when in a listening situation it

is not likely that an overt expression or action takes place, and therefore the subjective experimental dimension will be the only one activated.

## 2.3 Emotions and Music

The question behind what makes us like a certain type of music or what elements of the music which make us like it, are intriguing and are attempted to be answered in the field of music psychology. This field tries to understand the underlying elements that e.g. evoke emotions, affect or different moods. In this section a more detailed look is presented on the effect music has on the emotional human.

### 2.3.1 Emotions in Music

Darwin's view on music was that it was a likely precursor of language, having its origin in the vocal expression of emotions [Cross, 2009]. It seems that emotions are evoked in humans when listening to music but what kind of emotions people perceive or are induced by when listening to music, and what elements or combinations of elements in music represent what emotions? Research within detection of emotions in speech as well as music has been done for a long time. Two aspects of music comes into play when talking about the emotional content and that is the two major sources of information, the acoustical signal itself, and the later decoded information from the signal in the form of the lyrical content.

#### **Acoustical source**

The work has been dominated by music psychologists that have researched the psychoacoustical elements in sound with psychological experiments as well as physiological measurements (for review see [Juslin and Laukka, 2003] and [Scherer, 2003]). In [McCraty et al., 1998] the effect of different types of music is measured using self-report methods, where an evaluation is made before and after listening to music. They show that listeners across age groups and preferences reduce stress and increase mental clarity when listening to designer music. Grunge rock is shown to increase hostility, fatigue, sadness and tension, where new age music is decreasing or have no effect on the same factor, but increasing relaxation. They further conclude that the reduction of stress and tension is higher for music that people enjoy and know. These emotions are rather broad in terms and could be constructed by more core affective terms, and some researchers argue that emotions induced or perceived in relation to music is much different from "normal emotions". In [Juslin and Västfäll, 2008] this is reviewed, one could argue that emotions that are likely to be experienced with music is somewhat naturally limited. The scenario of listening to music and feel the same kind of fear as when meeting a lion on the savanna is not likely. The amplitude of the emotion could also be different in the sense that when experiencing an emotion described as disgust when using another sense such as smell, would be much different and weighted different when us-

ing the hearing sense. On the other hand, emotions experienced when listening to music such as sadness could be much higher than other situations. It is safe to say that emotional experience is influenced by the perception of music, speech and sound in general. What emotions and to what degree this is the case is not certain. In [Russell and Barret, 1998] as was discussed before, argues that prototypical emotions consist of a combination of core affect. In [Scherer and Oshinsky, 1977], they suggest that combinations of acoustical parameters may serve to differentiate attributions of subclasses of certain emotions. One could argue that core affects are combined into emotions that are experienced with sound, music and speech in particular, where other combinations are then experienced when e.g. tasting or feeling.

### Lyrical source

Another aspect which is not covered in great detail here is the lyrical content in music. One aspect is the way the performer sings the lyrics. It can be in a sorrowful or angry way, which is more a acoustical aspect linking it to the performers expression [Palmer, 1997]. Another is the lyrics themselves. It is obvious that there lies a great difference between a singer singing “I love you”, or “I hate you”, whether it be ironically or not. Nonetheless information is carried. In [Herbert et al., 2008] and in [Kissler et al., 2007] they investigate the *Early Cortical Response* (ECR) and other *ERP* during reading. It is clear that reading semantically valenced words does induce emotions in humans, which can directly be measured as demonstrated in their work. To which extend a person is more emotional captivated by the lyrics or by the music itself is highly subjective. Often it is difficult to understand the lyrics in a musical track due to masking by the music and the singers pronunciation, where a lot of the information is lost. In [Ali and Peynirciogly, 2006] they show that lyrics do influence the overall emotional valence of music, allowing music to more easily convey negative emotions when they are present, and allowing music to more easily convey positive emotions when they are absent. They further show that the melodies rather than the lyrics are the most dominant component when eliciting all four of the emotions they use in their experiments. Consisting of (Happy, Sad, Calm and Angry) one for each of the quadrants from the circumplex emotional model of [Russell, 1980]. This view is also supported by [Sousou, 1997] where they test undergraduate students by means of self-report methods using two lyrical states (Happy and Sad Lyrics) and three musical conditions (No Music, Happy Music and Sad Music). They show that the mood of the participants was influenced more dominantly by the music and not the lyrics.

### 2.3.2 Perception vs. Induction of emotion

If one has to model the emotions in music it is essential to know how to measure these, and more important to know what is measured. At this point it is therefore relevant to distinguish between the perception of emotions and the induction of emotions. The perception of an emotion in music draws on the ability

to describe qualities of music using a vocabulary that is defined in some emotional adjective space. This is often measured using self-report methods, where the subject is asked to rate or use some defined adjectives or descriptors, certain characteristics or features in music that they recognize. The distinction between perception and induction is related to the distinction between cognitivism and emotism, where emotivists hold that music elicits real emotional responses in listeners, cognitivists argue that music simply expresses or represents emotions [Scherer and Zentner, 2001].

When an emotion is induced the psychological and mental state of a person is changed due to a stimuli. A number of factors can be measured and prove the induction of emotions as was described in [Juslin and Västfäll, 2008] including, subjective feeling, psychophysiology, brain activity, emotional expression, action tendency, emotion regulation. These are measured using the methods which were described in section 2.2.3.

These factors serve to prove that emotions are induced in people when listening to music, where different reactions occur due to different structural, circumstantial, musical and listener circumstances around a musical percept. e.g. the musical signal itself, the coloring of this signal through a “channel” that the music is perceived through and where you are when you hear the music. In [Scherer and Zentner, 2001] a set of production rules is suggested that account for the actual experience of an emotional response in music. They suggest a number of factors including,

- **Structural** features both supra- and segmental. Where the segmental are the acoustical building blocks of the musical structure and the suprasegmental include melody, tempo, rhythm, harmony etc. which will be discussed later in section 2.3.4.
- **Listener** features, which include their musical experience, the familiarity, current motivation or mood, the learned associations and conditioning and the cultural context, where the listener is from.
- **Contextual** features, such as where the music is being played and which acoustical scene the musical experience is set in. That could be a choir singing in a church or in a studio which would lead to much different experience, or listening to music via headphones or in a live concert. Further if the musical experience is a part of a greater event such as a festival or carnival.
- **Performance** features, where the identity of a performer including the physical appearance, expression and reputation of an artist influence the preference. The technical and interpretive skills and finally the performance state which includes interpretation, concentration, motivation, mood, stage presence, audience contact, etc.

In order to find specific reasons or origins of an induced emotion due to music, [Juslin and Västfäll, 2008] further proposes, a hypothesis, a theoretical frame-

work consisting of six mechanisms that are involved in the induction of emotions due to music and can explain emotions complementarily.

- **Brain stem reflexes**, due to musical stimuli is caused by the physical acoustical signal such as the stapedial reflex when impulse sounds are present that might be harmful, or loud music that can cause pain in the ear.
- **Evaluative conditioning**, refers to the trained association between an auditory stimuli and an emotional response, that can be caused by many factors, such as hearing some music every time something pleasant happens.
- **Emotional contagion**, refers to an internal expression of music induced by feedback from peripheral muscles or activation of an emotional center in the brain. In this sense a person 'mimics' what the person hears through music or vocal expression or sees through facial expression or behavioral movement.
- **Episodic memory**, contrary to emotional contagion the episodic memory reflects on the past memory of when a song or acoustical signal has been heard in the past.
- **Musical expectancy**, refers to the listeners expectancy of a specific acoustical feature in music occurring, which results in an emotion being induced in the listener. Since music depending on the style is built by components arranged in a certain structure, due to the context the acoustical features are presented in, one might expect some acoustical event to occur. This prediction is solely built on the structure where it is natural that this ability could be trained, and is more developed by musicians and people that have trained with an instrument [Palmer, 1997]. It could also be an implicit knowledge about the genre or common features with other genres [Stevens and Byron, 2009].

This theoretical framework gives the possibility to break down the influences music has on inducing emotions and potentially accounting for the possible effects, it might have when measurements are made. Atomizing the elements in music emotion induction is rarely done and by no means agreed amongst researchers. To give an overview of where the different terms have relevance, the figure 2.1 is extended on figure 2.4. An addition is made here in the form of perceptive skills, which here covers a rather broad group of the participants abilities to hear, e.g. are they normal hearing or does there exist any damage to the auditory system that would hinder the perception of the music.

One can argue intuitively that indeed music does induce emotions as a combination of the single elements presented in the framework. The connection between

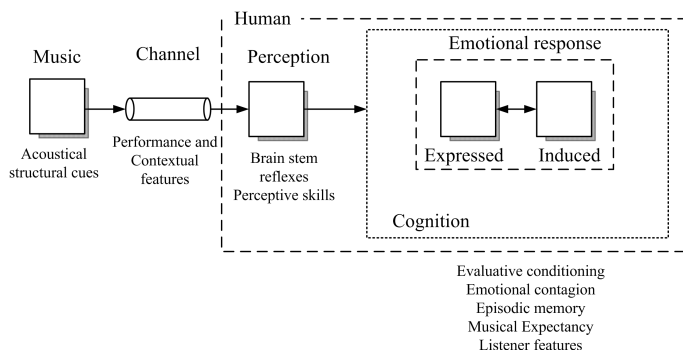


Figure 2.4: Extended schematic of the influences on the emotional response to music.

what an artist or musician expresses in the music and can be perceived, and the link to the induction of an emotion in a human is not simple and highly complex [Gabrielsson, 2002]. It is obvious that the induction of music is highly subjective and one person may have none or neutral induction of emotions due to specific musical piece where others may have strong emotional experiences as was shown in [Gabrielsson, 2001], so called peak experiences. Others may enjoy negative emotions expressed in music [Schubert, 1996]. This serves to show that a person might perceive an excerpt as being very negative and dark but actually liking the track, and the track inducing a positive feeling. People might prefer perceived negative music and naturally would be biased in their self-report about what the person feels about the track. It is obvious that many combinations could be made where people hate classical music or country music etc. although none of the acoustical features suggests so due to many other factors such as emotional contagion, episodic memory and evaluative conditioning. Nonetheless in [Evans and Schubert, 2006] is showed that by quantifying the relation between felt and expressed emotions in music, that there indeed there is a connection. Using self-report methods they show that, imagined and real music show the same valence and arousal shape for a concert pianist [Schubert et al., 2006], where music was chosen to be familiar for the test subjects. They show that in average in 70 % of the cases there is a positive connection between felt and expressed emotion, although further concludes that it is far from generalizable due to subjective differences. The results are also greatly influenced on the measuring technique and which rules are set up for the experiment.

### 2.3.3 Models of emotions in music

The distinction between general emotional models as described in section 2.2.2 and emotional models describing music, lies here in the semantic vocabulary used to describe the two. In section 2.3.1 it was argued that not all emo-

tions may be experienced when listening to music, hence limiting the choice of emotional models. Whether a dimensional or a categorical model is chosen, the anchoring of dimensions or labels used to describe the emotions elicited by music is crucial. The categorical method used in [Hevner, 1936] was indeed related specifically to music, but related to the dimensional models only few have been properly tested. Schubert suggests a two dimensional model that use valence (happy-sad), arousal (aroused-sleepy) space which he calls *2-Dimensional Emotions Space* (2DES) in [Schubert, 1999a]. In order to obtain a list of adjectives from all the emotional semantic words which exist and are suitable to describe music, he uses 91 words sampled from musical and non musical words. They are gathered from Hevner Adjective Circle (1936), Farnsworth's revision of Hevner (1960), Russel's circumplex model (1989), Whissell's dictionary of affect (1980) and Sloboda (1989) amongst others. He comes up with a reduced list of 37 words by frequency of use testing in musical description amongst 24 highly training musicians. The validity and reliability of this *2DES* were further tested and confirmed in [Schubert, 1999b]. Test subjects were asked to rate the words within the *2DES* given a developed interface, and the result shows that there is a good intuitive understanding of the *2DES*, and demonstrated good test-retest reliability. Further there was a high correlation between hypothesized responses from adjectives from Russell and Whissell.

In [Zentner and Eerola, 2010] a review of methods used to model emotions and they conclude that the best method for modeling emotions in music is the domain specific model using *Geneva Emotional Music Scales* (GEMS). Where Zentner et. al. uses questionnaires filled out by 801 participants at a festival primarily focused on classical music (72% classic, 11% rock, 10% world, 7% jazz). People are asked to fill out the questionnaires right after or during performance, whether they have felt any of the 66 emotional adjectives "somewhat" or "a lot". The most frequently reported include relaxed, happy, joyful, dreamy, stimulated, dancing (bouncy), enchanted and nostalgic etc. Using confirmatory factor analysis they reduce the list to a 9-dimensional model.

It is obvious that given so many synonyms of affect adjectives, many of them overlap where a solution could be to map these words from e.g. [Russell, 1980] or [Hevner, 1936] into a multi dimensional space. This could e.g. be done by using rated words from the *ANEW* database or as it was done in [Schubert, 1999a]. Problem lies in whether or not these adjectives are usable for the description of music, or as in the case of *ANEW* whether these adjectives have the same relation to music as they do to e.g. pictures, sounds or just the words themselves.

### 2.3.4 Acoustical cues and emotional response

Whether or not emotions are perceived or induced by music, it is the acoustical signal that is encoded by the musician and decoded by the listener that is the carrier of information. Disregarding other senses when listening to music such as the smell of a concert or the visual aspect of the e.g. back droppings, etc. The

goal here is to explore what acoustical features of the sound which are associated with what emotions. Like core affect could be responsible for prototypical emotions core acoustical features could be responsible for the communication of specific emotions. Different approaches have been taken in exploring acoustical features communicating emotions such as synthetic sounds, speech and music. It has to be mentioned that due to the fact that it is psychologists who are dominating this research field, the technical descriptions of the cues and methods to manipulate them is scarce. Musicians often use descriptors that was explained in section 2.1.2 and A.1 to account for what acoustically is the source of a change in emotion.

Rating scale	Acoustical parameter (main effects) and configurations (interaction effects) listed in order of predictive strength
Pleasantness	Fast tempo, few harmonics, large pitch variation, sharp envelope, low pitch level, pitch contour down, small amplitude variation (salient configuration: large pitch variation plus pitch contour up)
Activity	Fast tempo, high pitch level, many harmonics, large pitch variation, sharp envelope, small amplitude variation
Potency	Many harmonics, fast tempo, high pitch level, round envelope, pitch contour up (salient configurations: large amplitude variation plus high pitch level, high pitch level plus many harmonics)
Anger	Many harmonics, fast tempo, high pitch level, small pitch variation, pitch contours up (salient configuration: small pitch variation plus pitch contour up)
Boredom	Slow tempo, low pitch level, few harmonics, pitch contour down, round envelope, small pitch variation
Disgust	Many harmonics, small pitch variation, round envelope, slow tempo (salient configuration: small pitch variation plus pitch contour up)
Fear	Pitch contour up, fast sequence, many harmonics, high pitch level, round envelope, small pitch variation (salient configurations: small pitch variation plus pitch contour up, fast tempo plus many harmonics)
Happiness	Fast tempo, large pitch variation, sharp envelope, few harmonics, moderate amplitude variation (salient configurations: large pitch variation plus pitch contour up, fast tempo plus few harmonics)
Sadness	Slow tempo, low pitch level, few harmonics, round envelope, pitch contour down (salient configuration: low pitch level plus slow tempo)
Surprise	Fast tempo, high pitch level, pitch contour up, sharp envelope, many harmonics, large pitch variation (salient configuration: high pitch level plus fast tempo)

Table 2.1: Cross-Modal Patterns of Acoustic Cue for Discrete Emotions from [Scherer and Oshinsky, 1977]

In [Scherer and Oshinsky, 1977] they use 128 different synthetic sounds created by a *MOOG* synthesizer by variations of amplitude variations (small-large), pitch level (high-low), pitch contour (up-down), pitch variation (small-large), tempo (slow-fast), envelope (low attack/decay ratio-equal attack/decay ratio). Further they create 36 tone sequences by manipulating 4 tone sequences by: Two three level factors, filtration (lowpass-bandpass-highpass), tonality (major-minor), rhythm (even-uneven), tempo (fast-slow) totaling in 24 sequences. Test



subjects evaluated these stimuli for the emotional content using self-report scaling on a 3-point semantically differential scale. They can account for 66%-77% of the variance in the test data by manipulation of the acoustical parameters of the tone sequences. Where tempo seems to be the most powerful predictor accounting for a third of the variance. The main finding of the experiment is shown in table 2.1, where the acoustical features are organized by rating scale so that the acoustical feature that contributes most to the description of the variance for the given semantical differential dimension is listed first and sorted in descending order.

In [Gabrielsson and Juslin, 1996] nine professional musicians were instructed to perform short melodies to communicate specific emotions using different instruments - the violin, electrical guitar, flute and singing voice. The performances were grouped according to the physical characteristics e.g. tempo, dynamics, timing and spectrum. They showed that the performer's expressive intention had a marked effect on all analyzed variables. They further conclude that it is unlikely to find physical cues in the sound which are independent of instrument, musical style, performer or listener in communicating emotions. In [Laurier et al., 2009a] they use features which are commonly used in *MIR* and compare them with ratings given by 116 participants of 110 15-second excerpts from film soundtracks. They compare Dissonance or roughness, mode (major vs. minor), onset rate and loudness to the 5 mood categories of happy, sad, angry, fear and tender. They find a positive correlation between dissonance and anger and fear, whereas a negative correlation with sadness and tenderness. They also show that happy music is dominated by major mode, minor mode in sadness, fear mostly by minor, tenderness mainly in major mode whereas anger was ambiguous. Onset rate shows that happy music tends to be faster music with a high onset rate and for sad and tender they have lower values meaning slower music. Fear also has high onset values. With loudness they find that for anger and sadness it has a small variation of high loudness, relating it to arousal.

### 2.3.5 Temporal emotional dimension

As discussed previously emotions in general change as a result of events occurring, and in music being a dynamic media, change through time. First question is how fast is this? What timescale does the emotional response to music operate? In [Bigand et al., 2005a] they investigate what the duration of a musical excerpt affects on the emotional rating, with untrained and trained musicians. They use excerpts of different lengths (250 ms, 500 ms, 1 s, 2 s, 5 s and 20 s on average) on two groups of participants. They find that as little as 250 ms was enough to induce strong or weak "feelings" in listeners, whatever style was played. They conclude that less than 1 s of music is enough to instill elaborated emotional response in listeners suggesting that emotional responses are quasi-immediate as soon as music is played. Where emotions accumulate over the duration of a musical piece. This finding of music played less than 1 s in

order for an induction of emotion is also found in [Bigand et al., 2005b] where they also show that the musical experience of the participants has a very small influence on the results.

If listeners have an emotional response which is fast reacting, this could be measured. But the emotions that was investigated for very short intervals was not very complex emotions. Complex emotions could arise after a longer period of time, as the participant has time to use all the elements as discussed in 2.3.2. Music emotion measured as a continuous variable, meaning that emotions changing as a function of time due to musical stimuli, were investigated in [Schubert, 2010]. A great deal of research and tools have been developed in order to capture the emotional change (see [Schubert, 2010]). Common for most is that a dial or mouse cursor is moved within a dimensional space, often 2-D. Using more dimensions is tricky due to the cognitive load on test participants, since this task is very demanding, both understanding scales and continuously, “online” have to rate the music. This compared to measuring music in discrete intervals, e.g. 15 s or 30 s as seen often in the *MIR* community. In [Duke and Colprit, 2001] they investigate if the ratings given by participants when e.g. post-rating 15 s excerpts are the same as the mathematical mean of the continuous response of the same excerpt. This is not the case. In an extensive test they show that listeners post hoc or post performance overall perception is a result of a complex interaction between temporal, qualitative and dimensional variables. This means that the listeners memory of a musical excerpt is not merely a sum of all the emotional response during the excerpt, some single events are weighted highly where others are completely forgotten or distorted by previous or future events.

## 2.4 Problem specification

The objects presented in figure 2.1 was analysis within the scope of this project.

This section will present a summary and discussion of the analysis and a further specification of the problems presented in section 1.4. The major elements of modeling emotional content in music consist of the emotional model to use, the mathematical model to use, how to obtain the emotional data, how and which acoustical features to extract.

### **Emotional model**

The emotional model to use is chosen to be a two dimensional model of the dimension of Valence and Arousal. It has not with sufficient proof or consistency been shown in the work reviewed what other dimensions to add in order to model the emotional content better. The choice of a dimensional model lies in the the work which has been done in confirming its validity in representing appraisal with the human mind by e.g *fMRI* and other *ERP*. Descriptors of different core affect, musical specific emotions or combinations of affect should not be

further investigated, thus disregarding anchoring. The multidimensional model is kept as simple as possible, with the two dimensions of Valence and Arousal.

### **Mathematical model**

Within the reviewed work previously done in modeling of emotions in music one method does not stand out in particular, the problem as mentioned before is finding common comparative measure. The approach taken is to start with simple models and increasing complexity if found necessary. Due to the choice of bag of acoustical features, dimensional reduction methods have to be investigated and used together with the mathematical models, potentially incorporating them. Given that the emotional model is used, i.e. a two dimensional model, a regression model seems to be the logical choice.

### **Emotional data**

Due to the limitations of facilities and issue of scope for this work, self-report methods will be used. Issues of bias with the different measuring methods should be investigated and at all times minimized where possible. In figure 2.4 the chain from music to emotional response was illustrated. Since a very great number of influences exists for the induction of emotions, this cannot be modeled in one single model. Therefore a two-model strategy is viewed here, one for the expressed or perceived emotional expression by music and that of the induced emotions, the subjective impression. In this work the focus is on the first model. It is obvious that a completely clear distinction between the two is very difficult, using self-report methods on test participants. The theory here is that subjective variance does exist and this should be facilitated in the model. The large variations will be averaged out by asking a number of participants. The modeling of the temporal development of emotional content in music should be captured within the devised self-report listening experiment.

The choice between using post rating or continuous scales is partially made here. Common in the *MIR* community excerpts of 15-30 seconds has been used, only a few use a continuous rating method (sampling e.g. 1-4 Hz). To follow previous work post ratings is chosen and for reasons regarding the design of the listening experiment which will be discussed later in section 4.1. The main reason lies in previous work done by [Schmidt and Kim, 2010] where they use post ratings to predict the temporal evolution of emotions in time. The problem is that if temporal predictions of a potential mathematical model is made, these do not directly correspond to the actual emotions measured at that point, but are based on the post ratings. It was shown that a simple mean across an excerpt measured using continuous scales is not equivalent to the post rating. No work has been done within the mathematical modeling world to the knowledge of the author as to show the connection between a mathematical model of post ratings predicted on a shorter time scale and the connection to continuous ratings. So the aim is not to predict the actual emotional content, but to use the prediction to find a potential structure e.g. emotional dynamics, that can cluster data across genres and emotional ratings.

### Audio features

Although at present day lyrical features seem to be as semantically potent as acoustical information, the audio dimension is chosen here. It is the opinion of the author that previous work has not included the full scale and knowledge of acoustical information in previous attempts in the modeling of emotions expressed in music. Using acoustical data also ensures that a data foundation for the automatic prediction of the emotional evolution in music is always available. This is rarely the case for lyrical data. The approach to be used is a bottoms-up approach since there has not been sufficient work done in validating both algorithms to represent the musical features or work by musical psychologists in the technical description of their findings. So a wide variety of features should be extracted to cover as wide an area as possible.

### Musical data

A problem using self-designed experiments becomes what music should be rated. There is also an inherent limitation in the amount of participants to test. Approaches so far in *MIR* is a random sampling of datasets e.g. *USPOP2002*<sup>1</sup>, *CAL500*<sup>2</sup>, *AMC* or *MSD*<sup>3</sup> have been made, but this approach will result in redundant testing. Songs that are tested might be equal in emotional expression or audio data. Therefore an intelligent sampling of songs to be rated, which was used in [Schmidt and Kim, 2010] can be used. Acquiring 150.000 ratings is not a small matter as they did, so a new method is proposed as seen on figure 2.5.

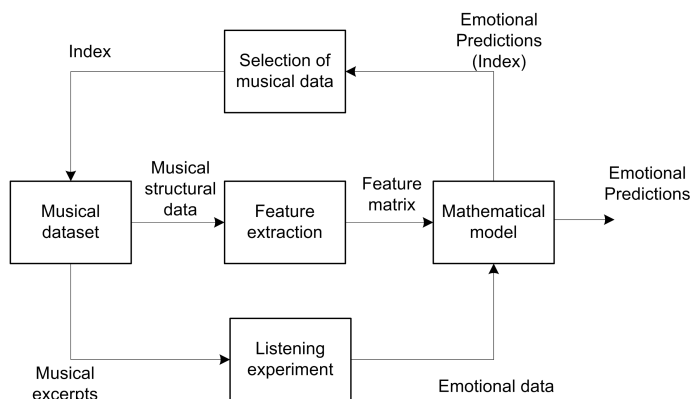


Figure 2.5: Schematic of the proposed sequential design, to intelligently sample musical data, for the purpose of modeling the emotional content expressed in music.

A sequential design, using an initial model constructed with emotional data based on audio features. This model can be used to sample large data sets of

<sup>1</sup><http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html>

<sup>2</sup><http://cosmal.ucsd.edu/cal/>

<sup>3</sup>Million song dataset - <http://labrosa.ee.columbia.edu/millionsong/>

music, which can subsequently be experimentally measured obtaining emotional data and used to update the model.

As musical data are hard to come by due to availability and copyright issues, a private archive is used for the preliminary testing. This is constructed using private archives, and webradio mining. For this reason the quality, encoding formats, etc. are varying which potentially can cause an error source when extracting acoustical cues. If the system should be used in the future for mobile players or computers where the data are of equivalent quality, this is acceptable.

In the rest of the thesis three different chapters will be presented that will explore the major aspects of modeling the expressed emotions in music. A section about audio features, the design of a listening experiment and the design of a mathematical model.



# Audio features

---

The audio features used for the modeling of the emotions expressed in music will be presented in this section.

## 3.1 Initial considerations

- The main goal is to be able to describe the descriptors that were mentioned in section 2.1.2 and A.1, which are often used in music psychology by psychologists to describe the influences of the acoustical signal on the expression of emotion in music. Further more the descriptors in table 2.1 in section 2.3.4 should also be accounted for by the features computed.
- A majority of the features found in the toolboxes listed on table 3.1 are correlated if not close to the same information. Feature selection algorithms or other types of information reducing measures can be used at a later stage to reduce these.
- Concerns such as computational time, complexity and redundancy should also be considered. Given that more algorithms computing the same feature where all are performing equally but one is faster to compute, this should be chosen for future use.
- Temporal alignment of features should be investigated, given that all feature extractors do not operate at the same output rate.
- The effect of lossy compression of audio data on the extraction of audio features should be investigated. Potentially some features could suffer greatly and are potentially unusable due to e.g. mp3-compression.

Given that the approach in acquiring and computing features is bottom-up, implementations were found in different academic DSP toolboxes and publicly available *Matlab*, *Python* and *C++* functions under the GNU license. These cover directly or indirectly the descriptors used by musicians and psychologists. Some of the feature extractors are considered “blackbox” since no information about the implementation methods were available. For this reason, only the information that was available in articles and technical documentation is presented here and within a level of detail that is appropriate for the problem given. The feature packs chosen are listed below.

- *YAAFE*<sup>1</sup> Yet Another Audio Feature Extraction toolbox, is a \*nix based feature extraction program which is shown to be very fast and efficient in computation. Features include Amplitude Modulation, Auto Correlation, Complex Domain Onset Detection, Energy, Envelope, etc.
- *MIR*<sup>2</sup> MIRtoolbox offers an integrated set of functions written in Matlab, dedicated to the extraction from audio files of musical features such as tonality, rhythm, structures, etc.
- *MA*<sup>3</sup>, Music Analysis (MA) toolbox is a collection of functions for Matlab. It contains functions to analyze music (audio) and compute similarities.
- *PsySound*<sup>4</sup> is software for the analysis of sound recordings using physical and psychoacoustical algorithms. It is an easy to use platform that does precise analysis using standard acoustical measurements, as well as implementations of psychoacoustical and musical models (such as loudness, sharpness, roughness, fluctuation strength, pitch, rhythm and running IACC).
- *ChromaToolbox*<sup>5</sup> Chroma Toolbox (CT) was developed by Meinard Müller, it contains MATLAB implementations for extracting various types of novel pitch-based and chroma-based audio features.
- *BCST*<sup>6</sup> Binaural Cue Selection Toolbox is a toolbox for calculation of Interaural differences based on psychoacoustical models.
- *ISP* Intelligent Sound Project toolbox, was developed as a cooperation between *IMM* at *DTU* and *AAU* amongst others. It consist of a collection of feature extraction functions for Matlab used within the project and inhouse at *IMM*.

---

<sup>1</sup><http://yaafe.sourceforge.net/>

<sup>2</sup><https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox>

<sup>3</sup><http://www.pampalk.at/ma/>

<sup>4</sup><http://psysound.wikidot.com/>

<sup>5</sup><http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/>

<sup>6</sup><http://www.acoustics.hut.fi/software/cueselection/>



In table 3.1 the packs and the resulting dimensions are shown. Meaning that not all features that can be computed by those pack are used, due to a multitude of reasons, e.g not working, producing invalid results, etc.

Feature pack	Total Dimension
ID	43
MIR	510
PSY	302
YAAFE	141
MA	64
CT	124
ISP	188
Total	1373

Table 3.1: The feature packs used to extract audio features from musical data. Specific features were selected and the resulting total dimensions of those features are shown in second column.

A complete list of all the features and their dimensions can be seen in section B.1.

## 3.2 Separation of audio data

A musical excerpt that is used for the testing and extraction of features is a dynamic media, being a non-stationary and non-linear temporal signal. The features used operate in different domains, e.g. time, spectral, modulation, etc. In each of these domains the signal must be separated in e.g. time frames, frequency bans or modulation bans, etc. Therefore two different types of separations is presented here that are common across all the features that are used.

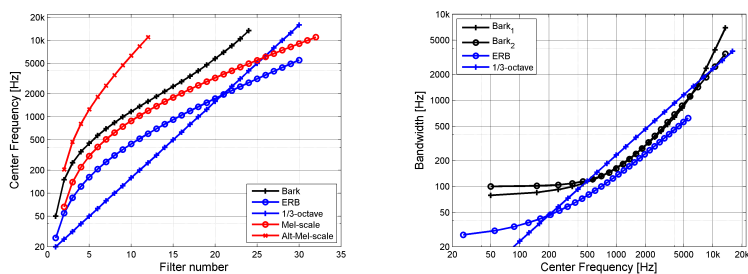
### 3.2.1 Temporal separation

In order produce multiple features across each excerpt temporal framing is used. Basic assumption on the time-series are made by the different features, some need stationarity to use Fourier analysis, other feature need a great deal of data to analyze the change of modulation spectra over time. Thus the temporal windowing is much different from feature to feature. A majority of features are based on the *Short Time Fourier Transform* (STFT) where the FFT is calculated for each short time window. In order for the *STFT* to be applicable the time series must be stationary. Depending on the audio signals content this lies between 20-100ms and thus the frame length lies in this window. Most features use a *Hanning* based windowing, with a certain number of samples overlapping between each frame, to create a smooth output. The overlap vary from 10% to 50% depending on features, including the different frame sizes misalignment can occur this issue will be dealt with in section 3.9.1, at present point it is noted

that a form of either alignment or integration of features into a feature vector is necessary for the modeling. For this reason temporal frame lengths are chosen to be the optimum for that particular algorithm where it is possible, equally for the amount of overlap of each frame.

### 3.2.2 Spectral separation

Across features, spectral separation is often used, using subband division by bandpass filtering. The width, shape and number of these filters is very different between implementations. Many of the filters are based on the non-linear properties of the inner-ear, specifically the basilar membrane. Different attempts have been made to account for these properties for normal hearing people. On figure 3.1(a) the center frequencies of different filter types are shown. The *Bark* scales relates to the Critical Bandwidth, another scale is the *Equivalent Rectangular Bands* (ERB), and the third octave band scale, which all are often used within the psychoacoustical community.



(a) Center frequency for different sub- (b) Bandwidth comparison for different  
band division scales. subband division scales.

Figure 3.1

Within the MIR community and speech recognition the so called *MEL*-scale is often used, which is a pitch based scale. The implementation of filters using this scale is very different. The two red curves on figure 3.1(a) show this fact, where the same filterbank script is used to compute filters from the *ISP* toolbox. It computes some user specified amount of filters to cover the frequency range also specified by the user. This changing the center frequencies drastically. Therefore between implementations based on the this scale, a direct frequency comparison cannot always be used.

On figure 3.1(b) the bandwidth of the filters again are shown on a logarithmic scale, where most a approximately linear on this scale. The *MEL*-scale is not shown here since the bandwidth of these filter vary a great deal, from different implementations.

Throughout this section a collection of features are presented divided into temporal, spectral, cepstral, perceptual, musical and misc. features. The exact division into these categories was made for convenience and often these categories overlap.

### 3.3 Spectral features

Features that use the spectrum of *Hanning*-windowed decomposed musical signal, computing features that describe each of the frames. Most of these features are *MPEG-7* features described in that standard.

#### Short Time Fourier Transform

All features are based in the *STFT* and is calculated as

$$X_{k,f} = \sum_{n=0}^{N-1} w_n x_n \exp(-2j2\pi kn/N) \quad (3.1)$$

for  $k=0,1,\dots,N-1$  where  $k$  corresponds to the frequency  $f_k = kf_s/N$  where  $f_s$  is the sampling frequency and  $w_n$  is the window function. The amplitude of the spectra  $X_{k,f}$  will be denoted  $a_{k,f}$  and the magnitude will be denoted  $M_{k,f}$ .

#### Spectral Rolloff

A way of estimating the amount of high frequency content is to find the frequency at which a certain fraction of the total power in a given frame is contained. In [Tzanetakis and Cook, 2002] they suggest a fraction of 0.85.

$$\sum_{k=1}^{R_f} M_{k,f} = 0.85 \sum_{k=1}^N M_{k,f} \quad (3.2)$$

where  $M_{k,f}$  is the magnitude of the  $k^{th}$  frequency component of the Fourier transform of  $x_n$  in the  $f^{th}$  frame.

In *YAAFE* they suggest a fraction of 0.99.

The *MIR* implements a supplementary method that measures the *brightness*, where the cutoff frequency fixed to 1500 Hz and then measuring the fraction of energy in the signal above this frequency.

#### Spectral Flux

The spectral flux gives an indication of the change in spectrum as a function of time, and is defined as the squared difference between the spectrum of two successive frames.

$$F_f = \sum_{k=1}^N (M_{k,f} - M_{k,f-1})^2 \quad (3.3)$$

where  $M_{k,f}$  is the magnitude of the Fourier transform of frame  $f$ , summing over all frequency components  $k$ .

In the *YAAFE* toolbox it is calculated as

$$S_{flux} = \frac{\sum_{k=0}^{N-1} (a_{k,f} - a_{k,f-1})^2}{\sqrt{\sum_{k=0}^{N-1} a_{f-1,k}^2} \sqrt{\sum_{k=0}^{N-1} a_{k,f}^2}} \quad (3.4)$$

where  $a_{k,f-1}$  is the amplitude of the  $k^{th}$  frequency component of the Fourier transform of  $x_n$  in the  $f^{th}$  frame.

$f_k$  is the frequency of the bin  $k$ .

$N$  is half of the *FFT* window size of frame  $f$ .

In the statistical world often different descriptors are used to describe a Gaussian distribution. These include the mean, variance, skewness and kurtosis. Which are the 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> moment of the mean. Assuming the spectra is a Gaussian distributed frequency distribution. Instead of using the whole spectra as a feature vector these descriptors can be used to reduce dimensionality and describe the shape of the distribution.

### Spectral Centroid

Is the first moment of the mean and is referred to as the centroid and gives an indication of the brightness of a musical piece and is defined as the center of gravity for the STFT.

$$C_f = \frac{\sum_{k=0}^{N-1} M_{k,f} k}{\sum_{n=0}^{N-1} M_{k,f}} \quad (3.5)$$

where  $M_{k,f}$  is the magnitude of the  $k^{th}$  frequency component of the Fourier transform of  $x_n$  in the  $f^{th}$  frame.

Notationwise here is another way of writing the same equation for later use. Here the  $f$  is omitted to make the notation easier.

$$\mu_i = \frac{\sum_{k=0}^{N-1} f_k^i a_k}{\sum_{k=0}^{N-1} a_k} \quad (3.6)$$

where  $a_k$  is the amplitude of the  $k^{th}$  frequency component of the Fourier transform of  $x_n$ .

$f_k$  is the frequency of the bin  $k$ .

$N$  is half of the *FFT* window size of frame  $f$ .

The  $i$  here indicates the moment of the mean.

### Spectral Spread

This spectral shape feature is often referred to as Spectral Spread or Spectral

Width. This second moment of the mean is a measure of the variance of the mean/centroid calculated.

$$S_w = \sqrt{\mu_2 - \mu_1^2} \quad (3.7)$$

where  $\mu_i$  is calculated as in (3.6) for the second moment the whole equation is written out as

$$S_w = \frac{\sum_{k=0}^{N-1} (f_k - \mu_1)^2 a_k}{\sum_{k=0}^{N-1} a_k} \quad (3.8)$$

### Spectral Skewness

This measure is also called the spectral asymmetry and is the third moment of the mean and is calculates as

$$S_s = \frac{2\mu_1^3 - 3\mu_1\mu_2 + \mu_3}{S_w^3} \quad (3.9)$$

where  $\mu_i$  is calculated as (3.6)

### Spectral Kurtosis

This features is often also referred to spectral flatness and is the fourth moment of the mean and is calculated as

$$S_f = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{S_w^4} - 3 \quad (3.10)$$

where  $\mu_i$  is calculated as (3.6)

### Spectral slope

It represents the amount of decrease of the spectral amplitude. It is computed my linear regression where the slope is calculated as

$$S_{slope} = \frac{N \sum_{k=0}^{N-1} f_k a_k - \sum_{k=0}^{N-1} f_k \sum_{k=0}^{N-1} a_k}{N \sum_{k=0}^{N-1} f_k^2 - (\sum_{k=0}^{N-1} f_k)^2} \quad (3.11)$$

### Spectral Variation

Spectral Variation is the normalized cross-correlation between two consecutive frames' amplitude spectra and is calculated as

$$S_{var} = 1 - \frac{\sum_{k=0}^{N-1} a_{k,f-1} a_{k,f}}{\sqrt{\sum_{k=0}^{N-1} a_{k,f-1}^2} \sqrt{\sum_{k=0}^{N-1} a_{f,k}^2}} \quad (3.12)$$

This method of computing the difference between two consecutive spectra is also the approach pursued in Spectral Flux. When it is close to 0 if the two spectra are similar and close to 1, when the two spectra are highly dissimilar.

### Spectral Decrease

The spectral decrease feature is similar to the spectral rolloff and is a measure of much the spectral amplitude decreases. The method comes from perceptual studies and should give a more correlated result to human perception.

$$S_{dec} = \frac{1}{\sum_{k=2}^N a_k} \sum_{k=2}^N \frac{a_k - a_1}{k - 1} \quad (3.13)$$

### Spectral Flatness

Some also calculate the spectral flatness in a different way, here as the ratio between the geometric and the arithmetic mean of the spectra.

$$S_{flatness} = \frac{\exp(\frac{1}{N} \sum_{k=0}^{N-1} \log(a_k))}{\frac{1}{N} \sum_{k=0}^{N-1} a_k} \quad (3.14)$$

## 3.4 Temporal features

In this section a description is given of some of the features that use the temporal waveform of the acoustical signal.

### Root mean square energy

A very simple way of calculating the energy

$$x_{rms,f} = \sqrt{\frac{1}{N} \sum_{n=1}^N x_{n,f}^2} \quad (3.15)$$

where  $N$  is the number of samples in the  $f_{th}$  frame. Calculating the RMS power in each frame gives a temporal evolution of the power, over time resulting in an energy calculation. No frequency division is made here.

### Low energy percentage

This feature uses a set of smaller frames to form a window to calculate what percentage of frames in one window have a power that is below the mean *RMS* value of that given window. It can be calculated as

$$L_e = \frac{1}{G} \sum_{f=1}^G H(\bar{x}_{rms,W} - x_{rms,f}) \quad (3.16)$$

where  $x_{rms,f}$  is the *RMS*-power of frame  $f$ , calculated by (3.15).  $\bar{x}_{rms,W}$  is the average *RMS*-power of the whole window.

$G$  is the number of frames in window  $W$ .

$H$  is the Heaviside step function defined by  $H(x) = 1$  for  $x > 0$  else 0.

As suggested in [Tzanetakis and Cook, 2002] vocal music with silences between each utterance will have a large low-energy value, while continuous instrumental playing will have a small low-energy value

### Zero crossings

The amount of time domain zero crossings of a waveform is said to give an idea of the noisiness of a signal. The higher rate the more noisy the music track is.

$$Z_f = \frac{1}{2} \sum_{n=1}^N |\text{sign}(x_n) - \text{sign}(x_{n-1})| \quad (3.17)$$

where the *sign* function is 1 for positive values and 0 for negative and  $x_n$  is the temporal waveform. This also indicates something about the articulation, i.e. the temporal distance between notes.

### Temporal Shape descriptors

Similarly to 3.4 the shape of the waveform is here described by the centroid, variance, skewness and kurtosis. For the temporal waveform similar to (3.6),(3.8), (3.9) and (3.6) the time samples are used instead of the frequency bins. This only makes sense if the histogram of the given time frame is Gaussian shaped. Often this is not the case for musical signals, nonetheless the feature might hold some information.

### Envelope Shape descriptors

The amplitude of the temporal envelope is extracted using a Hilbert transform, then low-pass filtered followed by a decimation. The methods as used in (3.6), (3.8), (3.9) and (3.6) are used on the envelope curve, where instead of spectral bins, it is samples of the amplitude of the envelope curve. As mentioned previously this only makes sense if each frame is Gaussian, which is true in many cases.

### Onsets

Another more simple method of detecting the tempo or rhythm in a musical track is to detect the peaks or onsets on the temporal envelope of the audio signal. The approach is to choose maxima or peaks on the temporal envelope curve in each frame, where the envelope is calculated using a simple low-pass filtering approach.

### Complex Domain Onset Detection

In [Duxbury et al., 2003] they develop a method of onset detection in musical signals. They combine phase and energy information instead of the more trivial energy-based approaches, in their onset detection function. The combination gives a more robust setup and provides sharp peaks at onset and smooth everywhere else.

## 3.5 Cepstral features

### Cepstrum

Is the Fourier transform taken of the logarithm taken of the spectrum. So a spectrum of a spectrum where the phase information can be preserved or removed by the log operation.

$$FT\{\log(FT(x(n)))\} \quad (3.18)$$

The Cepstrum results in giving information about the change in the spectrum and is often used in speech and music analysis. Often it is predecomposed in to frequency bins on the Mel-scale, which will be discussed later.

### Cepstral flux

Is simply the change in the Cepstrum as a function of time, where the resolution is dependent of the frame size the waveform is decomposed to. It is calculated similarly to (3.9) where the Cepstrum is used instead of the Magnitude spectrum.

### Cepstral centroid

Is similar to the spectral centroid where here it is the quefrency centroid calculated similarly to (3.5).

### Mel Frequency Cepstral Coefficients

Is the Cepstrum divided into Mel-frequency bands. The MFCCs have often been used in speech recognition and increasingly in music research. Here seven different implementation are used, four from *ISP* and single implementations from *YAAFE*, *MA* and *MIR*. Illustrative the implementation from *ISP* is shown. The first stage is to make a short time decomposition of the temporal waveform  $x_n$  and compute the discrete Fourier transform of each window. This is done as was shown in (3.1). Different types of windowing functions can be used, where the rectangle and hamming windows are often seen.

The magnitude spectrum is now scaled logarithmically and using the Mel filter bank ( $H_{k,m}$ ) it is additionally divided into frequency bins. The Mel filter bank used is different for different implementation with a “logarithmic” frequency scaling tendency, but in general it is used to mimic the frequency resolution of the inner ear and the basilar membrane particular. This gives

$$X'_m = \ln\left(\sum_{k=0}^{N-1} |X_k \cdot H_{k,m}|\right) \quad (3.19)$$

for  $m = 1, 2, \dots, M$ , where  $M$  is the number of filterbanks where the number of filterbanks is restricted to much less than the length of the waveform. The shape, type and center frequencies for the filterbank is implementation specific but in general an approximation to the center frequencies of the Mel scale is



given by

$$\phi = 2595 \log_{10} \left( \frac{f}{700} + 1 \right) \quad (3.20)$$

further the amount of MFCC filters  $M$  to use, and what the maximum frequency is different. Given a  $f_c = 22.050 \text{ Hz}$  the  $f_{max} = 11.025 \text{ Hz}$  dependent on the width of the filters  $M = 11 - 40$ . Finally to obtain the MFCCs a Discrete Cosine Transformation (DCT) is made on  $X'_m$

$$c_l = \sum_{m=1}^M X'_m \cos \left( l \frac{\pi}{M} \left( M - \frac{1}{2} \right) \right) \quad (3.21)$$

for  $l = 1, 2, \dots, M$ , where  $c_l$  is the  $l^{\text{th}}$  MFCC.

## 3.6 Perceptual features

### Pitch

Pitch may be defined as that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale. In other words variation in pitch gives rise to melody and is related to the repetition rate of a waveform. For a pure sinusoid this is referred to as the frequency and for complex tones it is called the fundamental frequency [Moore, 2004]. It is important to note that it is not just the spectrogram that describes the pitch, but rather the human perception of it.

Four different methods are used to extract the pitch of each musical track. An autocorrelation method and a spectral decomposition method that both return the frequency of the presence of a pitch component from *MIR*. The last is method from [Müller, 2007] that uses a multirate filterbank decomposition using elliptic filters with center frequencies corresponding to pitches A0 to C8 corresponding to MIDI pitches  $p = 21-108$ . They then compute the short-time mean-square power (STMP) for each band to indicate the amount of energy that is present at a given musical note.

### Roughness

Using the knowledge of the auditory system, that whenever two sinusoids are close in frequency and time, the auditory system cannot perceive them as two tones but rather a beating sound is perceived with the frequency of the spectral distance between them. This sensory dissonance gives a sensation of roughness in sound. The method here is to sum up all pairs of sinusoids with selective spectral peak picking.

### Fundamental frequency

In *ISP* a method of calculating the fundamental frequency is used, that is a maximum likelihood and order estimator. It calculates likelihoods using an FFT and is based on an assumption of white Gaussian noise based on MDL.

### Harmonics

In *ISP* based on the calculation of the fundamental frequency the number of harmonics, or overtones is calculated equally based on MDL. Thus the features relates to the musical descriptor of harmony.

### Inharmonicity

Is here defined as the amount of frequency components or partials that are not multiples of the fundamental frequency. So the divergence of the signal spectral components from a purely harmonic signal. In [Peeters, 2004] it is calculated as

$$I_h = \frac{2}{f_0} \frac{\sum_{k=0}^{N-1} |f_k - f_0 k| a_{k,f}^2}{\sum_{k=0}^{N-1} a_{k,f}^2} \quad (3.22)$$

where  $a_{k,f}$  is the amplitude of the  $k^{th}$  frequency component of the Fourier transform of  $x_n$ .  $f_k$  is the frequency of the bin  $k$ . It ranges from 0-1 since  $a_{k,f} - f_0 k$  is at maximum at  $f_0$ .

### Total/Main and specific Loudness

A general description of loudness was given in section 2.1.2, for all different feature packs most contain measures of loudness. The implementation from [Peeters, 2004] computes an approximation of the relative loudness in a bark band scale using a simplified approach than was suggested in [Moore et al., 1997], where some conditions for quiet signals are removed. In *PSY* two major different methods are used. A implementation of the method proposed in [Chalupper and Fastl, 2002] was made that computes the dynamic loudness model. These models use the Bark critical band rate scale to model auditory filters, and auditory temporal integration is included in the loudness model. A loudness fluctuation model is also included. The static loudness model of [Moore. et al., 1997] is also calculated, although it is a static model, it is applied to each analysis window as if it were a dynamic model.

### Perceptual Sharpness

The measure of sharpness is a perceptual equivalent to the spectral centroid where, in this implementation, it is computed using a Bark band scale. It is calculated according to [Peeters, 2004] as

$$S_{sharp} = 0.11 \frac{\sum_{b=1}^{nbands} g_b N'_b}{N} \quad (3.23)$$

where

$$g_b = \begin{cases} 1 & \text{for } b < 15 \\ 0.66 \exp(0.17b) & \text{for } b \geq 15 \end{cases} \quad (3.24)$$

### Perceptual spread

This feature is a measure of the distance from the largest specific loudness value

to the total loudness. It is calculated as

$$P_{spread} = \left( \frac{N - \max_b N'(b)}{N} \right)^2 \quad (3.25)$$

### Interaural differences

For humans to perceive spaciousness or to locate sound sources, temporal and level differences between our two ears are used, along with monaural cues. Using perceptual models, estimates of these differences can be computed. The auditory peripheral system, perceives temporal differences (*Interaural Time Difference* (ITD)), when low frequency sounds reach first one ear and then the next, creating a time difference. This occurs with little to no ambiguity at frequencies at 725 Hz and below. Sound pressure difference between two ears called the *Interaural Level Difference* (ILD) occurs with sounds containing high frequency contents, where the head acts as an acoustical shadow, thus dampening the sound. This only occurs at frequencies of 500 Hz and above. Due to these limitations only the bark bands that lie within these limits are used in the extraction of these features. The *Interaural Coherence* (IC) has also showed to be useful for audio source location, therefore this features is also used here.

## 3.7 Musical Features

### Chromagram

The chromatic scale has been used in music for many years, and consist of a 12 evenly spaced pitches, one semitone apart, ranging from C to B. This adding up all tones thus reduces the dimension of the pitch scale. The theory for the use in Music Information Retrieval is that perceptually there is no difference between a C in each octave. This is the case for single frequency components or complex tones, but the greater amount of tones played at the same time, the easier it is perceptually to distinguish tones across octaves [Moore, 2004]. The implementation approach taken in [Müller, 2007] is to simply appropriately add each *STMP* up so that A0, A1, A2, etc are added up.

**Hybrid-Chroma** The *Chroma Energy Normalized Statistics* (CENS) that were used in [Müller et al., 2005] provide short-time statistics across energy distributions within each of the chroma bands. The feature shows a high correlation with the short-time harmonic content of the temporal acoustical signal. It has a high level of robustness to variations of properties such as dynamics, timbre, articulation, execution of note groups, and temporal micro-deviations.

Another hybrid feature that is based on the Chromagram is the *Chroma DCT-Reduced log Pitch* (CRP) which was introduced in [Müller et al., 2009]. The idea comes from *MFCC* where it is said that the lower coefficients are very related to the concept of timbre. Thus if one wants to create a feature that is robust to just that, one removes this information, and that is the approach here.

Combining the idea of *Chroma* and *MFCC* one can then, instead of using the mel-scale one replaces it with a nonlinear pitch scale. Then a *discrete cosine transform* (DCT) is applied on the logarithm of the pitch representation to obtain *pitch-frequency cepstral coefficients* (PFCC). Removing the timber dependent data by only keeping the upper coefficients, then an inverse *DCT* is applied. The resulting pitch vectors are then projected onto the 12-dimensional chroma vectors.

### Fluctuations

In [Pampalk et al., 2002] a method of calculating a time-invariant representation of the rhythmic-pattern is proposed, which contains information about how strong and fast beats are played. It uses the amplitude of the modulation coefficients weighted using a psychoacoustical function, that models the fluctuation strength. The implementation in *MIR* is based on this approach and consists of a spectrogram computation transformed by an auditory model. The audio signal is decomposed to 23ms frames half overlapping, the *Terhardt* outer ear model is applied and spectrally divided into bark-band scale, where magnitude values are converted to dB values. Each bin for each frame a FFT is computed, from 0 to 10 Hz, corresponding to a rhythm pattern of up to 600 bpm. The modulation coefficients amplitudes are then weighted using a psychoacoustic model of the fluctuation strength. This measure is related to the computation of the rhythm in music.

### Tempo

The tempo is estimated using the onset curve, where the autocorrelation is computed and a peak picking algorithm is applied. In order for the tempo detection to be valid a frame size of over 1 sec is used, with a hop size of 50 %. For each frame a single value is returned as beats-per-minut(BPM). This measure is also related to the rhythm of a song.

### Pulse clarity

Pulse clarity is considered a high-level musical dimension that conveys how easily in a given musical piece, or a particular moment during that piece, listeners can perceive the underlying rhythmic or metrical pulsation [Lartillot et al., 2010]. This given method uses the maximum autocorrelation value for each frame of the audio signal giving a indication of the strength of the beat. A multitude of other values to pick out on the autocorrelation curve of the onset curve and can be compared at a later stage.

### Keystrength

Is computed as the probability of each key candidate, through a cross-correlation of the wrapped and normalized Chromagram and a set of similar profiles for each key. Where the resulting values, the cross correlations of each key result in a measure of the strength of each key.

**Key**

This feature is closely related to the keystrength, where the estimation of the tonal centerpoints is calculated using a peak picking algorithm on the keystrength cross-correlation curve. Here also the clarity is calculated using the ordinates of the curve. Thus the features relates to the musical descriptor of Register.

**Tonal centroid**

Also called tonality, where in [Harte et al., 2006] they suggest a new model for Equal Tempered Pitch Class Space, where a 6-dim tonal centroid vector is calculated. A 12-bin chroma vector is mapped into the so called interior space of a 6-D polytope, where pitch classes are mapped vertices of this polytope. Which should correspond to a projection of the chords along circles of fifths, of minor thirds, and of major thirds.

**Harmonic Change**

Is again proposed in [Harte et al., 2006] and is the flux or change of the the tonal centroid vector between each frame. Giving a way of calculating a rhythm related measure

## 3.8 Misc. features

**Linear Predictor Coefficients**

Linear Predictor Coding is normally used in the mobile communication world to code and compress a speech signal, which gives the best quality speech at low bitrates. Here the coefficients are computed for each frame. It uses an autocorrelation method in combination with a Levinson-Durbin recursion algorithm. Since it is used for speech signals the results on musical signals and singing speech voice is unknown.

**Linear Spectrum Frequency**

The approach in LSF is similarly to LPC which is a predictive method of coding a speech signal. The implementation is adapted from [Duxbury et al., 2003].

## 3.9 Post-processing

Given the great number of features some post-processing should be carried out in order to enable the use of them in the mathematical modeling. Further some meta analysis should be made in order to gain perspective prior to modeling. Three main areas will be investigated here, the temporal alignment of features, the issue of the potential degradation of information due to lossy encoding of audio source data, and last a look at corrections of the output of the feature extraction toolboxes.

### 3.9.1 Alignment of features

Given that most of the features operate in a frame based computational manner, the window size was chosen to be the default for all algorithms. For the sake of mathematical modeling these features have to be aligned in some manner. An example of the problem is illustrated on figure 3.2, where some features have e.g. frame sizes of several seconds where others exist in as little as 9 ms. One idea is to integrate short-time features as was suggested in [Meng et al., 2005] thus changing the features into alternative features that would align. The sheer number of features used in this project and further investigation of this is non-trivial. Another method is simply to resample all features to a common temporal frame length. Issues such as interpolation noise, when upsampling, and the removal of high frequency components, when downsampling, should be investigated further. When resampling the features, even though a degradation of information occurs, this degradation is not direct transferable to a potential performance of a mathematical model. But crucial information could potentially be removed due to this choice.

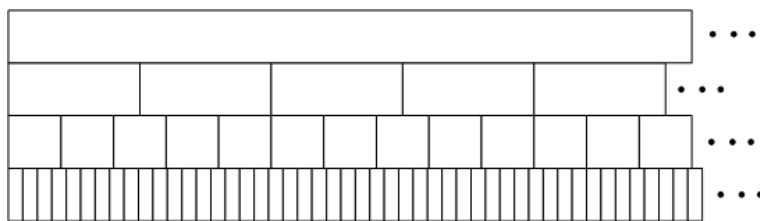


Figure 3.2: Illustration of audio features miss alignment in the temporal dimension.

In section B.3 a more thorough investigation is made. A sampling frequency was chosen to be a compromise between available computational power, for later mathematical modeling, and audio feature degradation. Ideally the smallest temporal frame, should be chosen. The compromise was a factor 8 of it, resulting in a sampling of 110 Hz, producing 1654 frames for 15 second excerpts. All features were then resampled, where all features that were downsampled was investigated in greater detail. 10 excerpts were used, and the  $r^2$  correlation coefficient was averaged across all excerpts. The result was that features that were computed across multiple audio frequency bands, were the features that suffered most. *IC*, *Chromagram* and *Loudness* measures suffered up to as much as  $r^2 = 0.4$ . The resulting error due to this degradation, in the mathematical modeling is not directly transferable, thus exclusion of features due to this measure cannot be made. Caution has to be made, if variance and co-variance analysis is made on resampled features, since variance is removed from these features.

### 3.9.2 Effect of Lossy-compression

Due to the varying source of quality present in the *pilot2* dataset and potentially future data, an investigation of the effect of different encoding types on the features extracted was made in section B.2. Two different aspects of the encoding of audio can reduce the quality of the audio signal. First a lowpass filtering is made of the audio signal, after that a sub-band division is made, within each of these bands a psychoacoustical model is applied, on this basis the quantization bits are chosen of the particular sub-band. If a sound is inaudible due to masking effects of the auditory system, the quantization bits allocated, will be reduced. Thus producing a lossy compression.

The conclusion of the investigation was that, common across most features the 128 kbit encoding is the worst of the four compared (128 kbit, 192 kbit, 320 kbit and Variable Bit Rate avg. 192 kbit), where the 192 kbit and VBR follow and as expected the best is the 320 kbit. For features showing a low correlation due to encoding, the general rule is that 192 kbit is the best of the rest, where 320 kbit is the clear winner over all features. Based on these results, 192 kbit and above is acceptable for the musical data. If features that have a low robustness to encoding get selected by a feature selection algorithm, caution has to be taken.

### 3.9.3 Output corrections

For reasons unknown to the author a number of the features that are used produce or output so called *NaN* values, which in *Matlab* is *Not a Number*. This in terms indicates that the algorithm produce missing data points. For these values, so called *NaN*-painting is used, which is often used within image processing. An investigation of this made in section B.4, where the feature of MFCC Flux was investigated. This feature was the one that produced the most missing values across all features and was therefore used for investigation. To test the performance of so called *NaN*-painting methods and to see the effect of potentially using this feature a test setup was made. MFCC Flux features calculated from 10 excerpts with no missing values was used. These were corrupted with an increasing amount of missing values using a distribution of errors similar to the existing errors of other missing values in feature vectors. The result showed that of up to 50 missing values was the maximum that will be allowed, resulting in a  $r^2$  statistics of 0.9 using a method that uses an average of the 8 surrounding values to calculate the missing value.

## 3.10 Conclusion

The objective was the obtain structural information about music, which was achieved. This was done using audio features extracted from musical data. Using a bottom up approach from an acoustical point of view, to gather features that describe both aspects as musical psychologists, musicians and features used and investigated in *MIR* and *DSP* community. 7 different feature extraction packs were found and working and relevant features were chosen from each.

Following were chosen and the dimensions they contain *ID* 33, *MIR* 540, *PSY* 280, *YAAFE* 143, *MA* 64, *CT* 124, *ISP* 189, totaling in 1373 dimensional feature vector. These features are divided into the domains of spectral, temporal, cepstral, musical, perceptual and a miscellaneous. Features where information and relevant information were available was presented.

The musical data used for the foundation of making a mathematical model for the prediction of emotions expressed in music was based on lossy compressed audio data. An investigation was made into the effect the compression had on the extraction of audio features. The result was that data that had a compression at no lower than 192 kbit could be used without any severe degradation of data. On the other hand no direct connection between loss of information in features and the resulting predictions of mathematical models could be made.

To align features a resampling method using a 100th order polyphase resampling method of *Matlab*. The effect of this on features was investigated when features were downsampled. For features which are frequency decomposed this showed to have a profound effect on the correlation of the features between original and downsampled.



# Listening experiment

---

In order to obtain emotional data for music, a listening experiment is devised.

## 4.1 Experimental considerations

Some issues should be considered before starting the design of a listening experiment.

Issues regarding the choice of musical data to test, and the length of these has to be taken into account. Following points are considered

- Enough music should be rated so that a mathematical model can be based on it. Thus be representative across genres.
- The music used should cover the whole emotional space when using a valence and arousal two-dimensional model.
- The emotional data obtained from participants should be reliable so a model is feasible and reflects the actual feelings expressed in music.
- The length of excerpt should have a length which on one side is not so short so as to demand a high amount of cognitive strain on participants to rate them but also not so long that large changes in expressed emotions, heavily influence the ratings.

Issues around the experimental planning that should be considered are the following

- The length and number of excerpts also come in to play in the experimental planning. A golden rule within listening experiments is that each participant can only be asked to concentrate on a single task, such as a listening experiment for no more than an hour. This not including introduction and instructions. (see e.g. [Zacharov and Bech, 2006]). Given musical data how long each participant take to rate in average, should be investigated.
- The order of the excerpts that are presented to the test subject can potentially have an effect on the ratings. No previous work has been found in documenting whether an effect is seen on the emotional rating if e.g. a heavy metal track is played before a classical piece or vice versa. It has been shown as discussed in section 2.3.5 that the ratings and experiences of listeners do in some way get added up in a non-linear way to some final post-rating percept, but nothing specific about order.
- A limitation of the testing, is the number of possible test participants is not being endless. A compromise has to be made between how many excerpts will be rated and how many ratings is needed for each excerpts in order for the data to be reliable. In [Schmidt and Kim, 2010] they rate 240 songs that is evenly distributed across the four quadrants of the A-V space, where they undergo “intense” rating. No information regarding how many different participants rated these excerpt or whether they rated themselves are available.

Measuring the expressed emotions is what is the aim using the listening experiment. The sources of emotional induction as was discussed in section 2.3.2 should be limited, to reduce bias. This can be done by controlling the experimental variables. Following points are considered,

- Listeners, Contextual and Performance features can be controlled by the listening room and reproduction system. These should be as neutral as possible, so not to influence the test participants. Furthermore test participants should be instructed to disregard any potential musical preferences.
- Musical expectancy should be limited so participants only rate what is being presented, and not what they know is about to come, they are familiar with the given musical song that the excerpt is taken from.
- Brain stem reflexes can partially be controlled by the volume of the music in which it is presented to the participant in.
- Emotional contagion can be limited in the visual sense by having a clean listening room with have monotone colors. If lyrics are learned or the music does result in some contagion, only instructions can be used to minimize this. The length of the excerpts can be chosen to be a length that ensures that participants are not heavily influenced by the learning of the lyrics.

However if long excerpts are needed, division of these into smaller segments and ordering them appropriately can reduce the contagion due to learning of lyrics.

- Episodic memory and Evaluative conditioning are two factors that are difficult to control and can only be done by instructions to the user as to focus what is being expressed in the music and what the person perceives.
- Perceptive influences can be partially controlled by determining whether or not people have normal hearing ability. This could e.g. be done by acquiring an audiogram or ask if participants have one. In this context it is within 20 dB HL which is considered normal hearing. Other more complex disorders should not be taken into account.
- A cognitive variable that has to be taken into account is whether or not participants are on any mood altering drugs, or suffer from disorder that might change their perception or induction of emotions. Simple questions to participants can reveal this issue.

The general bias that can occur when measuring emotions were discussed in section 2.2.3. To control these bias, following aspects should be taken into account,

- *Demand characteristics* is not a major concern for the primary experimental setup of rating excerpts on two emotional scales. The main objective is that the participants understand the underlying idea behind the test. The issue can come in to play at different pilot experiments where effects of the ordering of excerpts or ordering of scales.
- *Self – presentation bias* can be reduced by making all ratings completely anonymous, and position participants isolated so exchange of any kind can not take place that can influence their behavior.
- *Limitation of the awareness of ones emotions* is a major issue in the test, regardless of what scales chosen. The underlying premise is that participants know how they feel about any piece of music. If participants fundamentally do not know or can not express how they feel about a piece of music, then this could lead to biased data. This issue is difficult to take in to account other than data interpretation.
- *Communication bias* is an extension of the last point, which is whether or not participants can, given they know what they feel, rate this on two given scales. If they do not know what the scales mean or represent the data it becomes error prone as a result. Thorough specific instruction has to be given to participants prior to testing. But not so much as to lead them into being biased by the instruction themselves.

The choice of scales and self-report methods following initial considerations are made

- Whether to use a continuous measuring method or use post rating of excerpts in order to capture the temporal emotional dimension. It was previously argued that post ratings should be used. The main concern at present point is that the test should be intuitive and not much training can be made, if any at all. This means that the cognitive load on participants should be kept at a minimum to provide stable result. For this reason knowing the possible bias involved as discussed in section 2.3.5, post rating of excerpts is still the choice. The precise length should be comparable with those used in the *MIR* community or multiple of them.
- When rating two dimensions of Valence and Arousal a possibility is to use either two one-dimensional scales or one two-dimensional scale (2D valence-arousal). The use of two separate scales, can cause bias towards the order of presentation of these, e.g. if arousal is always presented before valence or vice-versa. Instead of presenting both scales for each excerpt one scale could also be presented per excerpt, but this would reduce the data acquired by a factor of two.
- Regardless of a two-dimensional scale (2D valence-arousal) or two single dimensional scales, the direction can potentially have an influence on the ratings, e.g. if positive valence is always to the right and the negative is to the left.
- The use of bipolar or unipolar scales to measure the two bipolar dimensions of Valence and Arousal could possibly increase the number of scales by a factor of 2. At present point no investigation is initiated to prove or disprove the dimensions or bipolarity of Valence and Arousal and therefore bipolar scales should be used.
- Issues like the scales used in the self-report method, are also elements that should be tested prior to the test. Since no previous work has been found documenting it, it should be investigated in a pilot experiment.

This multitude of initial considerations will be investigated throughout the following sections, and result in pilot experiments that will test and clarify these.

## 4.2 Design of listening experiments

To design the experimental procedure, both the independent and the dependent variables have to be discussed and defined. These variables are responsible for the question presented to the test participant and later the quantification of this answer. The response attribute is the question to the participant, the dependent variable is the answer the test participant provides, and the response format is the method of quantifying this answer.

## 4.3 Quantification of emotion

The quantification of emotion is done by means of self-report method. In [Zentner and Eerola, 2010] a great variety of method for quantifying the emotional content in music was discussed. Key issue is that the scale has to be easy to understand and intuitive. Another aspect is that it should be understood by a great variety of different types of people, e.g. education, age and nationality. For this reason valenced adjectives should be limited due to vocabulary limitation by the test subjects, to limit bias.

### 4.3.1 Response attribute

The vocabulary of use should be limited, in order to measure the two dimensions of valence and arousal for the response to music. One could as suggested in [Zacharov and Bech, 2006] develop a consensus or individual vocabulary and thereby let the test participants suggest what adjectives should be used being representative for the two dimensions. This however is very time consuming and therefore a costly affair. The question to participants therefore be a direct one to rate the music using the two dimensions of valence and arousal.

### 4.3.2 Response format

Which response format the expressed valence and arousal in music should be measured with is very much linked with the response attribute. As discussed previously in [Zentner and Eerola, 2010] a number of methods and scales are presented. The constraint there is on the use of adjectives, visual response formats are chosen. In [Bradley and Lang, 1994] they present the *Self-Assessment Manikin* (SAM) that consist of a series of drawn iconic images of a human representing pleasure, arousal and dominance (see figure C.6 in the section C.1.4). Where each series of images have a 9-point scale below them. Using this method, under the assumption that the images are understood, reduce the vocabulary bias. The manikins have effectively been used to measure emotional responses to pictures, images, sounds, advertisement, painful stimuli, etc. They have been used in tests on children, anxiety patients, analogue phobics, psychopaths and other clinical populations. Showing that the method has been widely tested in the psychological world. Within the *MIR* community or music testing with this method has not been found. Given the shortcomings the test is still used to its status as standard. As mentioned in section 2.4 only valence and arousal dimensions will be used in this experiment, and therefore the corresponding manikin will only be used. The biggest issue with using such manikins is whether or not the test participants understand them (i.e. communication bias), and furthermore find them appropriate in rating music. This should be investigated in the pilot experiments in order to make sure that people actually understand them, without any adjectives to describe the images.

Given the method of measuring expressed emotions in music, two pilot experiments are designed to gather appropriate information about the setup of an listening experiment.

- **Pilot 1** Will deal with the ordering of the stimuli, excerpt length, and appropriateness and understanding of scales giving no instruction about the scales. It will also deal with the time it takes to rate an excerpt and potential influence of mood prior to testing and musical experience and training of participants.
- **Pilot 2** Based on the findings in *pilot1*, an appropriate amount of excerpts will be rated. Ratings should be the foundation of a mathematical model. Investigations in to the length of excerpts, effect of presentation order of scales appropriateness and understanding of scales based on finding in the first experiment.

Common for the listening experiments are most of the independent variables, which includes calibration, reproduction system, listening room and user interface. The changing elements are the stimuli and the test subjects that vary between each of the tests. Statistical tool should be used where necessary in order for verification of the results.

## 4.4 Common experimental variables

In this section the instructions, user interface, reproduction system including the listening room and the calibration used for all the listening test is presented. The general requirement for these aspects of a listening experiment is to reduce the influences that were discussed in section 2.3.2.

### 4.4.1 User interface

Given that the manikins are self explanatory the interface should be easy, intuitive and require as little amount of work from the test subject as possible. To make the whole process automated a computer program should be devised. This should be designed to

- Minimize user interaction e.g. mouse clicks, typing on keyboard, etc.
- Timing of each participant to rate an excerpt.
- Use manikins from [Bradley and Lang, 1994]

The resulting interface for the rating of emotions expressed in music can be seen on figure C.6. To ensure that the interface is understood by participants, a

pre-test will be made prior to all experiments. This will be marked with a clear indication so participants are sure it is a test.

#### 4.4.2 Instructions

The instruction given to the participants should emphasize that what should be rated is the expressed emotions in music and not what is being felt. Therefore clear instructions as to how the test procedure will be should be made. The written instruction can be seen in section on figure C.1 and C.2. Taking into account that the written instructions were not understood a verbal repeat of the written instructions were also given. Participants were allowed to ask question within the scope of each test.

#### 4.4.3 Reproduction system

Two major different setups are evaluated for the reproduction system. One being a stereo speaker setup and the other using headphones. Whether or not music is indented for playback on a stereo or headphones is a long discussion. Being aware of the possible bias being enforced on the experiment, a setup using headphones is chosen for practical issues. Following equipment is used for the experiments

- ATCAHR4 computer using Windows XP.
- Matlab R2006b.
- Sennheiser HD 580 Precision headphones.
- *pa\_wavplay* matlab plugging is used for playback.

#### 4.4.4 Listening room

The listening room used for all listening tests is the *CAHR* Right booth at Center for Applied Hearing Research at the Acoustical department in the Technical University of Denmark. The room is chosen due to its very monotone appearance, with gray walls and a computer placed inside. The booth is sound insulated so that no exterior sound influences are present during testing.

#### 4.4.5 Calibration

The musical excerpts should be normalized so that the presented playback volume is the same. By playing one excerpt louder or softer could influence the subjective evaluation of the experienced emotion in the excerpt. It is clear that when using multiple musical genres from rock to classical and pop, the temporal evolution of the music is very different. A normalization using Loudness or

Specific Loudness could be an option. However changing the overall playback volume frequency dependent, could change the original intended expression by the artist, due to the change in frequency balance. An overall Loudness normalization could also be used, but the perceptual consequences are unknown and further investigations of this is out of scope and will not be covered. A widely used normalization method in the form of *RMS* is also perceptually unknown but chosen here for convenience.

#### 4.4.6 Order of presentation

Two different emotional scales exist, that of valence and arousal. Which of the two scales should be presented first to the participant, or should only one scale be presented to the participant at a time, to reduce correlation between the two emotional scales. Another issue is in which direction the two scales should be directed. Should the positive or negative valence be to the left or right. Due to time constraints and the limited number of test participants, both scales will be presented to the test participant for each excerpt. To balance the design the valence and arousal scales should be presented equally first and last. The issue of the direction of scales (i.e. happy to the left or right), will not be closer investigated within these listening experiments and the scales are as presented in [Bradley and Lang, 1994].

### 4.5 Pilot 1

The pilot1 experiment has the objective to make investigations regarding an experimental setup. The main issues of the test are

- Main concern for this experiment is to test the ordering of the stimuli, which would be the most pertinent to use in a potential test.
- The appropriateness and understanding of the scales for the use in music.
- The length of the musical excerpts: whether or not the test participants find it appropriate.
- The duration of the test and the time it takes to rate each excerpt should be measured and investigated.
- Questions describing the participant, e.g. musical experience, and demographic data.

Given these concerns to be able to have something to test, a testable paradigm is formulated. For all other aspects these will merely be information gathering.

#### 4.5.1 Test paradigm

Here the test paradigm is given for the main objective of the test.



- **Premise 1**, There is a change in subjects rating of emotion expressed in music due to the musical excerpt previously presented (Carry-over effect).
- **Premise 2**, The emotional carry over effect can be measured using listening experiments.
- **Conclusion**, A carry over effect of emotion exist due to the ordering of musical excerpts.

The experimental variables for the first pilot experiment is presented.

### 4.5.2 Subjects

For the first pilot experiment 24 participants are chosen amongst students and employees of the acoustical department and the department of Informatics at DTU. The first 12 participants should be used in the testing of the sequential presentation and the other 12 should be used for the balanced structural design.

### 4.5.3 Stimuli

The stimuli chosen here is three different excerpts that are different in genre and relative unknown to the author. Genres are pop/r&b, pop/rock and heavy rock. The total length of each clip is 30 seconds chosen in the middle of each song.

1. Metallica - Better than You.
2. U2 - In Gods Country.
3. Back Street Boys - Spanish Eyes.

Each clip is divided into 4 equal excerpts of 7.5 seconds with a Hanning shaped fade in and out of 0.2 seconds. The format is PCM at 44.1 kHz sampling, 16-bit stereo wave files.

### 4.5.4 Order of stimuli

To test the emotional carry over effect of musical excerpts, two different experimental order of stimuli must be constructed.

**Sequential ordering (SO)**, is ordered so that the 4 clips of each song is presented sequential, but each song should be ordered so that the carryover of the last clip in each song and the first in each song is balanced. The first excerpt presented to the participant should also be different so that no carryover exists from e.g. pre-testing. Here completely random or a simple Latin Square does not suffice.

**Balanced ordering (BO)**, is the possible combinations of the 12 clips in a balanced design. So that excerpt one and excerpt two are only succeeding each

other once. It should also hold that the starting sound excerpt should also be different for each test participant so that no emotional carry-over effect is present from pre-testing.

For both these cases a balanced Latin Square design is the solution to these requirements, the so called *Williams Latin Square* (WLS).

#### 4.5.5 Data processing

To compare the two set of results, the *BO* and *SO* of excerpt an objective measure should be used. Given that for each excerpt 12 ratings are made for each of the methods, these two groups will be compared with a *Two-sample Kolmogorov – Smirnov* (2KS) test. The null hypothesis for the 2KS test is that the two measurements are from the same continuous distribution. The alternative hypothesis is that they are from different continuous distributions. For each excerpt a test is made between the *SO* and the *BO* results. To test the effect of the *BO* and *SO* within each clip, the 4 excerpts within each clip will equally be tested with 2KS test.

An alternative method of testing the difference between the two is to simply compare the sample-mean of the given two distributions, not implying no specific underlying distribution at present point. Furthermore no regard is made on outliers on either of the datasets. Another aspect to investigate is that one could expect participants given a sequential ordering, would give more consistent ratings Therefore this will be investigated using variance analysis of ratings across excerpts for each participant.

#### 4.5.6 Meta data

Demographical data should be gathered about each test participants together with some personal information about musical background and medical information.

- Age
- Occupation
- Time listening to music per day
- Years of musical training
- Preferred musical genres
- Suffer from mood disorders
- On any anti depressive medicine

The questionnaire can be seen in section C.1.2 on figure C.3. To test scales and excerpts length questions should be asked after the experiment has completed, which can be seen in section on figure C.4.

- Understanding of scales
- Appropriateness of scales
- Appropriateness of excerpt length

To see if their mood prior to the test has an influence on their ratings, this should be rated prior to testing, this can be seen in section C.1.3.

### 4.5.7 Results

All participants were normal hearing and did not suffer of any mood disorders or were on any medication. All participants were PhD student or master students with an average age of 27 years. Using the  $2KS$  test between each of the 4 excerpts, that represent a 30 second clip, for all excerpts using both designs, the NULL hypothesis was accepted. It is likely that regardless of the ordering method within each of the excerpts, that they originate from the same underlying distributions for each of the clips.

To test the ratings across ordering designs each excerpt e.g. 1, 2, 3 and 4 excerpts from clip 1 in the *BO* design, was compared to the ratings of the same excerpts from the *SO* design.

Clip	Excerpt	1				2				3			
		1	2	3	4	5	6	7	8	9	10	11	12
$2KS$	Valence	0	0	0	0	0	0	0	0	1	0	0	0
	Arousal	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.1: Results of *Two-sample Kolmogorov – Smirnov* test between the results obtained in the *pilot1* experiment comparing results from a sequential and balanced design, to see the effect on emotional ratings of music using the scales of Valence and Arousal.

On table 4.1 the results of the *Two-sample Kolmogorov – Smirnov* test is shown, where all excerpts pass the test, except no. 9 which is the first excerpt in *Back Street Boys - Spanish Eyes*. A histogram of those ratings can be seen on figure 4.1. By visual inspection there is a clear difference between the two, where the sequential produces much more consistent results, than that of the balanced. Another thing is that the mean of the ratings has changed by 1.

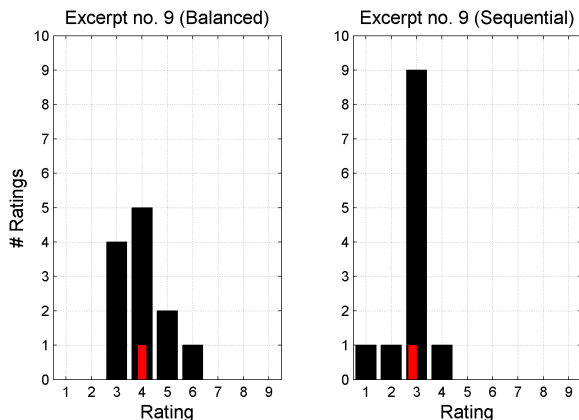


Figure 4.1: Histogram of the ratings of excerpt 9, in the balanced and sequential ordered experiment. The  $KS2$  test in the two set of ratings was the only that failed. The red bar marks the mean of the ratings in the given experiment.

The difference in mean and variance between the two ordering methods, is used to visually having a measure that is easier to interpret. The results are shown on figure 4.2. The difference between the means show that across the 12 excerpts, using the *SO* on the arousal scale, participants produce higher ratings compared to using *BO*. On the valence ratings, specially the ratings for *Back Street Boys* show that using *BO* produce ratings that are in average 1 rating higher than that of the sequential.

By using a *SO* design, it it was tested if the ratings given by participants would become more consistent. The results are shown on figure 4.3, where a slight change in the average variance over all participants, with 0.38 for *SO* and 0.75 for *BO* on the valence scales is observed. 0.38 for *SO* and 0.56 for *BO* on the arousal scales, which is not a huge difference. On arousal ratings of excerpt no. 5 or valence ratings excerpt no. 11 and 12, the *BO* has a lower variance than the *SO*. On the other hand valence ratings of excerpt no. 5 and arousal ratings of excerpt no. 4 and 9 show a much higher variance.

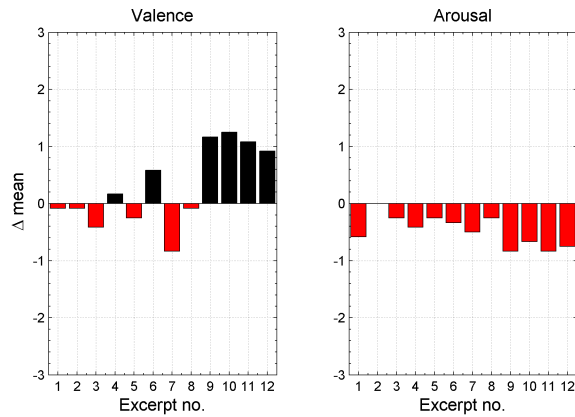


Figure 4.2: The difference between the mean of the ratings given in the pilot1 experiment. Bars that are red, indicate that the sequential design produced higher ratings than the balanced design. Vice versa the black indicates that the balanced design produced higher ratings than the sequential.

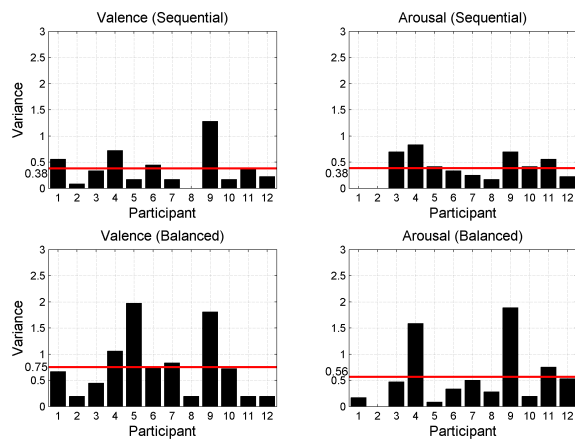


Figure 4.3: Variance of each participant's ratings. The variance was calculated for each clip consisting of 4 excerpts. The average was then taken between each of the 3 clips, used in the test. The red line indicated the average over all participants, where it is numerical indicated on the left of the line.

### Meta data results

An analysis of the meta data gathered was made in C.2. The analysis of the time it took to rate an excerpt with the two scales was analyzed in section C.2.1. It was shown that across participants only participant 1 in the *BO* design seemed to find the test very straining. Comparing the two methods, the time it took to rate for each participant in average was 6 seconds for *SO* design and 8.6 seconds for the balanced design. If the time taken to rate is taken as a sign of the cognitive load, it shows that using a balanced design uses a great deal of cognitive power compared to the *SO*. This trend is explained by looking at the time it takes to rate the individual excerpts where it can be seen that rating the first excerpt in each clip takes the longest, and after that decreases in time. Whereas the *BO* design is rather constant.

An analysis of the questions posted to the participants regarding the understanding and appropriateness of the scales was investigated, in section C.2.2 and C.2.3. Based on the results and post questioning of participants, it was found that the scale of arousal was difficult to understand. Especially for the use in music. Regarding the length of the excerpts, it was found that participants had to use a high amount of cognitive power to rate the excerpts of 7.5 seconds, and they would prefer something longer. They were prone to rate similar results, e.g. the middle of the scales, when they were in doubt because there was too little music to make a proper assessment.

### 4.5.8 Discussion

The effect of the ordering of excerpts in the rating of emotions expressed in music using the dimensions of valence and arousal was tested. The test method was the *2KS* test and visual comparison by comparison of mean. 24 participants were divided in two groups that used a *SO* and *BO* design. The *2KS* showed that within each ordering method, the excerpts were likely to originate from the same underlying probability distribution. That is, regardless of methods used to order the excerpts, the results were consistent within a margin. This result is very important as it shows, that both methods can be used and still obtain consistent results. The second test was to test whether or not there was a change in ratings given by users, due to the excerpt that came prior to the rated. Using *2KS* test one excerpt failed the test. One can suspect based on these results, that using the sequential ordering, participants are more prone to rate the same in succession. Thus producing a great deal of the same ratings, but on 4.3 a clear difference is not seen. In some cases the *BO* ratings have a slightly lower variance, but in some cases are also much larger than *SO*. It could be that by only using 3 clips, participants memorize their ratings when using the *BO*, thus still producing rather consistent ratings. On the other hand it could just be chance. Given the failing of *2KS* on excerpt no. 9, the general larger ratings on the arousal scale and in some cases much larger on the valence scale on e.g. *Back Street Boys*, there is a difference between the two types of

ordering of stimuli, as the purpose of this test was set out to investigate. “Pros” and “Cons” between the two methods is that *SO* does produce slightly more consistent results and results are somewhat different than using *BO*. *BO* does not produce a much larger variance between ratings, and by enough participants it could result in more neutral results. Depending on what is desired to model, e.g. if one wants to model the emotional buildup in a song, *SO* would be the method to use. If one wishes to model the emotions expressed in music and disregard any temporal buildup as it is in this case, the *BO* would be the method to choose.

Based on the findings in questionnaire and post questioning of participants, it was found that a more detailed explanation of the scales should be given, specifically the arousal scale. Another finding was that longer excerpt should be used, as participants became very cognitively exhausted, making them rate “neutral” (in the middle of the scale), not to account for errors in the test. This is confirmed by a temporal analysis of the ratings, where it shows that it takes a longer time to rate each excerpt in the *BO* design than in the *SO* design.

## 4.6 Pilot 2

The pilot2 experiment has the objective to make further investigations regarding an experimental setup. The main issues of the test is

- The experiment should acquire data for the basis of creating a mathematical model, that can predict emotional ratings of valence and arousal.
- To test whether or not the participants can be asked to rate on both scales every time an excerpt is presented. It could be that there is an enforced correlation between the two scales, if they are both used.
- Given the experimental emotional data, an analysis of it should be made, to clean the data for potential outliers.
- Acquire meta data regarding appropriateness and understanding of the scales used, when they are explained.
- Investigate meta data regarding the participants that can potentially influence the results from the experiment, e.g time to rate, listening habits, musical training, familiarity of the music and mood prior to the test.

### 4.6.1 Test paradigm

Here the test paradigm is given for the test of the change in participants ratings due to the order of presentation of scales.

- **Premise 1**, There is a change in subjects rating of emotion expressed in music, due to the order the scales are presented to participants.

- **Premise 2**, The effect the ordering of scales have on emotional ratings, can be measured using a listening experiment.
- **Conclusion**, An effect is present on emotion ratings due to the ordering of the rating scales.

## 4.6.2 Subjects

For the second pilot experiment 14 participants were willing to participate, amongst students and employees of the acoustical department at DTU.

## 4.6.3 Stimuli

The stimuli chosen here should cover the whole valence-arousal space. Due to the results of *pilot1*, a length of each excerpt is chosen to be 15 seconds. Another finding was that the response time for each rating was between 3-20 seconds. Although there was an innate lag of the timing in the Matlab system, so they are not completely reliable. Once a routine is built in after a long test session the response time is presumably going to go down, and the scales have been given a clearer definition. Thus estimating around 2-6 seconds of time to rate one excerpt. This totals of 17-21 seconds per clip. Aiming for 1-hour total testing time, it results in 200 excerpts of evaluation. This could exceed the golden 1-hour rule of auditory testing but it is seen as acceptable since the task at hand should not put the test participant under heavy cognitive load, when listening to music. Users will also be instructed to have a break if they become cognitively exhausted.

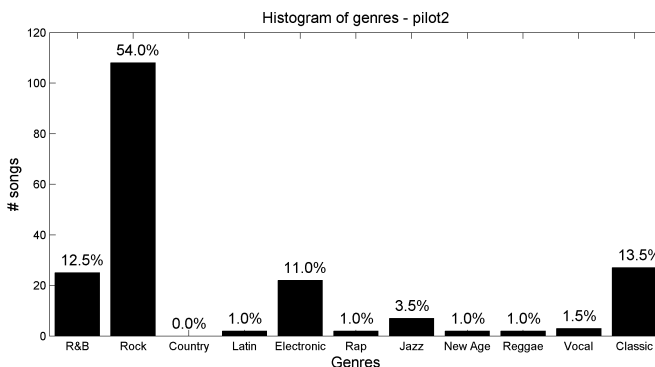


Figure 4.4: Histogram of the genres for each of the 200 tracks used in pilot 2 experiment. The genres used were those of the *AMG*, and the same genres as the *USPOP2002* including a classical, which contains all subgenres of classical genre, opera and scoring/film music. The number above each bar indicates the percentage of the total amount of excerpts



The 200 excerpts are chosen among all musical data available by the author using self-rating of excerpts and the number of excerpts in each quadrant is calculated post test. The same genres as the *USPOP2002* dataset is used to approximate the same distribution, where a second genre is added in the form of classical (see figure 4.4). This genre is a collection of opera, film music, and all varieties of classical music. These types of music are specially made to elicit emotional responses in listeners as was discussed in [Cohen, 2010]. The method obtaining the data has been mining webradios, and therefore the whole musical tracks have not been available and the quality has been changing from excerpt. The distribution of bitrates for the data can be seen in section C.4.

#### 4.6.4 Order of stimuli

A complete *WLS* design of the 200 excerpts would require 200 test participants. Since this is not practically possible a compromise should be made. Concatenated individual *WLS* of 20 test participants where the *WLS* has a randomized initial column is used. Using this method the number of excerpts in the test should be a multiplum of the number of test participants. Where each concatenated *WLS* should be increased with the number of test participants.

#### 4.6.5 Instructions

All instruction for the second pilot experiment was the same as *pilot1* with one exception. Given the results from *pilot1* a thorough verbal explanation of the scales should be given to each participant. The scale of valence was explained to be *positive* or *negative* using very common and yet neutral anchors. The scale of arousal was explained to be *excited* or *not excited*. Still emphasizing that it is the expressed not induced emotion.

#### 4.6.6 Data processing

As can be seen in 4.5 the data for valence and arousal on the two scales are illustrated. The two 9-point scales used for the manikins are in nature ordinal scales, that is, the distance between the first and second point does not have to be the same distance as the third and fourth point. Even within each subject the mapping of each point can be different. The ratings from each participant form two histograms on a 9-point scale. To model this, the underlying structure of data, a distribution should be fitted to this data. Although the data is ordinal, for the purpose of modeling this approach, it is used for practical issues. The distribution should only contain probability mass in the given interval of the 9 point, and should be 0 everywhere else. The truncated normal distribution or the *beta* distribution has those qualities, in this case the beta distribution is used and given by

$$prob(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt} \quad (4.1)$$

which is only defined in the interval  $[0; 1]$  with a mean of  $\mu_\beta = \frac{\alpha}{\alpha+\beta}$  and mode  $\frac{\alpha-1}{\alpha+\beta-2}$  for  $\alpha > 1$  and  $\beta > 1$ . The 9-point scale's intervals are defined so that 1<sup>st</sup> ordinal point of the scale is defined in  $x_1 \in [0.5; 1.5]$ , 2<sup>nd</sup> in  $x_2 \in [1.5; 2.5]$  etc. as seen on figure 4.5.

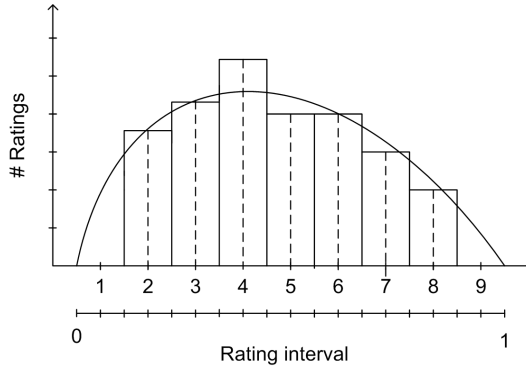


Figure 4.5: Illustration of the fitting of beta distribution to experimental data from listening experiments.

On the figure the dotted lines indicate the rating values given by the participants. This is so called grouped data where one discrete value represents all the data in the interval e.g.  $f(x_1)$  in  $x_1 \in [0.5; 1.5]$  as the participants given the ordinal scale could not rate in between. These intervals are mapped to the rating interval of  $x' \in [0; 1]$  defined by the beta distribution. For simplification reason the beta distribution is fitted on the grouped data using maximum likelihood estimates of the variables  $\alpha$  and  $\beta$ . Alternatively the area under the curve  $A_I = f(x_I + 1) - f(x_I)$  for  $I \in (1, 9)$ , should be used for the *MLE*, but is not seen as necessary here. The mapping is done by  $x_\beta = (2x_r - 1)/18$

**Outlier criteria** Given the beta distribution of the ratings of each excerpt, outlier criteria can be formulated. Two different scenarios are thought of here, for data being an outlier.

1. A test participant has fundamentally misunderstood one or both of the scales, and therefore would make error full ratings consistently.
2. A participant loses concentration due to a number of reasons, and thereby momentarily makes error prone ratings.

The fundamental assumption here is on the experimental data, that the underlying mechanism of objectively rating the expressed emotions in music is somewhat the same for test participants. If ratings are far from the majority,

then this would be considered an outlier. To formulate the two outlier criteria (OC1 and OC2), a center of opinion in the form of the mean of the beta distribution is used.

### OC1

To test if participants fundamentally misunderstood the test, the distance between the rating a participant rates on an excerpt and the mean of all ratings on that given excerpt, is summed up over all excerpts rated by the test participant. This can be written as

$$\Upsilon_j = \sum_{i=1}^N |y_{j,i} - \mu_{\beta_i}|, \quad \text{where } \mu_{\beta_i} = \frac{\alpha_i}{\alpha_i + \beta_i} \quad (4.2)$$

where  $\Upsilon_j$  is the summed error of the  $j^{\text{th}}$  participant of  $K$  total.  $y_{j,i}$  is the rating of test participant  $j$  of the  $i^{\text{th}}$  excerpt out of  $N$  total excerpts.  $\alpha_i$  and  $\beta_i$  are the *MLE* fitted parameters of the beta distribution for excerpt  $i$ . Using the summed error over all excerpts an outlier criteria can be formulated as

$$\hat{\Upsilon} = \sum_{j=1}^K H(\Upsilon_j - \phi_t) \quad (4.3)$$

$\hat{\Upsilon}$  is then the number of outliers of the total  $K$  possible outliers, and  $\phi_t$  is the criteria for which the participant is judged to be an outlier.  $H$  is the Heaviside step function defined by  $H(x) = 1$  for  $x > 0$  else 0. If the participant makes consistently error prone ratings  $\Upsilon_j$  will become large and be considered an outlier if greater than  $\phi_t$ .

### OC2

To test if participants loses concentration and thus makes an error on an excerpt by excerpt basis, the distance between a given rating to the mean of the beta distribution fitted on all ratings of that excerpt is used.

$$\Psi = \sum_{i=1}^N \sum_{j=1}^K H(y_{j,i} - (\mu_{\beta_i} + \sigma_u)) + H((\mu_{\beta_i} - \sigma_l) - y_{j,i}) \quad (4.4)$$

where  $\Psi$  is the number of outliers out of all ratings made in the test.  $\sigma_u$  is the upper and  $\sigma_l$  is the lower limit of which ratings are considered an outlier.

In statistics when estimating the mean of a distribution often so called *trimming* is used where 2 % of the ratings on each tail of a distribution is removed. Given the shape of the beta distribution and the position of ratings,  $\sigma_u$  and  $\sigma_l$  might not be the same to obtain this measure.

To calculate an objective outlier criteria for distributions for each of the  $N$  excerpt,  $L$  samples are drawn.  $\hat{y}_i \sim \text{Beta}(y_i|\alpha_i, \beta_i)$  where  $i = [1, 2, \dots, L]$  simulating  $L$  participants.

The probability of being an outlier, given the two outlier criteria, is then calculated as

- $p(\text{Outlier}|OC1) = \frac{\hat{\Upsilon}}{L}$ .
- $p(\text{Outlier}|OC2) = \frac{\hat{\Psi}}{LN}$ .

where  $L$  should be chosen so that stable results are obtained. The aim is then to find  $\phi_t$ ,  $\sigma_u$  and  $\sigma_l$  appropriately.

### Order of Scales analysis

To test if there is an effect of whether the arousal scale or the valance was presented first, a test should be made of the data. A simple *Two-sample Kolmogorov – Smirnov* (2KS) test is used again. Thus comparing distribution of ratings for each of the excerpts given by half of the participants to the ratings given by the other half, i.e. arousal presented first or valance presented first. The assumption here is that there is enough data to form the basis of comparison.

### 4.6.7 Results

The aim of the choice of musical excerpts by the author was to cover the entire Valence-Arousal space. On figure 4.6(a) the distribution of all ratings are seen. A clear dominance is seen around the center point of the valence scale as seen on figure 4.6(b), where the rating point donated 4, 5 and 6 contribute account for  $\sim 60\%$  of the total amount of ratings, whereas arousal only contribute  $\sim 45\%$  of the total ratings. It is also observed that some participants e.g. participant 1 uses the middle rating point very dominantly on the valence scale.

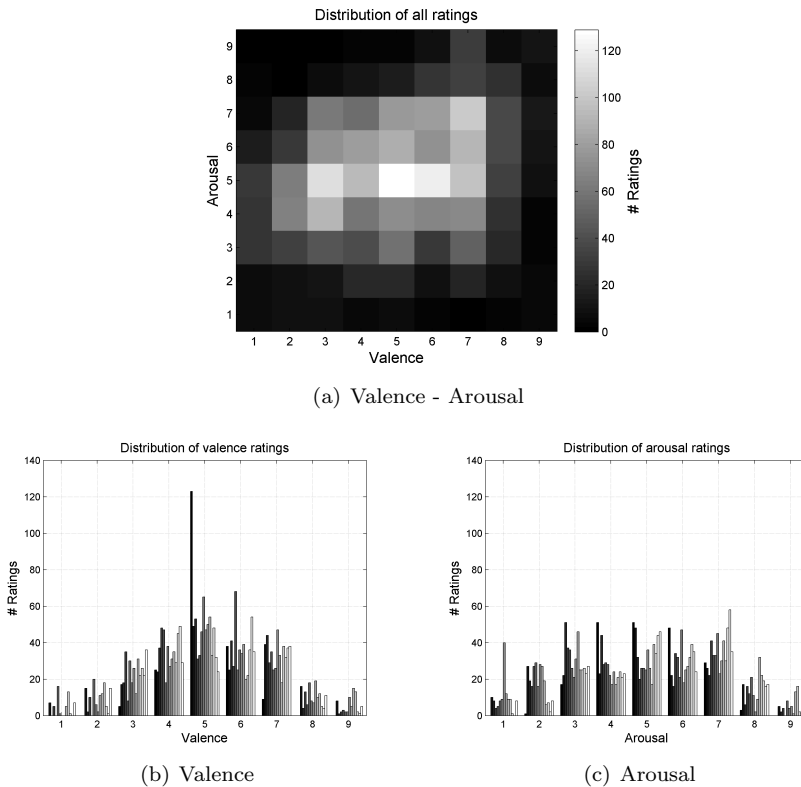
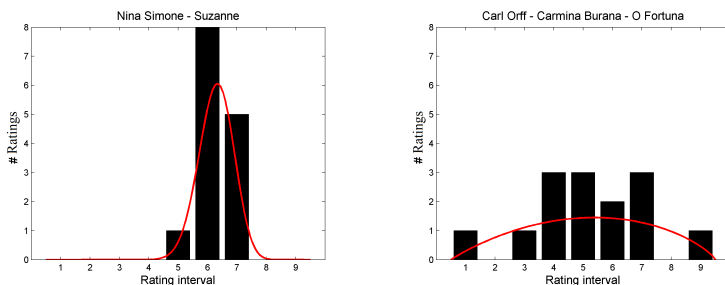


Figure 4.6: On (a) valence and arousal are plotted for the total number of ratings in pilot2 listening experiment. On (b) and (c) valence and arousal ratings accumulated through all rated excerpts are plotted.

Of all the 200 rated excerpts, the two with the highest and the lowest variance is shown on figure 4.7(b) and 4.7(a) respectively. All the histograms for all ratings and the corresponding fitted beta distribution can be seen in section C.5 on figure C.12, C.13, C.14 and C.15 .



(a) Valence rating for Nina Simone - Suzanne,  $\sigma_n = 0.61$  and  $\mu_n = 5.00$ . (b) Arousal ratings for Carl Orff - Carmina Burana - O Fortuna,  $\sigma_n = 2.02$  and  $\mu_n = 5.21$ .

Figure 4.7: On (a) and (b) the histogram and fitted beta distribution for the excerpts that have the highest and lowest normal variance

### Order of Scales

To compare the effect of presenting one scale first and subsequently presenting the second on the same excerpt a  $2KS$  test was made. Given that the ratings of half of the participants were compared to the other half using  $2KS$  test. The results of rating valence, comparing the data where the arousal scale was presented first and the valence scale was presented first, 8 did not pass  $2KS$  Null hypothesis for the valence scale, excerpt 2, 8, 32, 65, 82, 109, 123, 168 and using the same type of test scheme, 5 did not for the arousal scale, excerpts 74, 112, 159, 180 and 183.

### OC1

1.000 samples were drawn from each beta distribution fitted to the experimental data. The accumulated deviation from the mean was calculated for all samples drawn. The resulting histograms for valence and arousal can be seen in section C.8 in figure C.19. Based on visual inspection  $\phi_t = 240$  for valence, and  $\phi_t = 270$  for arousal, based on the fact that this number is much higher than any of the sampled participants. The results are presented in table 4.2.

Using  $OC1$ , test participant 8, 9 and 14 were considered an outlier with a accumulated deviation from mean of 311, 355 and 409 respectively for the arousal scale. Averaging around 1.5-2 ratings away from mean through all the excerpts. If this criteria would be used, this would exclude 300 ratings on the arousal scale equivalent to 21 %.

Participant	1	2	3	4	5	6	7
Arousal	153	196	218	193	196	213	206
Valence	141	226	147	193	150	234	172
Participant	8	9	10	11	12	13	14
Arousal	<b>311</b>	<b>355</b>	247	204	169	189	<b>409</b>
Valence	203	176	211	231	156	182	224

Table 4.2: Accumulated deviation from mean where outliers are considered when  $\phi_t = 270$  for arousal and  $\phi_t = 240$  for valence. Participants that are considered outliers based on the given criteria are marked with bold.

### OC2

The results of the *OC2* is presented in table 4.3 and 4.4. Drawing 1.000 samples from each of the 200 distributions,  $\sigma_l$  and  $\sigma_u$  are determined so that in average over all beta distribution of 2% ratings in each tail will be removed. Using  $\sigma_l$  and  $\sigma_u$  determined on emperical data on the experimental data, it is evident that fewer outliers are seen on the valence scale, than of the arousal scale. In average, if a participant rates 2.6 of a rating interval away from mean, it would be considered an outlier for valence and 2.82 for the arousal scale.

The valence ratings show that applying the same criteria that removes a total

	Value	Empirical		Experimental	
		Outliers	Percentage	Outliers	Percentage
$\sigma_l$	2.6	4.059	2.03%	29	1.04%
$\sigma_u$	2.6	3.937	1.97%	56	2.00%
Total		7.996	4.00%	85	3.04%

Table 4.3: Results of outlier criteria 2 on the valence ratings, Empirical data where  $\sigma_l$  and  $\sigma_u$  estimated to reach 2% in each tail. Resulting outlier removal on experimental data to the right

of 4% of the ratings using emperical data, removes 3.04% of the ratings on experimental data. On the arousal scale this removes 5.71% of the ratings.

Using the *OC2*, thus removing 160 arousal ratings and 85 valence ratings of

	Value	Empirical		Experimental	
		Outliers	Percentage	Outliers	Percentage
$\sigma_l$	2.82	4.060	2.03%	76	2.71%
$\sigma_u$	2.82	3.894	1.95%	84	3.00%
Total		7.954	3.98%	160	5.71%

Table 4.4: Results of Outlier criteria 2 on the arousal ratings, Empirical data where  $\sigma_l$  and  $\sigma_u$  estimated to reach 2% in each tail. Resulting outlier removal on experimental data to the right

the total 2800 ratings, how the outliers distribute over participants and specific excerpts is shown in section C.9 on figures C.20(b) and C.21(b). It is evident that participants 8, 9 and 14 have clearly the highest amount of ratings removed. An example of an outlier removed can be seen on figure 4.8.

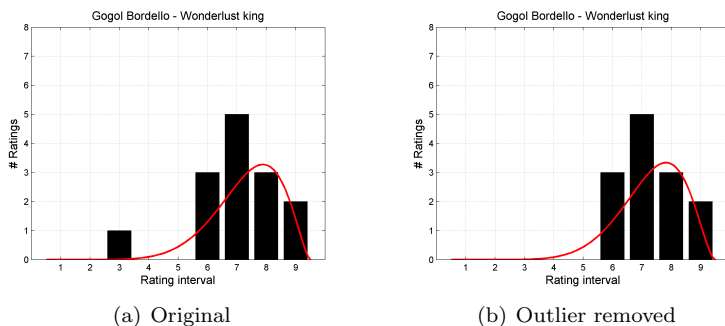


Figure 4.8: On (a) the histogram and fitted distribution for the ratings of the 15 second excerpt of *Gogol Bordello - Wonderlust king* is plotted and on figure (b) the *OC2* has been applied and removed 1 rating.

Using the data where outliers are removed, new beta distributions are estimated, the resulting  $\alpha$  and  $\beta$  of the beta distributions are plotted on figure 4.9. There is a clear clustering of the  $\alpha$  and  $\beta$  coefficients describing the beta distribution. The properties of the beta distribution is that when  $\alpha = \beta$ , the distribution is centered, in this case around 5 or 0.5 of the beta scale. The higher the coefficients become the more narrow the distribution becomes. Excerpt 154 is a narrow centered distribution meaning that people agreed on the valence rating on that excerpt. Once the coefficients are off diagonal, e.g.  $\alpha > \beta$  ratings are in the higher range e.g. very happy or excited and vice versa. The grouping shows that not very narrow distributions are fitted and that there is a tendency towards higher ratings on the valence scale.

The distribution of the beta means for both the arousal and valence scales were compared in section C.10 on figure C.22 and C.23. To see the effect of the outlier removal in the distribution of the beta distributions. It was shown that by using *OC2* the distribution for the arousal scale became wider thus covering a greater area of the emotional space. No significant changes could be seen on the valence scale.



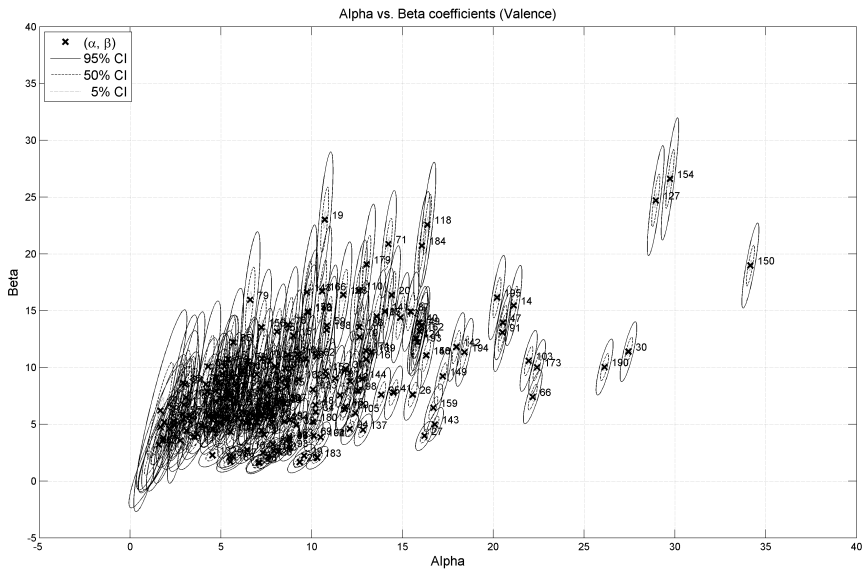


Figure 4.9: Alpha and beta coefficients of the fitted beta distributions plotted with the 5%, 50% and 95% confidence intervals. Numbers indicate the excerpt number from 1-200. The confidence intervals were calculated using the negative beta log likelihood function given  $\alpha$ ,  $\beta$  and the experimental data where outliers were removed using *OC2*, using the asymptotic covariance matrix.

### Prior mood

Prior to any emotional rating of excerpts, participants were asked how they would rate their emotional state at that present point. The results can be seen on figures C.16 and C.17 in section C.6. Visually there is a large distance between the prior emotional state and the ratings given during the test in some cases, and in others they seem very much similar. To test if there is a connection between participants prior mood and the ratings provided for all excerpts and analysis was made in section C.7. The distribution of all ratings provided by each participant was parameterized by using a variation measure of the different between the 50 th and the 25 th percentile. This was compared to each participants pre-emotional rating, using visual inspection of scatter plots and correlation coefficients. It was shown that no measurable connection between the two was present and participant seemingly were able to disregard their mood when ratings excerpts.

### Timing data

The time it took to rate an excerpt for each of the participants was measured using simple timing algorithms; the results can be seen on figure C.24. The time it took each of the participants to rate one excerpt as an average was examined. It showed that participant 9 and 14 took 4 and 3 seconds longer to rate each excerpt respectively. This could indicate that they found the task difficult. Analyzing across all excerpts it was found that some excerpts required a great deal of extra time to rate than the average. Thus indicating that they were more difficult than the others. Removing such excerpt based on this measure is difficult as the data acquired could potentially be very relevant for future mathematical modeling.

### Understanding of scales

In section C.11 a thorough analysis was made of the data acquired about participants during the test. The same set of questions regarding the scales and the length of excerpts as in *pilot1* were asked here. The result was that by explaining the scales better using verbal communication, the participants found it to be much easier to understand and use them. Post questioning revealed that in some cases there was still some confusion as to the scales. The length of the excerpt increasing to 15 seconds produced much higher appropriateness ratings than in *pilot1* from below average to above good. An analysis was made into the correlation between participants understanding of scales and their emotional ratings in section C.12. The results was that no measurable connection could be found.

### Musical background

The musical training, preference, time spend listening each day and the familiarity of excerpts in the test were also investigated in section C.11.3. To quantify

any potential influence all these aspects about participants could have on their emotional ratings an analysis was made in section C.12. The procedure was the same as done in analyzing the influence of prior mood to participants ratings. The results showed that no connection could be found between any of the meta data parameters and the participants emotional ratings.

### 4.6.8 Discussion

The clear center point on figure 4.6(b), post questioning of the test participants revealed that every time a test participant was in doubt of what to rate, center ratings were often used. Using a bipolar scale also communicates that when the test person does not feel that it is neither positive nor negative, e.g. they feel nothing. Then the middle score would be rated on the valence scale. This is specially seen for valence, where no particular emotions were expressed by the music, according to the participants. This could later in the mathematical modeling pose as a problem if the acoustical features are very different and the ratings are the same.

Another issue is the spanning of ratings across excerpts, the music chosen for *pilot2* should have been chosen to fill out the entire emotional space, but did not completely. One explanation is that often test participants do not like rating end-points. They save those extremes to that “special” excerpt where they really feel something. It could also be that given the scales/manikins, no music can actually fill the full extent of the emotional space due to the medium music and the scales themselves. Due to the relative few test participants no real conclusion on this matter can be made. For the purpose of acquiring data that fill the emotional space for the purpose of designing a mathematical model, the aim has been reached.

Using *SAM* as ratings scales, the aim was to find scales that were self-explanatory, but given users feedback both using post questioning and verbal communication after the test, participants had a problem with the scales. It was necessary to explain the scales to participants and in future testing alternatives could be investigated.

A comparison and investigation of whether or not there was an effect of presenting both scales for participants to rate each excerpt. The *2KS* test showed that on the arousal scale, 8 excerpt failed the NULL hypothesis and for the valence scale 5 failed out of the total 200. The comparison was made on a rather small dataset comparing 7 ratings to each other. The reason for making participants rate on both scales for each excerpt is the element of time. It would half the experimental time, since participants should only listen to an excerpt once, and producing 2 ratings. A simple correlation analysis was attempted between the two datasets, but due to the relative few ratings, nothing conclusive could be made. It is a fact that there seem to be a difference between the order of the scale presentation, and further investigations could be made at a later stage, with a higher amount of participants. To balance the effect, in future designs

equally the arousal and valence scale should be presented first as was done in this case.

Prior to the test start participants were asked to rate their mood on the same scales as the experiment. One could expect that the mood prior to test could have an influence on the subsequent ratings of excerpts. Using a variation measure of the emotional ratings provided by participants and comparing these to pre-emotional ratings no connection could be found between the two. This was confirmed by calculating  $r^2$  coefficients between two for both valence and arousal with a correlation of 0.218 and 0.227 for them respectively.

Musical experience was measured using two different variables, which are training and time spend listening to music every day. Lastly a similar measure was measured in the form of familiarity of the 200 excerpts measured. The same analysis as was performed on pre-emotional ratings were performed on the meta data for each participant. However no connection between any of the meta data for participants and their ratings could be found. One approach in future work could be to group data according to any or more of these meta data and subsequently training models on these data as was done in [Yang et al., 2007] to obtain more personalized models. For now this is not used.

Two outlier criteria were compared *OC1* and *OC2*. To make an empirical outlier measure for both the criteria, 1.000 samples were drawn from each of the fitted distributions, thus simulating 1.000 participants. Using a conservative measure where no participants were considered an outlier on empirical data 3 participants fell for this criteria using *OC1* on the arousal scale. Thus removing 300 ratings out of 1400 ratings equivalent to 21 %. *OC2* criteria was determined in the same way as *OC1* with 1.000 samples from each distribution. For arousal this results in average that 5.9% of the ratings are removed, and 3.46% for valence. *OC2* on one hand takes into mind that all participants can make mistakes or be mentally distracted. On the other hand imposing a grouping phenomena, that all people to a certain degree should agree on the rating. Using the 2% rule on the other hand does not impose this strictly, still giving room for variance within the ratings. One could suspect that the majority would rate conservatively and choose center ratings where a few actually rates the extremes. The outlier removal would then force rating to become centered, but this is not the case as was shown in section C.10 where it is evident that using *OC2* actually increases the variance of the distribution of beta means. The centering of ratings must then lie in the choice of music, or the scales themselves. The choice between outlier criteria is seen here regarding general experimental setups and participants, that all can make errors due to concentration and therefore *OC2* is used.

Participant 9 and 14 are distinguished in the temporal analysis where they use much longer time to rate excerpts than the other participants. This suggests that the temporal analysis could be used as a preliminary measure of participants understanding of the experiment, given that the same participants were also the ones who made consistent errors.

Given the data where outliers were removed, beta distributions were refitted, where it is clear that the *MLE* of the coefficients lie within the confidence intervals of multiple other estimates. This can be due to the selection of the musical excerpts, that they were not chosen to be separated enough in the valence-arousal space. Another issue is the number of participants of only 14, making the estimates poor, thus increasing the confidence intervals. Given these concerns on the data that should be used for the preliminary mathematical model, the data are used.

## 4.7 Conclusion

The main objective of the whole investigation into listening experiments was to develop a method to obtain reliable ratings of emotions expressed in music. A number of considerations were presented prior to testing, that were issues that either should be investigated or dealt with through experimental setup and instructions. Sources of emotional induction were limited using experimental setup and instruction to participants, as well as the general bias that can occur when measuring emotions. Choice of scales and self-report methods were based on limiting any vocabulary bias, to be visual icon images of humans by expression showing different emotions on the valence and arousal scale, using a 9-point scale. The length of excerpts, the appropriateness of the scales and the effect the ordering of excerpt have on ratings was investigated in the first of two pilot experiments *pilot1*. Using 3 excerpts of 30 seconds divided into 4 clips each resulting in 12 clips. These were rated by 24 participants where the ordering was either sequential, (i.e. excerpt 1 2 3 4 5 etc.) or a *WLS* design (e.g. 2 4 1 8 5 etc.) each design being rated by 12 participants. The differences were compared using *2KS* and comparison between variance and mean of ratings. It showed that only for 1 excerpt of the 12, that there was a significant difference between the two ordering methods. Inspection of the difference in variance between the two showed that even though participants were presented with clips in succession the variance of the ratings were not much smaller than when using a *BO*. Therefore a balanced design should be used in future testing. Questionnaire data showed that the length of excerpts was too short so this was extended to 15 seconds in *pilot2* and there was confusion as to the understanding and use of scales, so a verbal explanation was used in *pilot2*.

An investigation into whether both scales could be presented to participants when rating an excerpt was made, using a *2KS*. 8 did not pass *2KS* Null hypothesis for the valence scale and 5 did not for the arousal scale out of total 200 excerpts. Given the fact that the data foundation was rather scarce only comparing 7 ratings future testing has to be done, to test if there is an effect or whether the difference was due to the subjective differences.

A test was made of meta data of participants, e.g. their musical experience, familiarity of the music, if they understood the scales, etc. Beta distributions

were fitted to the experimental data obtained on the two scales of valence and arousal, grouping all participants ratings per excerpt. Outlier criteria were made where two were compared under the fundamental assumption that participants either misunderstood the whole experiment or the use of one of the scales. The other was that participants could become distracted and therefore rate incorrectly. The latter was chosen using a deviation from beta mean method and the result showed that greater variance of data was achieved for arousal. The data obtained for valence showed very little variance in ratings indicating that participants were hesitant to use the whole scale, where post questioning showed that participants used to middle rating as a “do not know” button. This could potentially pose a problem in future modeling.

# Mathematical model

---

In this chapter the models used for the modeling of the emotional content expressed in music will be presented.

## 5.1 Initial considerations

The ratings acquired from the listening experiment described in Chapter 4 should be modeled using the acoustical data acquired by the features described in Chapter 3. Given the ratings on two 9-point ordinal scales, different approaches can be taken in order to construct a model.

- It can be seen as a classification problem where each point of the 81 points is a class. Here the ordinality should be enforced since  $x_1 < x_2 < \dots < x_9$ . The downside to this approach is the amount of data that should be used. All ratings for all participants and acoustical features for each of the excerpts. Using around  $6.510^9$  floating points. The amount of point the scale is divided into could be reduced here the problem is the fact that there is a dominance of rating with the center. Rounding of mean ratings could be made, resulting in only one point per excerpt and not one for all participants reducing with a factor of 14, could be utilized.
- Since a model was constructed on the experimental data in the form of beta distributions, a regression model could be constructed which has an output of a beta distribution, and e.g. *Generalized Linear Model (GLM)*. Another option is to parameterize the beta distribution and use the  $\alpha$  and  $\beta$  coefficients as targets in a normal linear model. Using this approach

a considerable data reduction would be made, from 28 to 4. Further the mathematical model should not both model the mapping from acoustical features and model the ratings themselves, thus separating the two models.

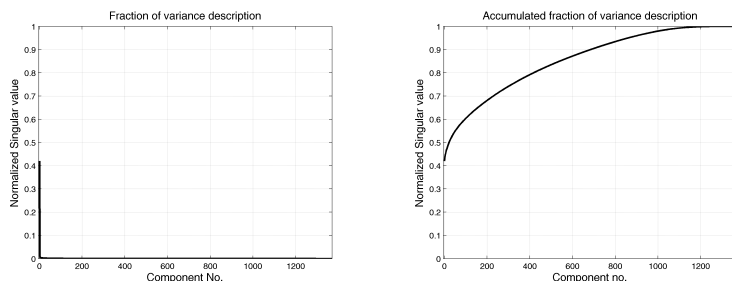
- Another issue to consider is the extensive amount of features compared to the relative few ratings. In the choice of features it was deliberately done so that as many features and even implementation of the same features were computed. At a later stage the features should be selected so that, the features providing the most information, be used for the model. A method of selecting feature should be found or be incorporated into the mathematical model.

## 5.2 Pre-analysis of data

The data obtained for *pilot2*, consisting of 200 songs, is the basis for a preliminary mathematical model.

### 5.2.1 Audio features

The 1373 dimensional feature vector, sampled so that each frame corresponds to 9 ms. This results in a matrix of 330k by 1373. Since it is expected that much of the data is correlated both in time, and between features a preliminary *Principle Component Analysis* (PCA) is made. Due to the sheer amount data a simplified method using Singular Value Decomposition was used.



(a) Fraction of variance using Normal- (b) Accumulated fraction of variance using Normalized Singular Value

Figure 5.1: Principle Component analysis using Singular Value Decomposition, both normalized to the sum of the total sum of eigenvalues.

Given the results on figure 5.1 it is evident that one feature seems to be responsible for the explanation of around 40% of the variance. The following tail increase very slowly meaning that accumulated they account for the data, but since the curve is so flat, a simple *PCA* cannot be performed to reduce the feature dimensions. Another problem in using *PCA* is that eventhough



features have no variance, it can still give a good indication as to what emotional response the given excerpt expresses, e.g. what key the song is written in.

### 5.2.2 Target labels

The initial considerations raised the question what labels should be used as targets for the modeling of the ratings of the participants from the listening experiments. For the sake of reducing the amount of data to use, a parameterization of the beta distributions is used. To include alternative labels, which potentially could be easier to predict than that of the  $\alpha$  and  $\beta$ , the mean and mode of the beta distribution is also used. The reason for choosing these two is that the mode, is the point at which there is maximum probability that people would rate. But this measure would punish the participants that rate in the tail. The mean is more centered, and does not lie at the point of maximum probability, so it could cause a small error for the majority. But would not punish the participants that potentially would lie in the tail of the probability distribution across ratings.

	Mean	Mode
Arousal	$1.1712 \pm 0.2377$	$1.2367 \pm 0.2693$
Valence	$1.0558 \pm 0.2284$	$1.0948 \pm 0.2650$

Table 5.1: A comparison of using mode or mean of beta distributions, using the average distance between each participants ratings and the mean or mode across all 200 excerpts from the *pilot2* experiment. Using the rating interval of [1; 9].

A simple way to see this is to compare all rating given by participants to the mean and mode of the beta distribution. Using the absolute average distance averaged over all rated excerpts. This can be seen in table 5.1. Comparing the mean and standard deviation of the distances of the mean and mode of each distribution, it is evident that using the mean of the beta distribution decreases the distance from all participants to the potential predicted label. This also shows that by using the beta distributions, if predictions should be made for a new participant, in average there would be a deviation of approximately 1-1.2 ratings from that participants actual rating.

## 5.3 Labels and features

Based on prior analysis the labels used in the modeling of the emotions expressed in music are

- $\alpha$  and  $\beta$  coefficients of the beta distribution
- $\mu_\beta$  - is the mean of beta distributions.

The features used are

- CT - 124 features from the Chroma Toolbox
- ISP - 188 features from the Intelligent Sound Project toolbox
- MA - 64 features from the Music Analysis toolbox
- MIR - 510 features from the Music Information Retrieval toolbox
- PSY - 302 features from the PsySound toolbox
- YAAFE - 141 features from the Yet Another Audio Feature Extraction toolbox
- ID - 43 features from the Binaural Cue Selection Toolbox
- *MIN* - 65 features selected based on frequent use in *MIR* which are, MFCC, Chroma, Loudness, Spectral- Decrease, Flatness, Flux, Rolloff, Variation, Center point, Variance, Skewness and Kurtosis.
- *ALL* - 1373 features where features are collected.
- *PCA*<sub>50</sub> - All features projected to 50 principle component vectors.
- *SFS* - 47 features chosen by *Sequential Feature Selection*.

## 5.4 Selection of Model

Given that the parameters  $\alpha$  and  $\beta$ , mean and mode of the beta distributions are used for the modeling, assumptions regarding the model type should be made. The choice of model depends on the assumed distribution of the error and predictions of the given model. The coefficients  $\alpha$  and  $\beta$  might very well be Gaussian distributed and therefore simple linear regression models can be used on these. The mean and mode of the beta distributions are naturally not Gaussian distributed, where the mean being more symmetric than the mode. Nonetheless since no *GLM* with a beta distributed output is found, regression models that assume Gaussian distributions is used, well aware of the potential problems this might result in. Moreover most of the data is not very skewed looking at all beta distributions. Popular regression models include the Least Squares models, which have a single output. Thus any correlation between  $\alpha$  and  $\beta$  is not taken into account, nor any correlation between valence and arousal. Furthermore since it is single output any optimization using error measures, has to be on the single coefficients, which will be discussed further in section 5.6.

## 5.5 Linear regression

Initially a simple *Linear Regression* (LR) model is attempted, the *Ordinary Least Squares* (OLS) estimate is widely used. Let  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$  be vectors of length  $n$ , representing  $n$  samples or observations of  $m$ , and let  $\mathbf{y}$  be a vector of  $n$  labels. Let  $\mathbf{w} = (w_1, w_2, \dots, w_m)'$  be a set of regression coefficients that produces the output  $\mathbf{y}$ . Including a bias term in the model  $w_0$  as a series of ones it gives  $m + 1$  variables. The feature matrix  $\mathbf{X}$  thus has  $n$  rows and  $m + 1$  columns and the target vector of observations is  $n$  long.

For all models used within this work normalization is made on the data so that,

$$\sum_{i=1}^n y_i = 0 \quad \sum_{i=1}^n y_{ij} = 0 \quad \sum_{i=1}^n x_{ij}^2 = 1 \quad \text{for } j = 1, 2, \dots, m. \quad (5.1)$$

The estimate by the linear model can then be described by

$$y_i = w_0 + \sum_{j=1}^m x_{ij} w_j + \epsilon_i \quad (5.2)$$

where  $w$  is the linear regression coefficients including the bias term  $w_0$  which can be seen as an offset for the model.  $\epsilon_i$  denotes the noise or error term for each instance of  $i$ , which is zero mean i.i.d. Gaussian noise. Due to this noise the problem cannot be solved directly, therefore a cost function is defined in order

to assess how well a set of regressors  $\mathbf{w}$  predict  $\mathbf{y}$  from  $\mathbf{X}$ . The Least Squared estimate is used which is a minimization of the sum of squared residual error.

$$\sum_{i=0}^n (y_i - \sum_{j=1}^m x_{ij}x_i)^2 \quad (5.3)$$

where here the bias term  $w_0$  is removed by adding a constant term of e.g. ones as a column to the matrix  $\mathbf{X}$  and letting  $i$  run from 0. We can write this in matrix notation as

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \quad (5.4)$$

We can differentiate (5.4) with respect to  $\mathbf{w}$  and setting the gradient to 0, subsequently equating for  $\mathbf{w}$  we get

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (5.5)$$

which is known as the normal equation of the least squares problem. We can identify  $\mathbf{X}^T\mathbf{X}$  to be the Hessian. Inverting the Hessian can cause numerical instability when solving this. Some methods exist to overcome this issue, one being to use the *Moore – Penrose pseudo – inverse* of the matrix  $\mathbf{X}$  which is given by

$$\mathbf{X}^\dagger \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad (5.6)$$

The normal equation is then given by  $\mathbf{w} = \mathbf{X}^\dagger\mathbf{y}$ .

### 5.5.1 $L_2$ -regularized regression

One thing is the inversion problem, another problem that can occur with an unconstrained linear model is the behavior of the regressors  $\mathbf{w}$ . In [Schmidt, 2005] an example is given, if two variables are highly correlated, this allows one coefficient of  $w_i$  to become e.g. very large in the positive direction where as another grows very large in the negative direction to cancel out the first. This can produce a model with very high variance in the weight-space, which can cause very different results of  $\mathbf{w}$ . Another issue is that with very high variance models, with low amount of observations  $n$  and a large number of variables/features overfitting can occur. The model producing very data specific results, consequently making poor predictions.

The *Tikhonov* regularization addresses the high variance of  $\mathbf{w}$ , by adding a so

called *weight decay* to the cost function.

$$\sum_{i=1}^n (y_i - w_0 - y_p)^2 + \lambda \sum_{j=1}^m w_j^2 \quad (5.7)$$

where for notational simplification the values predicted by the linear model is written as  $y_p = \sum_{j=1}^m x_{ij}w_j$ . The bias term is moved out from  $w$ , since no penalty of the regularization is made on it. Using the notation and arguments from [Schmidt, 2005] we include the bias term into  $\mathbf{w}$  by setting  $w_0$  to the mean of the target values, and use a *centered* target vector. We can then write in matrix notation as

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \quad (5.8)$$

where the  $\lambda$  value is a scalar that regulates the penalty upon the squared 2-norm of the regressors. Since this penalty is on the squared 2-norm this regularization is often also called *L2-regularization*. Similarly to (5.5) we now differentiate with respect to  $\mathbf{w}$ , and get

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y} \quad (5.9)$$

for  $\lambda > 0$ , where  $I$  is the identity matrix. Here it is evident that by applying a penalty on the regressors, we add a positive value to the diagonal of the Hessian, thus improving the numerical stability of the inversion and forcing  $\mathbf{w} \rightarrow 0$ . In *L2-regularization* the  $\lambda$  parameter then acts as a “smoothing” parameter of the regressors. The aim is then to minimize (5.8) while keeping  $\mathbf{w}$  small, not forcing it to become zero and thus enforcing systematic errors by the model.

### 5.5.2 *L1-regularized regression*

*L2-regularization* provides a good means of generalization by imposing a penalty on the regressors adjusted by  $\lambda$  thus reducing the joint 2-norm. But it does not result in a sparse model, meaning a parsimonious model. The aim here is that choosing the features extracted within the model, thus potentially resulting in a better model and more simple than that of the *L2-LS*. A method of doing this is to use a 1-norm penalty in the cost function. The advantage of this is that often these model outperform those of the *L2* regularized models [Schmidt, 2005]. The unconstrained cost function can be written as

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 \quad (5.10)$$

this is a unconstrained convex optimization problem, but this is non-differentiable, when  $w = 0$ . Thus it cannot be solved in a similar manner as with the *L2*

regularized method. Solving this problem by minimizing Least squared error subject to a 1-norm penalty has been attempted by numerous methods, and is often called the Least Absolute Selection and Shrinkage Operator (*LASSO*). Presently one method is used here called the LARS method.

### Least Angel Regression and Selection

Different L1-regularization methods exist, that within the model, features are selected by forcing regressors to become 0, producing a sparse model. This so called 1-norm constraint on the regressors is called the *LASSO*. Different approaches exist to obtain this solution.

A stepwise model is used here called Least Angel Regression and selection method that was described in [Efron et al., 2004] and [Mørup et al., 2008]. A simplified outline of the method will be made here, where the notation is changed to follow previously used notation. The implementation used was done by [Sjöstrand, 2005]. The same starting point as was used in (5.2) is used here,

the cost function is written similar to (5.10) and for convenience it will be denoted  $\mathbf{c}$  here. The gradient of the cost function is then given by

$$\mathbf{g} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \text{sign}(\beta) \quad (5.11)$$

omitting  $\lambda \text{sign}(\hat{\beta})$ , (5.11) is identical to the negative gradient of the unregularized problem as was used from (5.4) to (5.5).

Introducing an active set  $A$  and an inactive set  $I$ , where the initialization consist of an empty active set, and the inactive set consist of all elements in  $\mathbf{X}$ .

Calculating the gradient for all features in  $I$ , adding the element  $j$  with the highest gradient.

$$j = \text{argmax}(|\mathbf{g}_I|) \quad (5.12)$$

Element  $j$  is then added to the active set  $A \leftarrow A \cup j$  and  $I \leftarrow I \setminus j$ , where initially  $\hat{\beta}_A = 0$

$$\mathbf{w}_A = \frac{\mathbf{X}^T \mathbf{y} + \lambda \text{sign}(\mathbf{w}_A)}{\mathbf{X}^T \mathbf{X}} \quad (5.13)$$

where  $\mathbf{X}^T \mathbf{X}$  is the Hessian. To update  $\hat{\beta}_A$  a step of  $\mu$  is taken using a Newton-Raphson step, where the update is calculated as

$$\tilde{\mathbf{w}}_A = \mathbf{w}_A + \mu (\mathbf{X}^T \mathbf{X})_{A,A}^{-1} \text{sign}(\mathbf{g}_A) \quad (5.14)$$

Since the inverse Hessian is given by  $(\mathbf{X}^T \mathbf{X})^{-1}$  the step  $(\mathbf{X}^T \mathbf{X})_{A,A}^{-1} \text{sign}(\mathbf{g}_A)$  will be in the direction such that the amplitude of the gradients in the active set are identical, i.e.  $|\mathbf{c}_{A_1}| = |\mathbf{c}_{A_2}| = \dots = |\mathbf{c}_{A_n}|$ ,  $\mu$  is then calculated using three

different criteria .

- There exist an  $\tilde{\beta}_A = 0$  out of  $q$  possible. Where  $\tilde{\beta}_A$  is the  $\hat{\beta}_A$  at the step of size  $\mu$ . Then element  $A_q$  is removed from the active set,  $I \leftarrow I \cup A_q$  and  $A \leftarrow A \setminus A_q$ .
- There exist an element  $l$  in the inactive set  $I$  where the gradient of that element in the inactive set equals any gradient in the active set  $A$ ,  $|\tilde{c}_l| = |\tilde{c}_A|$ .
- The gradient at the step by  $\mu$ ,  $|\tilde{c}_A| = 0$ .

where  $\hat{\beta}_A = 0$  is the so called LASSO condition for more information see [Mørup et al., 2008].

After taken the step  $\mu$  the process is repeated as long as there exist an element  $j$  in the inactive set  $I$  such that the gradient is greater than zero,  $|\tilde{c}_j| > 0$ . The algorithm then processes all elements in  $X$  one at a time, making the method ideal for problems with a great number of features and observations.

### 5.5.3 Stepwise regression

The stepwise regression model used here is the so called *Forwardstepwise* regression approach. Using *Forward* instead of *Backward* means that initially in the approach used here, an empty model is used. It proceeds to systematically add and remove features from the multilinear model. Each iteration it calculates the statistical significance of adding or removing a feature for regression, using  $F$ -statistic. The  $p$ -value is computed with and without each of the features, if a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. If a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model [Mathworks, 2010]. It follows after fitting the initial model.

1. If any terms not in the model have  $p$ -values less than an entrance tolerance, add the one with the smallest  $p$  value and repeat this step.
2. If any terms in the model have  $p$ -values greater than an exit tolerance ( $p_{remove}$ ), remove the one with the largest  $p$  value and go to step 1, otherwise end.

The approach used here although using the same initial empty model, could produce different results based on the order in which features are added and

excluded. So  $p$ -values should be determined to be the optimum, using e.g. cross validation. Potentially a different sequence of steps could lead to a better fit, thus not guaranteeing that the model reaches a global minima, but just a local minima. Therefore multiple runs are made for each fit, to obtain the best model.

### 5.5.4 Sequential Feature selection

Sequential feature selection is a brute force method of obtaining sparse models. The method used here sequentially evaluates all features that is not within the model and adds the feature that reduces the error the most. Initially an empty model is used, using Least Squares a model is calculated, the  $RMSE$  is then calculated using (5.15) on all features. The feature that produces the smallest error is then chosen.

1. Train Least squares models on all features that are not present in the model.
2. If any features not in the model have a  $RMSE$ , by adding it to the current model, lower than a given tolerance level, go to step 3, otherwise end.
3. The feature, by adding it to the model, decreases the model error the most is added to the model, go to step 1.

The method will potentially find a global minima using the error measure implemented. The problem with this method is that it has to compute a great deal of models and errors, where with datasets with a great deal of dimensions and features, becomes computational very expensive.

## 5.6 Error measures

It is crucial for all model evaluation that an error of performance is found. By the choice of error measure, implicitly the underlying model is chosen. Four different variables is sought to be model, the two coefficients of the beta distributions and the mean and mode or maximum probability of that distribution. The  $\alpha$  and  $\beta$  parameters predicted by the linear regression model, are Gaussian distributed around the true mean, making it symmetrical. Error measures could be made directly on the models ability to predict the coefficients, but this is not that comparable, and does not including the robustness of the beta distribution. As was mentioned previously, since a single output algorithm is used, any optimization cannot be made using any distribution comparing measures, but has to be made directly on the coefficients. This is also the case for the mode and mean of the distributions. The error measure chosen is the *Root Mean Squared Error* (RMSE).



### 5.6.1 Root Mean Squared Error

It is a well known error measure, the root mean squared error that, as the name suggests is calculated as,

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \quad (5.15)$$

where  $y$  is the target and  $\hat{y}$  is the predicted value. The measure is useful within parameter comparisons and model selection. Using this measure on the mean and mode gives a direct measure as to how many ratings the error is across all participants and excerpts, keeping in mind the symmetry of the *RMSE* and the potential asymmetry of the beta distribution.

To compare models and data sets measures should be made on comparing the target and predicted distributions. Three different measures are used here, euclidean distance between means as was used in [Schmidt and Kim, 2010], *Kullback-Leibler* divergence.

### 5.6.2 Euclidean distance between means

Given that two beta distributions for each of the dimensions of valence and arousal should be compared, three different Euclidean distances are made. The distance between target beta mean and predicted beta mean for both scales and is calculated as

$$d(\hat{\mu}_\beta, \mu_\beta) = \sqrt{\left(\frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}\right)^2 - \left(\frac{\alpha}{\alpha + \beta}\right)^2} \quad (5.16)$$

where the  $\hat{\mu}$  and  $\hat{\alpha}$  and  $\hat{\beta}$  denotes the predicted values of the beta mean and the beta coefficients. To produce results comparable to [Schmidt and Kim, 2010] using simple Pythagoras the distance between the means of the predicted beta distributions and the targets of the two dimensions of valence and arousal is calculated as

$$d(\hat{\mu}_{\beta(val,aro)}, \mu_{\beta(val,aro)}) = \sqrt{d(\hat{\mu}_{\beta(val)}, \mu_{\beta(val)})^2 + d(\hat{\mu}_{\beta(aro)}, \mu_{\beta(aro)})^2} \quad (5.17)$$

the measure is somewhat comparable given that the distributions are defined within the same numerical space.

### 5.6.3 Kullback Leibler divergence

*Kullback Leibler* (KL) divergence is a non-symmetric measure of the divergence between two probability distributions  $P$  and  $T$  and is written as

$$D_{KL}(P \parallel T) = \ln \frac{B(\alpha, \beta)}{B(\hat{\alpha}, \hat{\beta})} - (\alpha - \hat{\alpha})\psi(\hat{\alpha}) - (\beta - \hat{\beta})\psi(\hat{\beta}) + (\alpha - \hat{\alpha} + \beta - \hat{\beta})\psi(\hat{\alpha} + \hat{\beta}) \quad (5.18)$$

where  $B$  denotes the *Beta* function and  $\psi$  is the *psi* or *digamma* function. It is non-symmetric so that the distance between  $T$  and  $P$  is not the same as  $P$  and  $T$ . In this case  $P$  is the predicted distribution using only test data and  $T$  is the target distribution, so the distance is from the predicted distribution to the target. The measure is rather difficult to interpret and is therefore seen as a way of comparing models and data sets. In [Schmidt and Kim, 2010] they also compute this, but use it on 2D-Gaussian distributions, so direct comparison can not be made.

### 5.6.4 Baseline error

Given any type of error measure to have something to compare with, given the data used for modeling, a series of baseline errors are proposed. In [Schmidt and Kim, 2010] they suggest within the 50-fold *CV*, to compare the predicted distribution to another randomly selected distribution within the test set, using *KL* divergence. This method is in their work referred to as *Averaged Randomized KL divergence*. By using this method, the overall distribution of the input data space is taken into account. The weakness of this method is that the baseline error measure can change since it is computed randomly.

Another method proposed here is when using the two 9-point emotional scales and the fact that a great deal of the training and test data is centered around the middle point, e.g. 0.5 on the beta transformed axis used for beta distributions or point no. 5 on the scale. For any model the safe bet would then be to predict this value all the time, which is illustrated on figure 5.2.

In principle the absolute averaged distance would then approximate the variation of the training and test data, if the data was completely centered. A proposed model will only be better if it can perform better than the safe bet predictions. The same is the case for comparing the euclidean distance between the mean of the beta distributions.

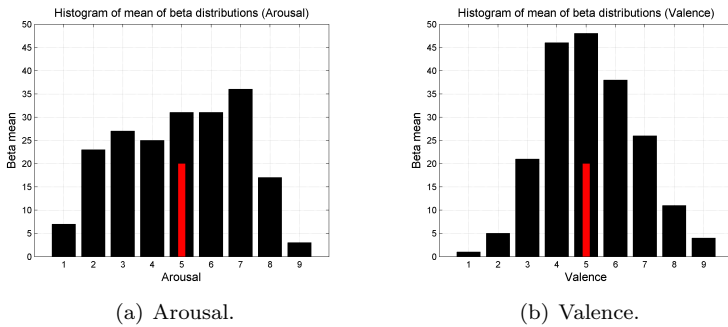


Figure 5.2: Histogram of mean value of beta distributions, the red bar indicates the rating of 5 which should be used as a baseline error measure.

## 5.7 Cross-validation

A problem when choosing parameters for e.g.  $L_2$ -regularized least squares is how to choose the regularization coefficient. Another issue when training linear models is when reporting an error, then how generalizable it is or if it just is representative for that particular chosen training and validation set. For this reason for models that are trained, a *Cross-Validation* (CV) scheme is used.

### 5.7.1 K-fold

To produce generalizable error and to optimize parameters within the linear models *CV* is used. For parameter optimization in e.g.  $L_2$ -regularization or stepwise regression a grid search is made, regularization coefficient or  $p$ -values for the two respectively. Each of these parameters are then traced through using *CV* and the optimal parameters are used. For the final testing a variety of different folds are used in the testing, i.e. 10-50. For convenience the 10-fold cross validation scheme is explained here, but can be scaled to to any amount. Prior to testing the total data set is divided into 10 equal folds, 1-fold is used for testing (red color) after the cross validation. See the top on figure 5.3. The 9 other folds is then used for a 10-fold cross validation, see bottom on figure 5.3.

For each parameter a 10-fold cross validation error is computed, meaning, training the model on 9 of the 10 folds (light blue color), and then validating on the last (green color). This produces 10 different so called validation errors. The mean of the 10 validation errors is then computed, representing the generalized error for that given parameter value. All parameters are then computed and based on the validation error the optimal value is chosen. When the parameter value is chosen, the model is then trained on the whole 9-folds and subsequently tested on the 1-fold that has not been used so far. This representing “unseen”



Figure 5.3: Illustration of cross-validation scheme.(top) all data divided into 10 equal bins. (bottom) remaining 9-folds divided into 10 equal bins.

data. This procedure is then repeated 10 times to obtain a true generalizable test error.

For  $L_2$ -regularization values the regularization coefficient is traced through the interval of  $2^{[-10;10]}$ . The optimum is used for each fold of the 50-fold  $CV$

For stepwise regression the p-values are chosen in 0.05 increments from 0.05 to 1.00, where  $p_{enter} > p_{remove}$ .

### 5.7.2 Temporal issues

Given the resampling of features, a great number of features exist for each obtained rating from users as described in section 4. Using a regular form of cross validation, where each sample included in each fold is chosen randomly (Gaussian), can cause overfitting. The problem lies in the fact that potentially a set of samples from a given excerpt might end up in both test and validation set. The labels are the same, but the features are different. Nonetheless the features are most likely highly correlated, thus giving the wrong indication of performance. For this reason each excerpt and all samples from that given excerpt are in the training or validation set at all times, treating it *blockwise*. Using  $CV$  to generalize model performance using this block based approach limits the folds used for testing if the aim is to have an even amount of excerpts in each fold. If using a 50-fold  $CV$ , there is trained on 196 and predicted 4 excerpt at a time. Thus the  $\lambda$ -regularization optimization has to be done on those 196 excerpt, limiting it to a factorization of this number. So 28-folds is used for grid search of coefficients.

## 5.8 Results

### 5.8.1 Baseline error

On table 5.2 the mean and standard deviation of the baseline errors using *RMSE* are shown for mean and the mode of all beta distributions. It is expected that using this kind of baseline error the mean is lower than when using the mode, simply due to the fact that the mean is a measure that is always close to the center than the mode.

Valence	Arousal
$1.0412 \pm 0.7278$	$1.3643 \pm 0.8218$

Table 5.2: Mean and standard deviation of the baseline *RMSE* for beta mean of distributions for valence and arousal, for the 200 excerpts used in the *pilot2* experiment. Measures are given on the rating scale from 1-9.

To have a baseline when comparing euclidean distances between the mean of beta distributions, distance for valence, arousal is presented in table 5.3 together with a measure using both fitted distributions.

Interval	Valence	Arousal	Valence-Arousal
(0;1)	$0.1157 \pm 0.0809$	$0.1516 \pm 0.0913$	$0.2057 \pm 0.0942$
(1;9)	$1.5412 \pm 0.7278$	$1.8644 \pm 0.8218$	$2.5157 \pm 0.8516$

Table 5.3: Euclidean distance between means of beta distributions, for valence and arousal and the two dimensional distance between the two using methods from section 5.6.2. Distances are given in two different defined spaces, the beta distribution  $[0; 1]$  and the scale space  $[1; 9]$ .

### 5.8.2 Sequential feature selection

The feature selection method of *SFS* was calculated on the entire feature set of 1373 features. Due to time constraints only the first 47 features were found. These can be seen on figure 5.4.

A list of the names and the feature pack they came from can be found in section D.1 in table D.1. It is evident that some features are often selected using this feature selection method, *CENS*, *MFCC*, modulation and envelope based features together with loudness, etc.

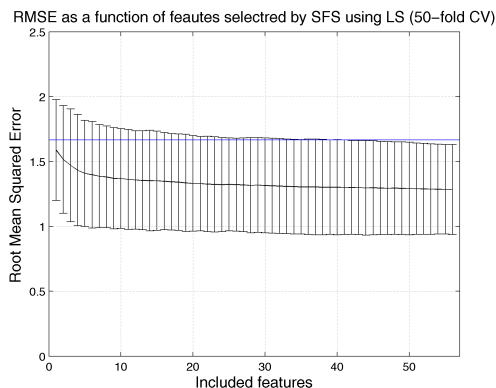


Figure 5.4: *SFS* used on all 1373 features using *OLS* with a 50-fold *CV* scheme. Black curve is the mean of *RMSE* of the 50-folds with error bar of one standard deviation. The blue line indicates the baseline from 5.2. The training was performed on the *mode* of beta distributions of the arousal data.

### 5.8.3 Predictions and errors

An example of predictions using *OLS* on *SFS* are shown in figure 5.5. It is evident that there is a great difference between predictions of emotional labels between excerpts. The variation of the predictions within each excerpt can be seen as an error in the model, since the labels were a constant value. Whether it is an error or it is an effect of the changing emotional expression in music will be dealt with in section 5.9.2. For the time being, this is considered an error by the model. The aim of the model is then to make predictions of emotional labels for each excerpt (i.e. 15 seconds).

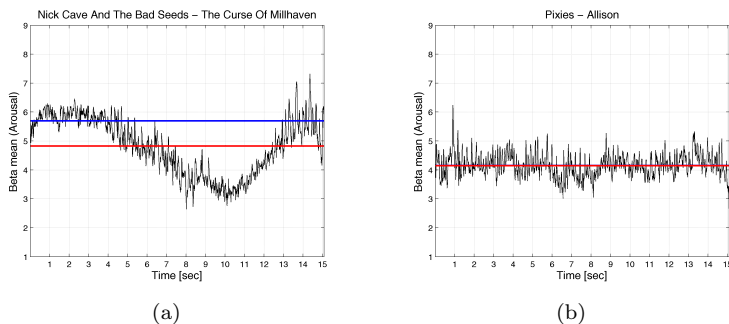


Figure 5.5: Predictions of beta mean for two excerpts using 50-fold *CV* with *OLS* on *SFS* features. Red line indicates the mean across the predictions and the red blue line indicates the label.

The error consist of a varying part (i.e. difference between black and red curve on figure 5.5) and a bias part (i.e. the difference between the red and blue line on figure 5.5). By using the mean of the 1654 predictions (red curve), resulting in one prediction, would then remove the variance and reduce the prediction error. For the sake of predicting emotional ratings for 15 second excerpts this approach is used for all error measure.

#### 5.8.4 *LR* - Beta mean

Here the results of the *LR* models predicting beta mean using both *OLS*, *L1* and *L2* regularized *LS* and *stepwise* regression.

To measure the performance of the 4 different regression models trained on 11 different features selections, predicting the mean of the fitted beta distributions to the emotional data of valence and arousal obtained in *pilot2*, the *RMSE* is used. On figure 5.4 the *RMSE* results for the valence ratings are shown. The baseline error is indicated in the bottom of the table.

Most of the models and selected features have results that lie above the baseline measure. The model that performs best of the four is the *OLS* and the selection that produces the best predictions is using *ALL* features where *MIR* and *SFS* are very close. *stepwise* outperforms *L1* in nearly all cases, where *L1* by *LARS* is the worse performing method.

Feature pack	Valence			
	<i>OLS</i>	<i>L2</i>	( <i>L1</i> by <i>LARS</i> )	<i>stepwise</i>
CM	1.011 ± 0.702	1.030 ± 0.702	1.055 ± 0.708	1.004 ± 0.694
ISP	1.029 ± 0.692	1.030 ± 0.701	1.046 ± 0.719	1.023 ± 0.690
MA	1.027 ± 0.693	1.031 ± 0.702	1.043 ± 0.714	1.023 ± 0.692
MIR	0.908 ± 0.626	0.937 ± 0.638	1.046 ± 0.713	0.934 ± 0.637
PSY	1.007 ± 0.690	1.022 ± 0.700	1.045 ± 0.712	1.021 ± 0.683
YAAFE	0.980 ± 0.653	0.992 ± 0.672	1.043 ± 0.710	0.968 ± 0.650
ID	1.058 ± 0.746	1.046 ± 0.712	1.049 ± 0.716	1.051 ± 0.743
<b><i>ALL</i></b>	<b>0.887 ± 0.625</b>	0.902 ± 0.588	1.030 ± 0.705	0.950 ± 0.607
<i>MIN</i>	1.015 ± 0.702	1.022 ± 0.700	1.047 ± 0.715	1.010 ± 0.701
<i>PCA050</i>	0.997 ± 0.695	1.013 ± 0.701	1.048 ± 0.715	0.991 ± 0.692
<i>SFS</i>	0.945 ± 0.677	0.968 ± 0.672	1.048 ± 0.715	0.938 ± 0.675
Baseline	1.041 ± 0.728			

Table 5.4: *RMSE* for 7 different acoustical feature packs trained using *OLS*, *stepwise* regression, *L1* and *L2* regularized *LS* (*L1* by *LARS*). Values are an average of the predicted test data over 50-fold *CV* on the  $\mu_\beta$  of fitted beta distributions. Training was performed on each channel of the excerpts separately. Each excerpt was kept exclusively in either the test, training or validation data.

Looking at the *RMSE* for beta mean predictions of the arousal scale on table 5.5, the results look much better compared to the valence results. Using *OLS* the best performing selection of features is *SFS* with an error of 0.734 ratings averaged over all excerpts, on a 9-point scale only using 47 features. *ALL* using 1373 features predicts in average 0.760 ratings away from target labels. The worst performing is the *ID* and *PCA*<sub>50</sub> where the *ID* is the only to have a higher *RMSE* than baseline. The *L2*-regularized regression method produces slightly worse results than the *OLS* which is rather surprising when using only test data for error measures. The *SFS* is the highest performing selection of features even performing better than *ALL* features using *L2*.

Feature pack	Arousal			
	<i>OLS</i>	<i>L2</i>	( <i>L1</i> by <i>LARS</i> )	<i>stepwise</i>
CM	1.032 ± 0.653	1.117 ± 0.691	1.377 ± 0.792	1.030 ± 0.651
ISP	0.950 ± 0.657	0.960 ± 0.666	1.284 ± 0.774	0.944 ± 0.654
MA	0.974 ± 0.684	0.982 ± 0.682	1.304 ± 0.811	0.972 ± 0.680
MIR	0.835 ± 0.631	x ± x	1.204 ± 0.740	0.826 ± 0.635
PSY	0.902 ± 0.627	0.938 ± 0.654	0.925 ± 0.633	0.885 ± 0.618
YAAFE	0.867 ± 0.600	0.888 ± 0.617	1.183 ± 0.725	0.879 ± 0.604
ID	1.404 ± 1.088	1.365 ± 0.831	1.373 ± 0.824	1.397 ± 1.088
<i>ALL</i>	0.760 ± 0.598	x ± x	x ± x	x ± x
<i>MIN</i>	0.955 ± 0.668	0.980 ± 0.668	1.282 ± 0.776	0.953 ± 0.662
<i>PCA</i> 050	1.042 ± 0.651	1.061 ± 0.660	1.339 ± 0.797	1.039 ± 0.649
<i>SFS</i>	0.734 ± 0.560	0.732 ± 0.562	1.026 ± 0.668	<b>0.727 ± 0.558</b>
Baseline	1.364 ± 0.822			

Table 5.5: *RMSE* for 7 different acoustical feature packs trained using *OLS*, *stepwise* regression, *L1* and *L2* regularized *LS*. Values are an average of the predicted test data over 50-fold *CV* on the  $\mu_\beta$  of fitted beta distributions. Training was performed on each channel of the excerpts separately. Each excerpt was kept exclusively in either the test, training or validation data. Results marked with x could not be computed due to time issues.

Surprisingly using *L1*-regularized regression it produces results which are inferior compared to *OLS*. The complexity of the model has to be investigated at a later stage to see if this has an effect or influence on the model performance. In one single case does *L1* outperform *L2* and that is when using *PSY*.

The *stepwise* regression method produce in general inferior results compared to the other three methods, except when trained on *SFS* which is the best performing combination of all when prediction beta mean of arousal.

To illustrate the variation in prediction across excerpts and thus the resulting error the *RMSE* is illustrated on figure 5.8. The results are obtained using the predictions of *OLS* on *SFS*. The mean of the *RMSE* across all excerpts



is compared to the baseline error which is equivalent to only prediction center ratings

*Nick Cave - As i sat sadly by her side* with a  $RMSE$  of 3.63 on the arousal scale seem to be the hardest to model where *Roxette Vulnerable* with a  $RMSE$  of 0.02 is one the excerpt the model describes the best.

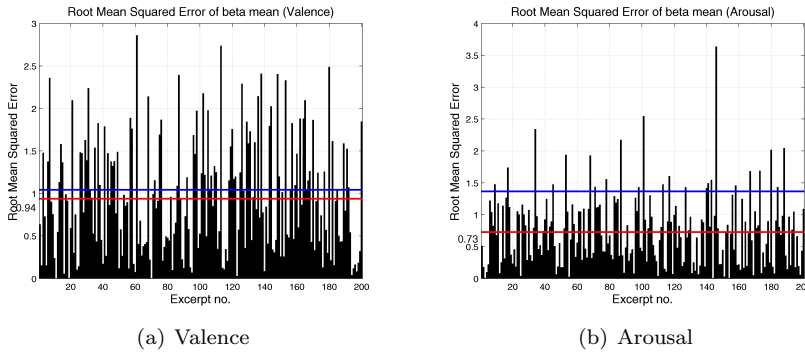


Figure 5.6:  $RMSE$  of predictions by a  $OLS$  model trained using 50-fold  $CV$  on  $SFS$  features (i.e. 4 excerpt predicted per fold). Red line indicates the mean across all excerpts. Blue line indicates the baseline error, calculated as explained in 5.6.4.

To illustrate the general difference between the performance of models used to describe valence and arousal ratings, the distributions of target and predicted beta means for valence ratings is shown on figure 5.7 and arousal in figure 5.8. The distribution of target labels for valence is rather narrow as are the predictions of beta means, when using rounding to fit the 9 bins.

Looking at the distribution of beta means for arousal ratings both target and predictions show a broader distribution.

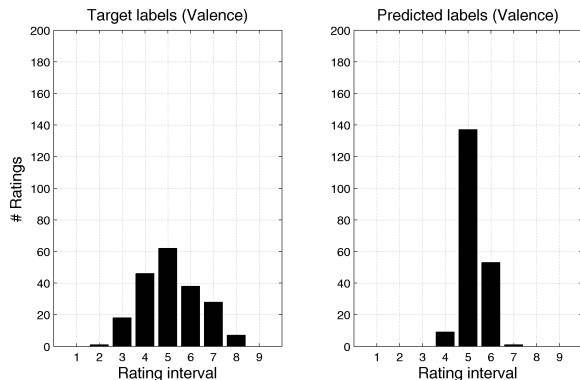


Figure 5.7: Comparison between the distribution of target labels of beta mean for the valence scale and the predicted labels using a *OLS* model trained using 50-fold *CV* on *SFS* features (i.e. 4 excerpt predicted per fold). Labels are rounded to fit 9 bins.

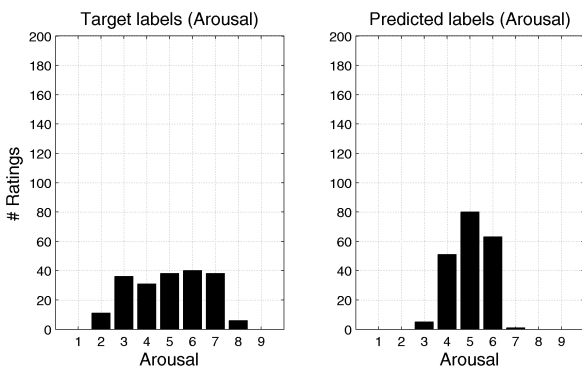


Figure 5.8: Comparison between the distribution of target labels of beta mean for the arousal scale and the predicted labels using a *OLS* model trained using 50-fold *CV* on *SFS* features (i.e. 4 excerpt predicted per fold). Labels are rounded to fit 9 bins.

### 5.8.5 LR - Beta distributions

In this section the results of modeling the distributions fitted on the ratings obtained in *pilot2* on the emotional scales of valence and arousal is presented. The parameterization of the beta distribution  $(\alpha, \beta)$  are the target labels, where *OLS*, *L2*, *L1* and *stepwise* were trained on these using 11 different selection of features.

#### KL divergence

To measure the distance between target and predicted beta distributions the *KL* divergence is used between distributions of arousal and valence separately. These results can be seen on table 5.6 for valence. The baseline error, the so-called *Averaged Randomized KL divergence* was computed for each of the models and selection of features. Given that it is based on random selection of other distributions in the test data the results between each run differ.

Feature pack	<i>OLS</i>	Baseline	<i>L2</i>	Baseline
CT	1.625 ± 2.389	1.854 ± 2.805	0.738 ± 0.777	0.783 ± 0.806
ISP	1.530 ± 2.365	1.741 ± 2.931	0.740 ± 0.780	0.800 ± 0.812
MA	1.471 ± 2.016	1.727 ± 2.281	0.739 ± 0.779	0.766 ± 0.810
MIR	1.358 ± 1.896	1.495 ± 2.443	x ± x	x ± x
PSY	1.330 ± 2.110	1.510 ± 2.302	0.740 ± 0.781	0.784 ± 0.843
YAAFE	1.318 ± 1.860	1.560 ± 2.159	0.728 ± 0.777	0.755 ± 0.826
ID	1.519 ± 2.910	1.716 ± 4.896	0.738 ± 0.782	0.783 ± 0.783
<i>ALL</i>	1.484 ± 2.402	1.746 ± 2.612	x ± x	x ± x
<i>MIN</i>	1.614 ± 2.635	1.812 ± 2.507	0.739 ± 0.784	0.763 ± 0.815
<i>PCA</i> <sub>50</sub>	1.682 ± 2.545	1.759 ± 2.629	0.724 ± 0.774	0.738 ± 0.761
<i>SFS</i>	1.502 ± 2.269	1.597 ± 2.830	0.722 ± 0.770	0.819 ± 0.788

Feature pack	( <i>L1</i> by <i>LARS</i> )	Baseline	<i>stepwise</i>	Baseline
CT	0.956 ± 0.832	0.988 ± 0.858	0.701 ± 0.762	0.821 ± 0.888
ISP	0.784 ± 0.772	0.822 ± 0.648	0.734 ± 0.805	0.666 ± 0.779
MA	0.752 ± 0.782	0.646 ± 0.676	0.732 ± 0.798	0.691 ± 0.723
MIR	x ± x	x ± x	0.642 ± 0.740	0.764 ± 0.813
PSY	0.747 ± 0.724	0.687 ± 0.724	0.791 ± 0.900	0.838 ± 1.020
YAAFE	0.731 ± 0.784	0.702 ± 0.773	0.677 ± 0.726	0.656 ± 0.770
ID	0.743 ± 0.784	0.748 ± 0.826	N/A	N/A
<i>ALL</i>	x ± x	x ± x	x ± x	x ± x
<i>MIN</i>	0.741 ± 0.783	0.699 ± 0.734	0.716 ± 0.779	0.690 ± 0.706
<i>PCA</i> <sub>50</sub>	0.742 ± 0.783	0.747 ± 0.814	0.699 ± 0.749	0.769 ± 0.792
<i>SFS</i>	0.742 ± 0.783	0.813 ± 0.832	0.678 ± 0.815	0.809 ± 0.880

Table 5.6: *KL* divergence between predicted and target beta distributions from the Valence scale. Models were trained using *OLS*, *stepwise* regression, *L1* and *L2* regularized *LS* on 11 different feature selections. Values are an average of the predicted test data over 50-fold *CV*. Training was performed on each channel of the excerpts separately. Each excerpt was kept exclusively in either the test, training or validation data. Results marked with x could not be computed due to time issues.

All models and selection of features achieve a better  $KL$  divergence than the baseline measure for valence. The model that performs best is the *stepwise* regression model, where the best performing set of audio features are the *MIR*. Again the *SFS* performs very close to the larger selection of features. It is observed that the variance of the  $KL$  divergence is rather high indicating that some excerpts are harder to model than others. The worst performing of the four models is the *OLS* even using *ALL* features. The worst performing feature is the  $PCA_{50}$  in combination with *OLS* and using *ALL* features does not improve performance much, where many reduced feature selections perform much better. The  $L1$  model seems to perform very similar on all different feature selections.

Looking at the  $KL$  divergence results for the arousal scale on table 5.7, again *stepwise* is the best performing model when trained on the *SFS* features. The same picture is present here on the  $KL$  divergence as was the case when modeling the beta distributions fitted to the valence data. *OLS* is the worst performing model of the four where number two and three are  $L2$  and  $L1$  respectively. The N/A results indicate that the divergence was not possible to compute due to the fact that prediction were negative and the beta distribution is not defined for negative  $\alpha$  and  $\beta$  coefficients.

To illustrate how the  $KL$  divergence is across all excerpt an illustrations is made on figure 5.9

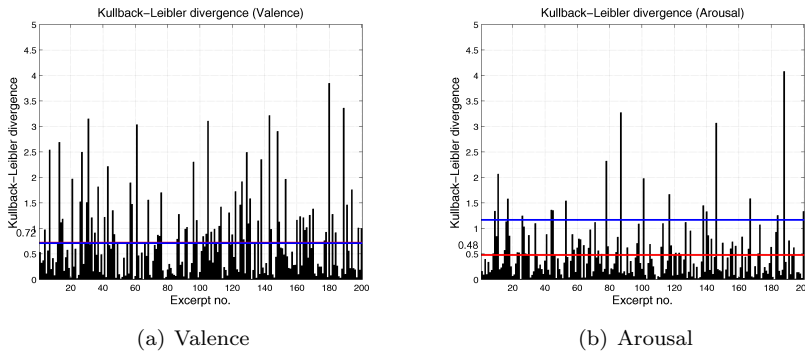


Figure 5.9:  $KL$  divergence between target and predicted beta distributions where  $\alpha$  and  $\beta$  coefficients were predicted by a *OLS* model trained using 50-fold *CV* on *SFS* features (i.e. 4 excerpt predicted per fold). Red line indicates the mean across all excerpts. Blue line indicates the baseline error, calculated as explained in 5.6.4.

Feature pack	<i>OLS</i>	Baseline	<i>L2</i>	Baseline
CT	N/A	N/A	$0.777 \pm 0.803$	$1.040 \pm 1.074$
ISP	$0.953 \pm 1.324$	$1.366 \pm 2.069$	$0.624 \pm 0.692$	$1.137 \pm 1.297$
MA	$0.951 \pm 1.265$	$1.339 \pm 1.612$	$0.640 \pm 0.717$	$1.104 \pm 1.219$
MIR	$0.900 \pm 1.280$	$1.280 \pm 1.601$	x $\pm$ x	x $\pm$ x
PSY	$0.848 \pm 1.169$	$1.207 \pm 1.806$	$0.585 \pm 0.693$	$1.247 \pm 1.443$
YAAFE	$0.875 \pm 1.353$	$1.082 \pm 1.444$	$0.536 \pm 0.623$	$1.315 \pm 1.794$
ID	N/A	N/A	$0.980 \pm 1.040$	$0.958 \pm 0.921$
<i>ALL</i>	$1.530 \pm 7.004$	$3.360 \pm 27.071$	x $\pm$ x	x $\pm$ x
<i>MIN</i>	$0.914 \pm 1.365$	$0.982 \pm 1.539$	$0.634 \pm 0.692$	$1.091 \pm 1.237$
<i>PCA</i> <sub>50</sub>	$0.923 \pm 1.337$	$1.214 \pm 1.716$	$0.694 \pm 0.723$	$0.959 \pm 1.193$
<i>SFS</i>	$0.908 \pm 1.386$	$1.534 \pm 2.918$	$0.480 \pm 0.577$	$1.186 \pm 1.481$

Feature pack	( <i>L1</i> by <i>LARS</i> )	Baseline	<i>stepwise</i>	Baseline
CT	N/A	N/A	N/A	N/A
ISP	$0.982 \pm 1.052$	$0.993 \pm 1.122$	$0.542 \pm 0.668$	$1.181 \pm 1.259$
MA	$0.970 \pm 1.033$	$0.986 \pm 1.131$	$0.570 \pm 0.688$	$1.107 \pm 1.462$
MIR	x $\pm$ x	x $\pm$ x	$0.530 \pm 0.922$	$1.275 \pm 1.591$
PSY	$0.902 \pm 1.101$	$0.988 \pm 1.001$	$0.529 \pm 0.704$	$1.515 \pm 2.236$
YAAFE	N/A	N/A	$0.469 \pm 0.607$	$1.109 \pm 1.288$
ID	$0.985 \pm 1.043$	$0.968 \pm 1.051$	N/A	N/A
<i>ALL</i>	$0.977 \pm 1.033$	$0.989 \pm 1.066$	x $\pm$ x	x $\pm$ x
<i>MIN</i>	$0.977 \pm 1.033$	$0.982 \pm 1.025$	$0.554 \pm 0.683$	$1.118 \pm 1.238$
<i>PCA</i> <sub>50</sub>	$0.982 \pm 1.039$	$0.990 \pm 1.064$	$0.621 \pm 0.657$	$1.172 \pm 1.259$
<i>SFS</i>	$0.981 \pm 1.038$	$0.874 \pm 0.881$	$0.386 \pm 0.523$	$1.511 \pm 1.833$

Table 5.7: *KL* divergence between predicted and target beta distributions from the Arousal scale. Model was trained using *OLS*, *stepwise* regression, *L1* and *L2* regularized *LS* on 11 different feature selection. Values are an average of the predicted test data over 50-fold *CV*. Training was performed on each channel of the excerpts separately. Each excerpt was kept exclusively in either the test, training or validation data. Baseline refers to the Averaged Randomized *KL* divergence as explained in section 5.6.4. Values indicated x due to time they were not computed.

### Euclidean distance

To measure the distance between both the beta distributions of valence and arousal, the distance between their means are computed (see section 5.6.2). The baseline is similar to beta mean predictions baseline, to always predict middle ratings (see section 5.6.4 and 5.8.1). The results can be seen on table 5.8

Feature pack	Euclidean distance Valence-Arousal			
	<i>OLS</i>	<i>L2</i>	( <i>L1</i> by <i>LARS</i> )	<i>stepwise</i>
CT	0.203 ± 0.137	0.177 ± 0.115	x ± x	0.163 ± 0.116
ISP	0.200 ± 0.132	0.164 ± 0.110	0.189 ± 0.121	0.156 ± 0.106
MA	0.202 ± 0.134	0.165 ± 0.110	0.191 ± 0.122	0.159 ± 0.108
MIR	0.194 ± 0.133	x ± x	x ± x	0.144 ± 0.102
PSY	0.185 ± 0.128	0.161 ± 0.108	0.181 ± 0.138	0.156 ± 0.109
YAAFE	0.187 ± 0.132	0.156 ± 0.105	0.193 ± 0.141	0.146 ± 0.099
ID	0.201 ± 0.148	0.191 ± 0.122	0.192 ± 0.122	0.195 ± 0.159
<i>ALL</i>	0.200 ± 0.143	x ± x	x ± x	x ± x
<i>MIN</i>	0.198 ± 0.136	0.165 ± 0.110	0.191 ± 0.122	0.156 ± 0.107
<i>PCA</i> <sub>50</sub>	0.207 ± 0.134	0.169 ± 0.110	0.192 ± 0.122	0.162 ± 0.106
<i>SFS</i>	0.194 ± 0.134	0.151 ± 0.102	0.191 ± 0.122	<b>0.135 ± 0.098</b>
Baseline	0.205 ± 0.094			

Table 5.8: *MED* for 7 different acoustical feature packs trained using *OLS*, *stepwise* regression, *L1* and *L2* regularized *LS*. Values are an average of the predicted test data over 50-fold *CV* on the  $\mu_\beta$  of fitted beta distributions. Training was performed on each channel of the excerpts separately. Each excerpt was kept exclusively in either the test, training or validation data. Results marked with x could not be computed due to time issues.

The best performing model is the *stepwise* using *SFS* features producing an euclidean distance between the two beta distribution of valence and arousal of 0.135. In general most of the models and selection of features are better than the baseline error chosen for this measure. The worst performing model is the *OLS* where again *PCA*<sub>50</sub> is the selection of features that perform the worst.

On figure 5.10 it is illustrate how the euclidean distance between the beta means of valence and arousal is distribution across all excerpts. *Craig David - Hidden Agenda* with 0.351 seem to be the excerpts that the model is the hardest to model, where *Backstreet Boys - No One Else Comes Close* is the track that the model predicts best with an distance of 0.001.

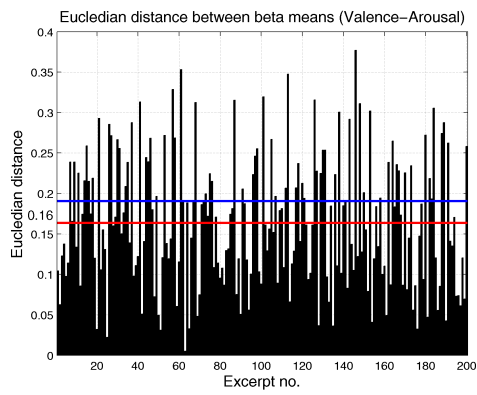


Figure 5.10: Euclidean distance between arousal and valence beta mean using  $\alpha$  and  $\beta$  coefficients predicted by a *OLS* model trained using 50-fold *CV* on *SFS* features. Red line indicates mean across all excerpts. Interval given by  $[0; 1]$ . Blue line indicates the baseline error, calculated as explained in 5.6.4.



## 5.9 Post data analysis

Two different aspects will be investigated in this section and that is the connection between audio features and emotional ratings. The other is the temporal changing emotional predictions by the model investigated in previous sections.

- The comparison of audio features selected by *SFS* and the resulting emotional ratings, to see if there are any general tendencies that if features have a certain value then the resulting ratings are e.g. happy.
- Is there any pattern or structure in the time varying emotional predictions made by the mathematical models. E.g. can these variations be used to group musical excerpts into different categories crossing genre boundaries.

### 5.9.1 Emotional ratings and audio features

To find the relation between audio features and emotional ratings, features selected by *SFS* that are listed in section D.1 in table D.1 are used. The approach is similar to the one used in [Laurier et al., 2009a]. The approach taken is to reduce the dimensionality of each feature vector for each excerpt by averaging it across time samples. For spectrally decomposed features such as loudness the average is taken across all coefficients.

Due to the limited amount of ratings obtained in *pilot2* (i.e. 200), to obtain a more simplified data foundation for the analysis, ratings are divided into four quadrants of the two dimensional emotional model. The division criteria and resulting groups can be seen in section D.4 on figure D.1.

This results in one feature value for the 108 excerpts used in the analysis for each feature. These values are then compared using *boxplots* where each feature is grouped into the corresponding 4 quadrants.

The results can be seen on figures D.2 and D.3 in section D.4. It shows that 5 out of 7 features, *CENS*, inharmonicity, average loudness, average of *MFCCs* and flatness show a good separation of the arousal dimension. For the valence dimension only pulse clarity and tempo show a difference for the Positive-Excited quadrant. Other features were inspected but showed no or similar results to the ones analyzed.

### 5.9.2 Temporal modeling of emotions

The results of the predictions made by the regression approach in the modeling of expressed emotions in music showed a variation in predictions. The label which is compared to using e.g. *RMSE* is a constant value, yet the prediction vary around this value with different amounts of offset. This variation could be seen as an error in the model or be seen as the measure of the time varying emotional

expression of music. Some examples of the predictions were shown in figure 5.5. As was argued in section 2.4 the values predicted are based on post-ratings and therefore are not necessary a measure of the actual time varying emotional expression. Nonetheless the approach taken in [Schmidt and Kim, 2010] was to use a regression model trained on 15 second excerpts and later use this model to predict time varying emotional ratings (i.e. 1 second).

As the models are designed in section 5.5 they predict emotional ratings every 9 ms producing 1654 predictions for each excerpt of 15 seconds. No ratings are given to what the actual time changing emotional ratings are for the 200 15 second excerpts so the only verification of the model is the post-ratings given in *pilot2*. Therefore what is interesting here is not to model the temporal changes in music directly, since no verification is available, but to see what the predictions can be used for. It could be that musical excerpts could be grouped based on the temporal structure, e.g. some songs start of excited and happy and end up being sad and not excited.

The approach taken to see if there is any possibility of grouping excerpts is to

- Collect predictions of beta mean for all 200 excerpts based on *stepwise* on *SFS* (see section 5.8)
- Smooth predictions using moving average filter, subtract mean and down-sample
- Attempt non-supervised machine learning for grouping (e.g. *KNN* or *GMM*)

The reason for smoothing is simply to make the grouping easier. What should be grouped is excerpts based on the temporal structure of emotional predictions, so the offset in ratings is of no concern and the mean is therefore subtracted for each time series. The initial results of the investigation are shown in figure D.4 in section D.5. Where the smoothing of each excerpt is made. Using the smoothed emotional predictions curves, a naive clustering is made by *KNN* using the average of within-cluster sums of point-to-centroid distance averaged over 20 runs to find the optimum  $K$ , and visual inspection of the clusters. The resulting clusterings can be seen on figures D.5 and D.6 in section D.5. Some general tendencies was found where the valence data showed smaller variation in predictions compared to arousal data. Intuitively the results make sense that some excerpt start being very excited and end with a low excitement or build up as the track progresses. But no verification of the results can be made.

## 5.10 Discussion

The modeling of emotions expressed in music was investigated with the use of 4 different linear regression models trained on 11 different selections of features. To parameterize ratings obtained in *pilot2* beta distributions where used where

the coefficients  $\alpha$  and  $\beta$  and the mean of the beta distributions was used as labels of the models. To select features, out of the 1373 dimensions used, that were useful for the modeling, four different methods of selecting features were investigated. *PCA* using 50 dimensions, *SFS* using 47 features, *stepwise* and *L1* by *LARS*. Initially concerns were raised as to the distribution of beta means for the valence data, as it was rather centered with little variance. This observation follows in the performance of all models of beta means for valence. They are slightly worse than the arousal scale, where predictions have a broader distribution than the valence.

In predicting beta means for valence the benefit of using *ALL* features is rather small compared to the use of e.g. *SFS*. The *SFS* is the best of the feature selection methods employed here and the initial concerns regarding the use of *PCA* was confirmed as it was the poorest performing selection of features. The models describing arousal ratings have a very high performance where all models and selection of features lie well below the baseline. *stepwise* on *SFS* is the best performing which is odd combination of two methods used to select features combined produce the best result. In general the best performing models performed 15 % and 47 % better compared to the pessimistic baseline using *RMSE* on beta mean predictions for valence and arousal respectively. Looking at the *RMSE* across all excerpts there seem to be a great variation, where some tracks have a zero error and some have up to a deviation of 4 ratings. The cause of this is argued to be due to the combination of audio features and emotional ratings for those excerpts. A series of different amount of folds used in *CV* show that there is a need for high number of folds e.g. 50 and above, since predicting emotional ratings for e.g. 4 or more excerpts at a time when training on e.g. 196 seems to be difficult.

Comparing the predictions of beta distributions fitted to the experimental data, the same setup as used for beta means was used. Where the coefficients  $\alpha$  and  $\beta$  were models separately and optimized using *RMSE* in a 50-fold *CV* scheme. For both valence and arousal data across all models and selection of features the models performed better than the baseline measure, using *KL* divergence. The fact that the baseline measure is calculated based on randomly selected other distributions in the test data set seems rather odd. Given the distribution of the ratings across excerpts this measure would change. If excerpt were chosen to lie very separated and cover the whole emotional space the baseline would be much worse and thereby the performance of the models would seemingly perform much better. Looking at the best performing model and selection of features using the *KL* divergence again *stepwise* on *SFS* is the best performing combination for both valence and arousal.

To have a measure to compare the total performance of the predictions of valence and arousal ratings for emotions expressed in music, the euclidean distance between both beta means were computed. The best performing combination is again *stepwise* on *SFS* with a euclidean distance of 0.135 equivalent to a 34 % gain in performance compared to baseline. This measure is also

the only measure that can be compared to the results that were obtained in [Schmidt and Kim, 2010]. They achieved a distance of 0.140 using *SVR* trained on *MFCC* audio features. The *KL* divergence cannot be directly compared since their divergence was computed between 2-D Gaussians.

## 5.11 Conclusion

The aim was to create a mathematical model to model the emotions expressed in music. 4 different linear regression models were compared and 11 different selections of features. Predicting beta means of the parameterized beta distribution fitted on experimental data models performed 15 % and 47 % better compared to the pessimistic baseline measure (i.e. rating in the middle) using *RMSE* for valence and arousal respectively. Considering complexity the best performing combination was *stepwise* on *SFS* where 47 features were used, where the *stepwise* across a 50-fold *CV* in some cases removed some features and in others selected all.

Modeling the whole beta distribution was also done by using the coefficients  $\alpha$  and  $\beta$  as labels. The best combination of model and features was again *stepwise* on *SFS*, where using euclidean distance showed a 34 % gain in performance compared to baseline. Comparing results with [Schmidt and Kim, 2010] there was an improvement, they achieved a distance of 0.140 using *SVR* trained on *MFCC* audio features and *stepwise* on *SFS* had a distance of 0.135. The goal of creating a model that can predict the emotions expressed in music has been met. The model achieves better results given any of the baseline measures, both for predictions of the mean of the beta distributions or the entire distribution. Using the features found by *SFS* to model expressed emotion in music, an exploratory investigation was made to see what the tendencies across grouped excerpts of the audio features were. 5 out of 7 features showed a good visual separation of the arousal dimension (*CENS*, inharmonicity, average loudness, average of *MFCC*s and flatness). For the valence dimension pulse clarity and tempo was found to show a difference for Positive-Excited excerpts.

Under the assumption that the variation in predictions of the *stepwise* model were not in fact error but a measure of the changing emotional expression in music modeled using post-ratings, an exploratory investigation was made. Since no data is available for verification it only remains exploratory. These vectors representing each excerpt was grouped using *KNN*, where 4 and 5 groups were found for valence and arousal respectively. Using the average across all excerpts within each cluster a comparison was made in the temporal development of expressed emotions in the excerpts. Some grouping were found to start being very excited and slow down towards the end, where others build up towards the end.

# Conclusion

---

In this section a discussion of the project as a whole will be made and lastly a summary of conclusions and achievements.

## 6.1 Discussion

The aim of the project was to make a model that can predict the emotional content in music. Using a systematic approach an analysis was made into what type of model should be used to represent emotions. Given that a great deal of different models exist, a two dimensional model using the dimensions valence and arousal was chosen. Through the analysis of emotions and their connection to music it was found that the mathematical modeling of emotions in music should be separated into two different models. A model that accounts for the expressed emotion and another that describes the induced emotion, in other words a personalized model. The focus in this work is on the model of expressed emotions. This is very important since this also defines the specific research question, and subsequently the mathematical model to model the expressed emotions by music.

The heart of the project is to extract emotional information from a musical source. A method of choosing musical data was suggested which was to create a simple model based on a well defined experimental setup, using structural information about music to predict emotional ratings from larger musical datasets and use these ratings to sample for future testing. This simple model and the necessary steps taken to create it was used to explore all aspect surrounding the topic.

A self-report method was used to obtain ratings from participants, which rated 200 musical excerpts of 15 seconds duration on two *SAM* scales, representing valence and arousal. Using two pilot experiments exploration was made into experimental setup issues using statistical and quantifiable measures. The possible effect of the presentation order of these scales was investigated and the direction of them was taken into account by balancing the design. The length of musical excerpts, the order they are presented in and the effect it has on ratings were also investigated. Often no thought or concern is made into these aspects or are argued to have no effect on results. The analysis made shows results that it does have an effect of e.g. what order musical excerpts are presented in and the emotional ratings participants provide. Prior measures have been taken into reducing familiarity of music but using simple measures the effect of familiarity and a multitude of other quantifiable measures were investigated. No quantifiable connection was confirmed by any of the meta data gathered to the ratings provided by participants.

All emotional ratings obtained in the listening experiment using the ordinal *SAM* scales were analyzed. In order for modeling of this data and to consider outlier removal a simplification of the data was made. This was done by fitting beta distributions for each excerpt on both the valence and arousal ratings. Since 14 participants participated in the experiments each distribution was fitted to 14 data points using *MLE*. The assumption was that it is actually possible to model the data using these distributions but given the few data points the estimates was not optimal. The estimates would become better if more participants had participated. Using these beta distribution parameters outliers of the experimental data were removed under the idea that participants would rate differently from the majority if participants were distracted or mentally absent. Creating outlier criteria based on empirical data and subsequently applying the method on the experimental data, outliers were removed and new beta distributions were refitted. Measuring how long it takes participants to rate each excerpt showed a correlation to the amount of outliers their ratings contained, showing it could be a good preliminary measure of errors. These beta distributions form the data foundation for the modeling of emotions expressed in music. A compromise is made of avoiding the ordinality of the scales by the modeling of these scales using beta distribution, making an underlying assumption that this is possible.

To simplify the beta distributions the mean was also used as a label, which was chosen over the mode since when using the mean the distance for all participants ratings would be lower compared to the mode, eventhough the mode has the highest probability.

The data foundation to make predictions of the emotional ratings was chosen to be audio features since music as a media always contains a sound signal. It was also argued that lyrics communicate emotions but are not always available and the way lyrics are song would be captured by audio features. The approach in finding audio features was to find as many features as possible and let

feature selection methods find the most suitable features to model emotions expressed in music. Four different feature selection methods were compared *PCA*, *L1*-regularized *LS*, *SFS* and *stepwise* regression. Using *SFS* 47 features were found that were compared with emotional ratings and visually they show a good separation of the arousal dimension and the valence dimension.

The mathematical modeling was approached by choosing simple models and methods of reducing dimensionality of the audio features used. The approach based on the results of the modeling shows to be the right way. Only using e.g. features from one feature pack shows that these do not perform very well compared to features that are selected across all the 7 different feature packs used in this work. *LS*, *L1* using *LARS*, *L2* and *stepwise* were compared all trained using 50-fold *CV* on excerpts which were grouped so that each excerpt was only in test, training or validation set. Initial investigations showed that by using a normal *CV* scheme the performance of all methods were considerably better compared to using the one chosen here, due to the fact that by sampling an excerpt and having samples in training, test and validation would artificially boost performance of methods in a very simple way.

*stepwise* on *SFS* was the best performing combination of model and features, both when measuring predictions of beta means where *RMSE* is used, and beta distributions using *KL* divergence and euclidean distance between beta means. A pessimistic baseline measure was formulated that the safe prediction would be to predict 5 on the 1-9 *SAM* scale. Compared to this measure the best model performed 15 % and 47 % better compared to the baseline.

The concern regarding the distribution of beta means could directly be seen on the performance of predicting these ratings. The target distribution was rather narrow illustrating that participants did not rate using the whole scale. This could be an artifact of the scale in combination with music or the fact that the music was not chosen to fill the whole range of valence. In future experimental testing, sampling of musical excerpts should be attempted to cover the whole valence dimension using these scales to see if the results are consistent or just due to musical selection (i.e. limitation of scales, musical selection or emotional model). It is a considerable problem that participants used the middle ratings as a "do not know" button for the sake of emotional modeling. A button that would indicate that participants are unsure about what to rate could be used or a scale of how sure they are of their rating. Another method to reduce the centering of ratings could be to use a two-dimensional scale. Future work has to be performed to explore these issues.

An analysis into the effect of lossy audio compression and using resampling of audio features for alignment was made on all audio features. No direct connections can be made to the performance of the models due to this but further investigations could be made into the excerpts with the highest error and looking at the musical source data. Using the predictions of the designed mathematical model an investigation of whether or not a grouping of excerpts could be made based on the temporal changing emotional predictions. In other words to see if

musical excerpts can be grouped based on the emotional dynamics, well aware that the predictions of the model are based on post-ratings of musical excerpts. Using non-supervised machine learning principles, 4 and 5 grouping were found for valence and arousal respectively. No data was available for validation of these results and future work could look into what the connection is between the predictions based on post-ratings and continuous measured emotions expressed in music.

## 6.2 Summary

The problems presented in 1.4 has all been investigated throughout this work and solved satisfactory within the scope of this thesis. The development of a method of extracting emotional information from music was achieved using a developed listening experimental design.

An emotional model to represent emotions expressed in music was found using a two dimensional model of valence and arousal.

The extraction of structural data from musical tracks was achieved by using audio features from the spectral, cepstral, temporal domains and musical and perceptual features. Features which describe emotions in music were found by using feature selection approaches, where features such as MFCC, Pulse Clarity, Main Loudness, Pulse Clarity, Spectral Flatness per. band, Inharmonicity and *CENS* were found to be some of the most suitable. The development of a mathematical model using audio features extracted from music to predict emotions expressed in music was achieved. A stepwise regression model trained on features selected by Sequential feature selection method was the best performing combination. Compared to defined baseline measures the model performed 15 % and 47 % better. Resulting in an average error of 0.727 ratings on the arousal scale and valence of 0.887 ratings given that participants rated on a 9 point scale. An approach to cluster musical data based on their emotional dynamics within each excerpt was investigated and can show some tendencies that should be confirmed in the future.



# Analysis

---

All supplementary notes to the analysis section is presented in this appendix.

## A.1 Musical Descriptors

In this section some more detailed descriptions of the musical terms often used by music psychologist is given.

**Pitch** (low-high), can be defined as that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale, which when changed gives rise to a melody (ASA 1960). The concepts of pitch is different from signal to signal, for pure tones it is directly related to the repetition rate of the waveform measured as frequency. Complex tones due to its harmonic structure pitch is related to the fundamental frequency. This is although not a directly measurable variable since pitch is a subjective matter, where the condition of the auditory system and the stimuli can change the perception greatly (from [Moore, 2004]).

**Ambitus** (small-big), is the distance from the highest to the lowest note in music. Thus it is a measure of the spectral dynamical range.

**Register** (low-high), is a measure of tonal position, and is highly correlated with the pitch.

**Harmonics** (few-many), complex tones are build up by a fundamental frequency and a number of overtones. These overtones are also referred to as harmonics since they are a multiplum of the fundamental frequency.

**Harmony**, is in principle the same concepts as harmonics but rather than overtones it is using "‘overnotes’" meaning a complex combination of notes that gives the sensation of something being more harmonious, where frequency is not used but rather pitch since it is a subjective perceptual measure.

**Tonality**, refers to the ordering and systematic hierarchical structure of pitch.

**Brightness** (dull-sharp), refers to the amount of high frequency acoustical energy in music or speech. A given cutoff frequency is given and the degree of brightness is then a measure of the amount of acoustical energy above relative to below this frequency.

**Timbre**, can be defined as, that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar. Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus (ASA 1960).

**Loudness** (soft-loud), is the subjective measure of the perceived sound level of an auditory event measured in *sones*. In the psychoacoustical world this is often confused with the intensity of the sound that is measured in dB, therefore confusion can occur between units. A distinction is made between the specific loudness or loudness level and loudness. The relation is that the first is a frequency specific measure in units of *phone* and last is the total loudness, measured in *sones*. The reason behind using *sones* is the novel idea that a measure is needed to reflect the frequency different hearing in the human auditory system. So the relation between a sound of 1 *sones* and another at 2 *sones* is so that they are perceived double as loud. The measure of intensity of a sound is not enough since the loudness level of that sound is frequency dependent and more complex psycho acoustical models has to be used in order to measure this.

**Roughness** (consonant-dissonant), in music, is the impression of stability and repose (consonance) in relation to the impression of tension or clash (dissonance) experienced by a listener when certain combinations of tones or notes are sounded together. In certain musical styles, movement to and from consonance and dissonance gives shape and a sense of direction, for example, through increases and decreases in harmonic tension [Encyclopedia of Britannica, 2008].

**Tone attack/voice onset** (short-long), this refers to the time it takes for a tone to reach its maximum amplitude / rise time or rate of rise of amplitude for voiced sounds and is often seen in the temporal envelope of a musical track.

**Tempo/Speech rate** (slow-fast), tempo is often measured as the number of beats per minute (BPM), the tempo can be set by a number of instruments and can be computed by methods such as, temporal envelope peak picking, fluctuation patterns and correlation methods. Speech or song rate can be calculated in a number of way, one method is to measure the number of utterances, voiced segments, words, phonemes, etc. per minute.

**Articulation/pauses** (staccato-legato), is measured as the time of a tones onset to the onset of the next tone divided by the time from onset to offset of the same tone. So music is legato when there is no silence between each tone and its said to be played more smoothly, where the opposite is true for staccato. In speech the the silence between each word can be measured either as an amount of time or ratio between time spoken and time with to speech.

**Rhythm/meter/mode**, refers to the repetition of acoustical events, the rhythm can be described by meter which is deviation of the music into rhythmic units. Often it is said that the music is in 4/4 or 3/4 time scale or rhythm.

**Jitter/vibrato** (low-high), refers to the temporal micro changes that are in the formants in vowels. This is also extracted using tracking algorithms similar to formant extraction. Vibrator also refers to microstructural changes in the pitch or loudness of a tone, which can be calculated using same type of algorithms as with jitter.

These are often used in western tonal classical music, where a wide range of instruments from strings, to wooden and brass pipes etc. are used and therefore give a differentiated sound.



## APPENDIX B

# Audio features

---

All supplementary notes to the audio feature section is presented in this appendix.

### B.1 Overview of features

In this section a complete list of all the features that is used is presented. The dimensions of each feature is shown and what feature pack contains what features.

Feature pack	Feature	Dimension
ID	Interaural Time Difference	8
ID	Interaural Level Difference	15
ID	Interaural Coherence	20
Total		43

Table B.1: List of features of the *Binaural Cue Selection* toolbox and the dimensions each consist of.

Feature pack	Feature	Dimension
MA	MFCC	40
MA	Sone	24
Total		64

Table B.2: List of features chosen from the *Music Analysis* toolbox and the dimensions each consist of.

Feature pack	Feature	Dimension
CT	Pitch	88
CT	Chroma	12
CT	CENS	12
CT	CRP	12
Total		124

Table B.3: List of features chosen from the *Chroma* toolbox and the dimensions each consist of.

Feature pack	Feature	Dimension
ISP	Chromagram	12
ISP	chromaIF - Chromagram	12
ISP	Frequency of instantaneous frequency gram	19
ISP	Magnitudes of instantaneous frequency gram	19
ISP	MFCC - Auditory toolbox	30
ISP	MFCC	30
ISP	MFCC - SIG	30
ISP	MFCC - Mike Brookes' Voicebox	30
ISP	Fundamental frequency	1
ISP	Number of harmonics	1
ISP	Spectral bandwidth	1
ISP	Spectral centroid	1
ISP	Spectral flatness	1
ISP	Temporal Voicing	1
Total		188

Table B.4: List of features chosen from the *ISP* toolbox and the dimensions each consist of.

Feature pack	Feature	Dimension
MIR	Envelope (mean, variance, skewness, kurtosis)	4
MIR	Spectral Flux	1
MIR	Spectral Centroid	1
MIR	Cepstral Flux	1
MIR	Cepstral centroid	1
MIR	Cepstral peaks	1
MIR	Spectral RMS	1
MIR	Temporal RMS	1
MIR	Low energy percentage	1
MIR	Fluctuations	15
MIR	Envelope (Klapuri06)	410
MIR	Tempo	1
MIR	Autocorrelation of onset detection curve	1
MIR	-Maximum correlation (Pulseclarity)	1
MIR	-Minimum correlation	1
MIR	-Average of the local maxima	1
MIR	-Entropy of the autocorr. curve	1
MIR	-Tempo related to the highest autocorr.	1
MIR	-Gammatone decomposition	1
MIR	Zero Crossings	1
MIR	Spectral rolloff	1
MIR	MFCC Flux	1
MIR	MFCC	20
MIR	Roughness (Sethares)	1
MIR	Roughness (Vassilakis)	1
MIR	Brightness	1
MIR	Inharmonicity	1
MIR	Fundamental frequency	1
MIR	Mode (Major vs. Minor)	1
MIR	Key (Best)	1
MIR	- Key clarity	2
MIR	- Key strength	2
MIR	Key strength (major and minor)	24
MIR	Tonal centroid	6
MIR	Harmonic Change Detection Function (flux of tonal centroid)	1
Total		510

Table B.5: List of features chosen from the *MIR* toolbox and the dimensions each consist of.

Feature pack	Feature	Dimension
PSY	Tempo	1
PSY	Cepstral 1st movement	1
PSY	Cepstral 2nd movement	1
PSY	Cepstral 3rd movement	1
PSY	Cepstral 4th movement	1
PSY	Cepstral std dev.	1
PSY	Cepstral skewness	1
PSY	Cepstral kurtosis	1
PSY	Puretonality	1
PSY	Multiplicity	1
PSY	Chroma Saliency	12
PSY	Chord likelihood	1
PSY	Pitch	1
PSY	Pitch strength	1
PSY	Loudness level	1
PSY	Specific loudness pattern	73
PSY	Excitation pattern	73
PSY	Loudness	1
PSY	Specific loudness pattern	1
PSY	Sharpness A-weighted	1
PSY	Sharpness Z-weighted	2
PSY	Timbral Width	1
PSY	Volume	1
PSY	Tonal Dissonance (HK)	1
PSY	Tonal Dissonance (S)	1
PSY	Loudness	1
PSY	Main loudness	24
PSY	Specific loudness	47
PSY	Sharpness	1
PSY	Roughness	1
PSY	Specific roughness	47
Total		302

Table B.6: List of features chosen from the *PSY* toolbox and the dimensions each consist of.



Feature pack	Feature	Dimension
YAAFE	Amplitude modulation - Tremolo 4 - 8 Hz	
YAAFE	- Frequency of maximum energy in range	1
YAAFE	- Energy difference between mean energy over all frequencies and energy at max freq.	1
YAAFE	- Energy difference between mean energy over frequencies in range and energy at max freq.	1
YAAFE	- Product of the two first values	1
YAAFE	Amplitude modulation - Grain 10 - 40 Hz	
YAAFE	- Frequency of maximum energy in range	1
YAAFE	- Energy difference between mean energy over all frequencies and max freq.	1
YAAFE	- Energy difference between mean energy over frequencies in range and energy at max freq.	1
YAAFE	- Product of the two first values	1
YAAFE	Complex Domain Onset Detection	1
YAAFE	Energy (RMS)	1
YAAFE	Envelope (mean, variance, skewness, kurtosis)	4
YAAFE	Linear Predictor Coefficients	2
YAAFE	Line Spectral Frequency	10
YAAFE	Loudness (Bark)	24
YAAFE	Octave band signal intensity	9
YAAFE	log of OBSI ratio	8
YAAFE	Perceptual Sharpness (Sharpness of Loudness)	1
YAAFE	Perceptual spread (Spread of Loudness)	1
YAAFE	Spectral Crest Factor Per Band (1/4 oct. band)	19
YAAFE	Perceptual spread (Spread of Loudness)	1
YAAFE	Spectral Decrease	1
YAAFE	Spectral Flatness Per Band (1/4 oct. band)	19
YAAFE	Spectral Flux	1
YAAFE	Spectral Rolloff	1
YAAFE	Spectral Slope	1
YAAFE	Spectral Variation	1
YAAFE	Spectral (mean, variance, skewness, kurtosis)	4
YAAFE	Zero Crossings	1
YAAFE	Temporal (mean, variance, skewness, kurtosis)	4
YAAFE	MFCC	20
Total		141

Table B.7: List of features chosen from the *YAAFE* toolbox and the dimensions each consist of.

## B.2 Effect of MP3 encoding on audio feature extraction

In this section a more thorough investigation of the effect lossy encoding of audio has on the extraction of features based on this data. On figure B.1 the spectra of a wideband white noise signal is shown, where the signal has been encoded using *Lame 3.97* codec at different bitrates using standard settings.

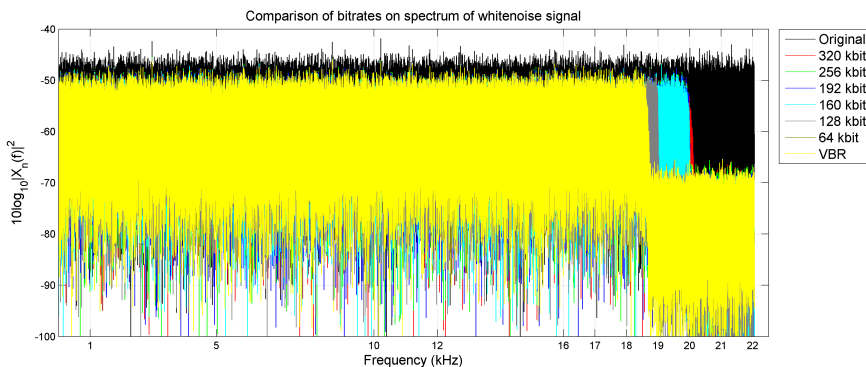


Figure B.1: Comparison of spectra of white noise signal encoded using *Lame 3.97* at different bitrates.

It is evident that the lower the quality is, the lower the lowpass cutoff frequency is used, when looking at the figure. Given that the audio feature extraction of the audio signals used for the experiments are of varying sampling frequencies (22-44.1 kHz) this fact may not contribute significantly to the degradation of the features. Further most of the energy of frequency content of music is not present at above 19 kHz. The other aspect of mp3 encoding comes from applying a psychoacoustical model on a frame and subband basis, thus reducing the quantification based on criteria set by the user and the mp3 encoder. To test the effect this has on the features extracted, 10 different songs of different genre, 30 seconds of duration, all of CD quality is extracted and encoded at 320 kbit, 192 kbit, 128 kbit and a variable bitrate version available from the *Lame* package, set to average 192 kbit.

To determine whether or not a significant change has been made to the audio signal due to the use of encoding, the Pearson's squared correlation coefficient  $r^2$  is used as was done in [Sigurdsson et al., 2006]. Given two signal  $x$  and  $y$  of length  $n$  the correlation is given by

$$r_{xy} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n s_x s_y}, \quad (\text{B.1})$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means,  $s_x$  and  $s_y$  are the sample standard deviations of  $x$  and  $y$ . It is given that an absolute offset in magnitude can be made by the different codec, for this reason the  $r^2$  is used since it will still show a high correlation. Another issue is that when encoding with an mp3 codec, an offset in time can occur, due to headers or other features of the encoder. To eliminate noise caused by the codec time lag, each signal is aligned using cross-correlation between the original signal and those of all the encoded versions.

The interpretation of the  $r^2$  coefficient is that when close to 1 a high correlation is present and when no correlation is present the value drop to 0.

### B.2.1 Results

Due to the great number of features extracted, the multi coefficient features that have above 0.97 of  $r^2$  will not be shown as there are assumed to have perfect correlation. Here multi coefficient refers to features i.e. MFCC that has 20-40 coefficients, or loudness that is divided into many spectral subbands, therefore producing multiple values per feature.

#### MFCC

Here a comparison of all the different implementations of MFCC that is being used, is made.

Given the 7 different MFCC implementation given here, *Pampalk* shows the highest  $r^2$  of up to coefficient 11 and from thereon the implementation of *YAAFE* is the most robust to MP3 encoding, when looking at figure B.2(a). Even at 128 kbit a high correlation is evident, and when using higher bit rates of 192 kbit and 320 kbit, almost perfect correlation is present. Due to the way MFCC is calculated a maximum frequency for each method is given, based in that the number of coefficient is calculated as the number of filters designed. Most of the methods do not surpass a maximum frequency of 16 kHz, thus this is not the reason for the lower correlation, but the psychoacoustical model applied and the subsequent quantification.

#### Interaural differences

The interaural differences here made as a inter-channel difference, only the relevant spectral subbands were chosen and for this reason the  $f_c$  along the abscissa on figure B.3 is different for each feature. A monaural outer ear model is applied to each signal and subsequent some comparison of the two signals across critical subbands. It is evident that the psychoacoustical model that is used for the encoding, and the subsequent quantification of the musical excerpts have an influence of the Interaural differences. The psychoacoustical model reduces the quantization bit allocated for inaudible sounds du to masking, whether or not these inaudible differences can be detected by the binaural processing in the

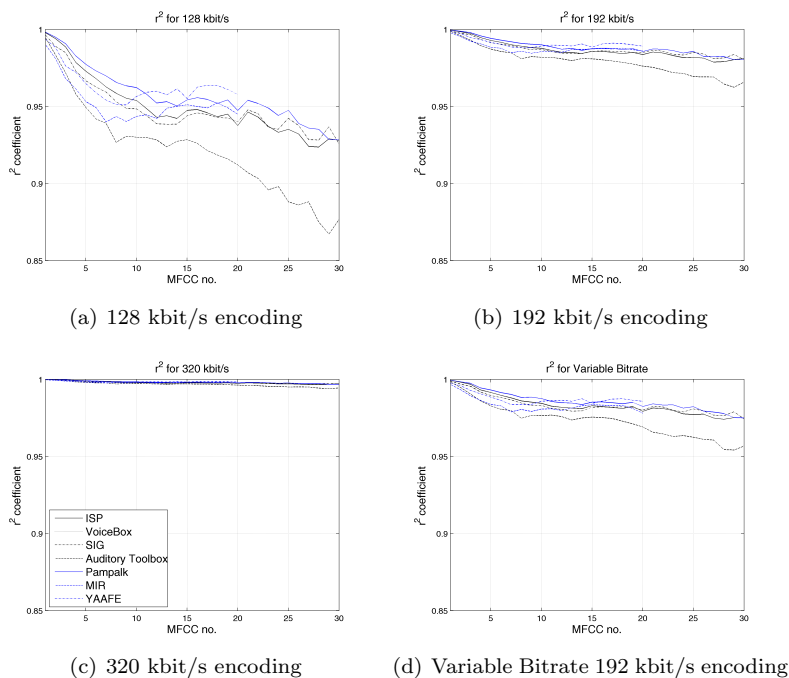
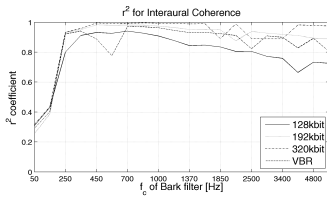
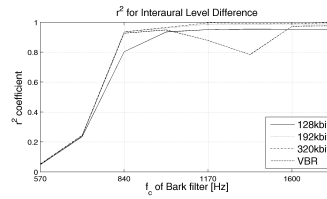


Figure B.2:  $r^2$  correlation of MFCC encoding using different implementations of music encoded at different bitrates. An average over 10 excerpts were made, each of 30 second duration. 4 different *ISP* implementation were used *SIG*, *ISP*, *VoiceBox* refers to Mike Brookes Voicebox and *AuditoryToolbox*. *Pampalk*, *YAAFE* and *MIR* are all compared.

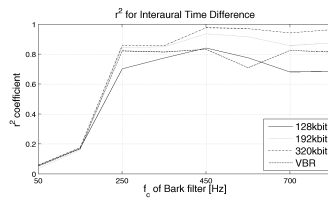
brain is not known.



(a) Interaural Coherence



(b) Interaural Level Difference



(c) Interaural Time Difference

Figure B.3:  $r^2$  correlation of Interaural Differences, the *IC* (a), *ILD* (b) and *ITD* (c). Since the *ITD* and *ILD* was only used in the frequency range where they are physically usable, the center frequency  $f_c$  of the Bark filter for the given Critical Band is used along the abscissa.

Common for *IC* (figure B.3(a)) and *ILD* (figure B.3(b)) is the decrease in correlation at lower frequencies, which cannot be due to any lowpass filtering but rather loss of information due to the quantification.

### Intelligent Sound features

Here the features extracted with the *ISP* toolbox is presented. Using the instantaneous frequency gram, the chromagram was derived, here it should be said that the resulting matrix of the instantaneous frequency gram is sparse. Giving a presence feature, where it is a numerical value when it can be computed, else it is 0. For this reason the  $r^2$  can be somewhat misleading since whether or not a presence can be computed will have a large effect on the correlation.

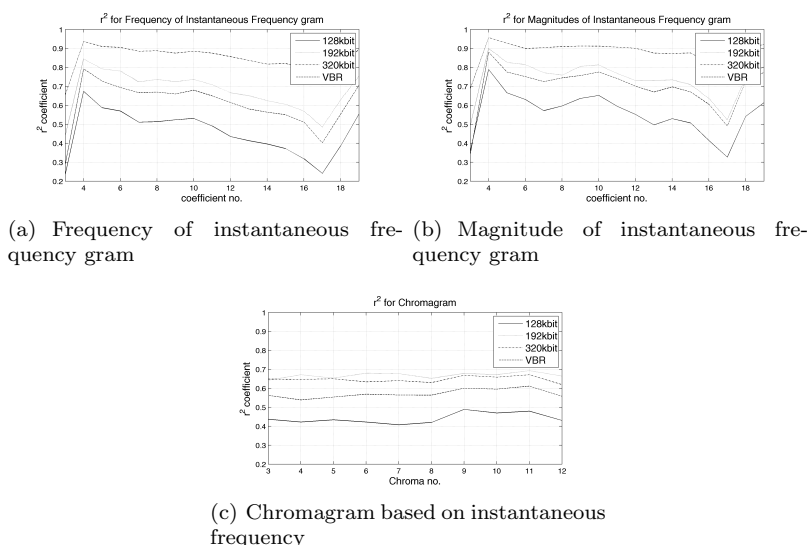


Figure B.4:  $r^2$  correlation of the Chromagram and instantaneous frequency gram.

All single coefficient dimensions from the *ISP* are presented in table B.8.

Feature name	Feature pack	128 kbit	192 kbit	320 kbit	VBR
Temporal Voicing	ISP	0.9892	0.9982	0.9996	0.9970
Fundamental Freq.	ISP	0.7422	0.8639	0.9634	0.9041
no. of Overtones	ISP	0.8479	0.9282	0.9738	0.9195
Pitch presence	ISP	0.4798	0.4441	0.5640	0.3792
Spectral BW	ISP	0.9922	0.9959	0.9965	0.9980
Spectral Centroid	ISP	0.9937	0.9979	0.9988	0.9983
Spectral Flatness	ISP	0.8838	0.8989	0.9017	0.9655

Table B.8:  $r^2$  correlation of features in the *ISP* toolbox, comparing different bit rates of encoding with Lame MP3 codec.

The pitch presence feature was the feature in the test that gave the greatest variety across the 10 musical excerpts chosen. For some songs the correlation

was 0.9 and others 0, giving a skewed image of the feature. The rest of the features in table B.8 have a high correlation even at low bit rates.

### Chroma toolbox features

Here three of the features from the toolbox dubbed *CT* here is presented on figure B.5, common for the other features is that a very high correlation exist and therefore the features are not shown.

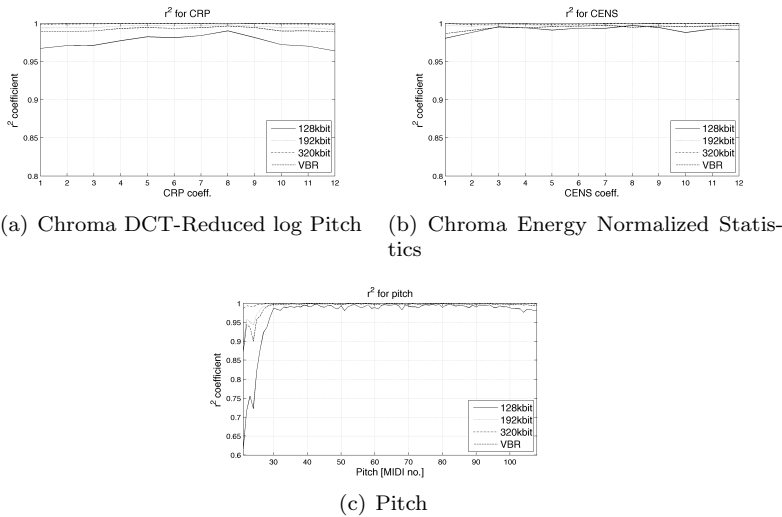


Figure B.5:  $r^2$  correlation of the statistical abstraction of Chroma features from *CT* and the pitch.

It seems common for the these spectral, features that they are very robust to MP3 encoding, where the pitch feature shown on figure B.5(c) is used to derive the other features.

### Music Information Retrieval features

A great number of features are available in the *MIRtoolbox*, where a subset of these are used in this thesis. The only multi coefficient feature that had a poorer than average correlation was that of the sound onset detection made by Klapuri and implemented in the toolbox.

The 128 *kbit*, 192 *kbit* and 320 *kbit* drop off at subband 300 whereas the *VBR* drops off at 350, although as shown on figure B.1 it has a lower cutoff frequency.

All single coefficient dimensions from the *MIR* are presented in table B.9.

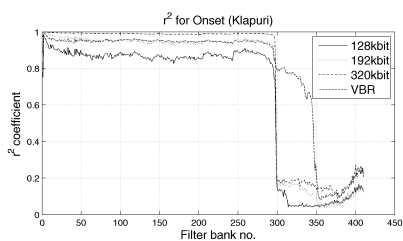


Figure B.6:  $r^2$  correlation for the onset curve decomposed into subbands.

Feature name	Feature pack	128 kbit	192 kbit	320 kbit	VBR
Zero Crossing Rate	MIR	0.9707	0.9845	0.9919	0.9839
Brightness	MIR	0.9938	0.9975	0.9982	0.9983
Cepstral centroid	MIR	0.4677	0.5730	0.6259	0.5693
Cepstral flux	MIR	0.2603	0.3548	0.4302	0.3570
Envelope kurtosis	MIR	0.5165	0.9100	0.9674	0.9087
Envelope mean	MIR	0.8782	0.9677	0.9854	0.9724
Envelope skewness	MIR	0.5232	0.9124	0.9683	0.9111
Envelope variance	MIR	0.7428	0.9542	0.9829	0.9561
HCDF	MIR	0.7588	0.9371	0.9899	0.9016
Key (Best)	MIR	0.5921	0.7888	0.9122	0.7479
Key clarity	MIR	0.8824	0.9702	0.9952	0.9566
Key strength (BK)	MIR	0.5921	0.7888	0.9122	0.7479
Low energy %	MIR	0.9676	0.9886	0.9947	0.9820
Mode	MIR	0.9375	0.9874	0.9979	0.9774
Roughness (Set)	MIR	0.8334	0.9331	0.9583	0.9220
Roughness (Vas)	MIR	0.7754	0.8978	0.9377	0.8812
Spectral centroid	MIR	0.9813	0.9834	0.9838	0.9952
Spectral flux	MIR	0.9904	0.9986	0.9997	0.9979
Spectral RMS	MIR	0.9955	0.9994	0.9998	0.9990
Spectral rolloff	MIR	0.9729	0.9766	0.9769	0.9939
Auto corr ODF	MIR	0.7439	0.6763	0.9843	0.6919
Tempo	MIR	0.7676	0.6774	0.9875	0.6851
Temporal RMS	MIR	0.9969	0.9995	0.9999	0.9994

Table B.9:  $r^2$  correlation of features in the *MIR* toolbox, comparing different bit rates of encoding with Lame MP3 codec.



Based on the results in table B.9 the cepstral features derived are not very robust to encoding even at very high bit rates. Similarly the Autocorrelation of the Onset Detection function and the Tempo, which is derived from the autocorrelation is also seen not to be so robust. Although at 320 kbit near perfect correlation is present.

### Psychoacoustical toolbox features

The multi coefficient features that had the lowest  $r^2$  are shown on figure B.7, where all others had a near perfect correlation of above 0.97.

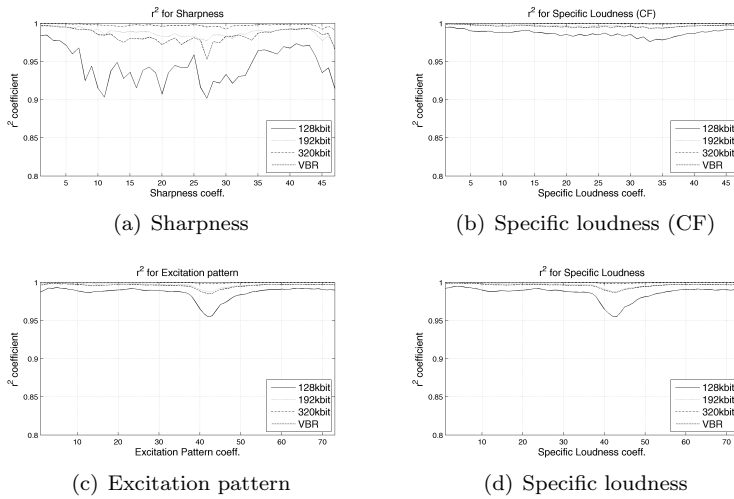


Figure B.7:  $r^2$  correlation features from the *PSY* toolbox.

The only feature of the four shown on figure B.7 that of the sharpness, is the one that stands out, with a slight lower correlation for 128 kbit. In table B.10 all single coefficient features from the *PSY* toolbox are shown. Here as was seen for features from the *MIR* toolbox, cepstral coefficients are not very robust to encoding with MP3. All features based on the cepstrum suffer, even at 320 kbit encoding, with a  $r^2$  of around 0.3 and below as seen on table B.10. Only other feature that suffers is that of Tonal Dissonance (HK) with correlation below 0.2.

Feature name	Feature pack	128 kbit	192 kbit	320 kbit	VBR
Roughness	PSY	0.9905	0.9982	0.9997	0.9973
Cepstral 1st mov	PSY	0.3255	0.3758	0.4069	0.5550
Cepstral 2nd mov	PSY	0.2520	0.2941	0.3485	0.4949
Cepstral 3rd mov	PSY	0.2670	0.3107	0.3624	0.5041
Cepstral 4th mov	PSY	0.2669	0.3127	0.3669	0.4978
Cepstral kurtosis	PSY	0.5581	0.5942	0.5895	0.7617
Cepstral skewness	PSY	0.4824	0.5260	0.5281	0.7190
Cepstral std	PSY	0.3320	0.3737	0.4085	0.5767
Loudness level (CF)	PSY	0.9990	0.9998	1.0000	0.9997
Loudness (MG)	PSY	0.9968	0.9993	0.9999	0.9991
Spectral Disso. (HK)	PSY	0.9777	0.9952	0.9993	0.9935
Spectral Disso. (S)	PSY	0.9596	0.9903	0.9985	0.9860
Sharpness (A)	PSY	0.9966	0.9992	0.9999	0.9990
Sharpness (Z)	PSY	0.9962	0.9991	0.9999	0.9989
Tonal Dissonance (HK)	PSY	0.0812	0.1872	0.4788	0.1982
Tonal Dissonance (S)	PSY	0.4608	0.6209	0.8127	0.6130
Timbral Width	PSY	0.9796	0.9956	0.9992	0.9934
Volume	PSY	0.9936	0.9988	0.9998	0.9984
Pitch strength	PSY	0.9827	0.9979	0.9997	0.9965
Pitch	PSY	0.6883	0.8271	0.9788	0.8482

Table B.10:  $r^2$  correlation of features in the *PSY* toolbox, comparing different bit rates of encoding with Lame MP3 codec.

### Yet Another Audi Features Extraction Toolbox

The multi coefficient features that did not have a  $r^2$  correlation of less than 0.97 across all coefficients from the *YAAFE* toolbox are shown in figure B.8. The only feature that stands out is that of the Spectral Crest Factor per Band, or the peak-to-average ratio as it is called. By removing energy from the acoustical signal based on a psychoacoustical model in the encoder, the peak and rms value of the spectra would change within each subband and naturally change the Crest factor, thereby causing lower correlation.

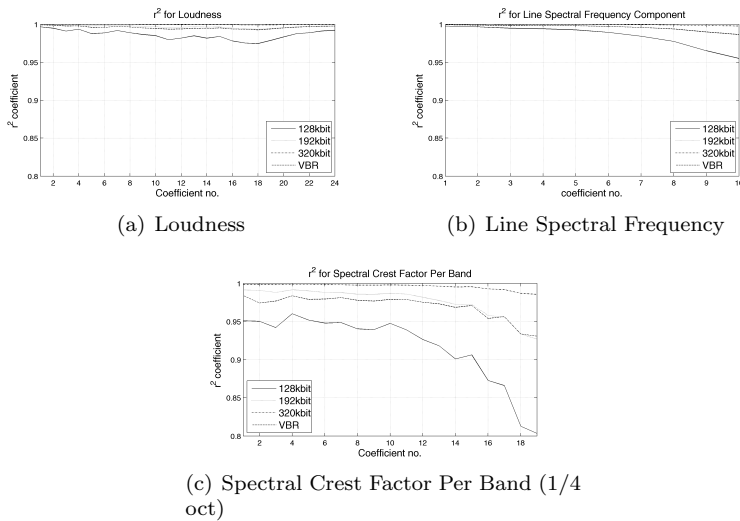


Figure B.8:  $r^2$  correlation of the features in the *YAAFE* toolbox.

All the single coefficient features of the *YAAFE* package are shown on table B.11. In general the features chosen by the *YAAFE* authors are very robust across all features, where all features shows have perfect correlation, even at low bitrates.

Feature name	Feature pack	128 kbit	192 kbit	320 kbit	VBR
Comp. Dom. Onset Det.	YAAFE	0.9962	0.9994	0.9999	0.9992
Energy (RMS)	YAAFE	0.9980	0.9998	1.0000	0.9996
Env. shape stat mean	YAAFE	0.9998	0.9999	0.9964	1.0000
Env. shape stat var	YAAFE	1.0000	1.0000	0.9996	1.0000
Env. shape stat kur	YAAFE	1.0000	1.0000	1.0000	1.0000
Env. shape stat skew	YAAFE	1.0000	1.0000	0.9990	1.0000
Perceptual Sharpness	YAAFE	0.9972	0.9994	0.9999	0.9992
Perceptual spread	YAAFE	0.9958	0.9994	0.9999	0.9989
Spectral Decrease	YAAFE	0.9025	0.9873	0.9979	0.9757
Spectral Flatness	YAAFE	0.9914	0.9974	0.9995	0.9974
Spectral Flux	YAAFE	0.9945	0.9993	0.9999	0.9987
Spectral Rolloff	YAAFE	0.9910	0.9975	0.9995	0.9970
Spectral Slope	YAAFE	0.9971	0.9994	0.9999	0.9991
Spec. Shape Stat. mean	YAAFE	0.9971	0.9949	0.9961	0.9963
Spec. Shape Stat. var	YAAFE	0.9994	0.9987	0.9991	0.9992
Spec. Shape Stat. kurt	YAAFE	0.9999	0.9998	0.9999	0.9999
Spec. Shape Stat. skew	YAAFE	0.9991	0.9985	0.9989	0.9989
Spectral Variation	YAAFE	0.9909	0.9988	0.9998	0.9979
Temp. Shape Stat. mean	YAAFE	0.9915	0.9863	0.9898	0.9996
Temp. Shape Stat. var	YAAFE	0.9989	0.9983	0.9987	0.9999
Temp. Shape Stat. kurt	YAAFE	0.9998	0.9997	0.9998	1.0000
Temp. Shape Stat. skew	YAAFE	0.9978	0.9964	0.9974	0.9998
Zero Crossings	YAAFE	0.9794	0.9904	0.9961	0.9890

Table B.11:  $r^2$  correlation of features in the *YAAFE* toolbox, comparing different bit rates of encoding with Lame MP3 codec.

**Music Analysis Toolbox** On figure B.9 the correlation for the *MA* toolbox is presented, where again we see that Chroma and Sonograms are very robust to encoding. To display the fluctuation pattern, the matrix across modulation channels and frequency channels has been stretched out to a single vector for each time frame of the excerpts. This is seen on figure B.9(a) which is the  $r^2$  for the entire matrix. Given that information across both modulation spectra and spectrum is calculated, some degradation of the information is expected. As with all previous the 320 kbit encoding ensures almost perfect correlation across all coefficients.

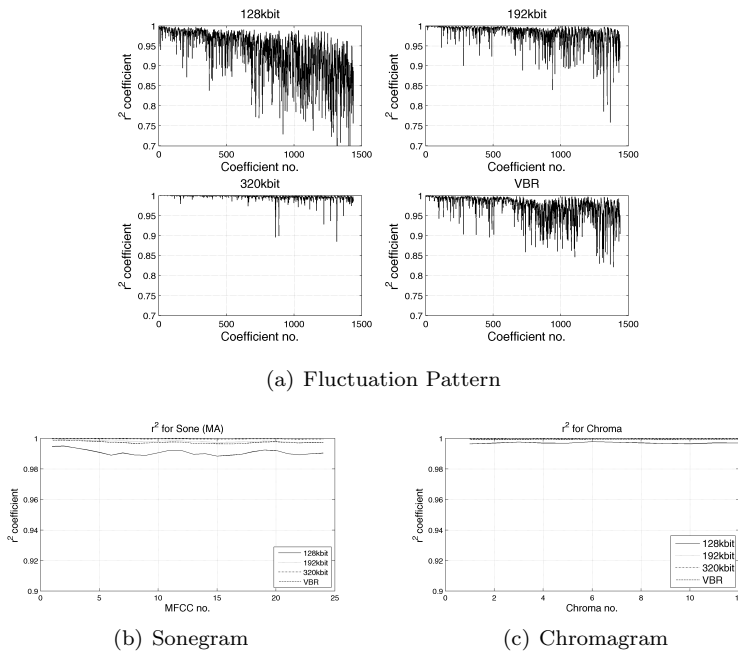


Figure B.9:  $r^2$  correlation of the features in the *MA* toolbox.

## B.2.2 Discussion

Common across most features is that 128 kbit encoding is the worst of the four compared, where the 192 kbit, VBR follow and last the best is the 320 kbit which is expected. Where features have a low correlation due to encoding it is a common aspect across the three lowest bit rates, where 192 kbit is of all of them, the one that rates the best of the rest. Based on the results 192 kbit and above is acceptable for the musical data. If features that have a low robustness to encoding get selected by a feature selection algorithm, caution has to be taken.

### B.3 Effect of resampling on audio features

The problem of aligning the features in order to make a mathematical model is attempted by simply resampling the features. By upsampling the potential error lies in the interpolation method used. Another issue is that the features are calculated in a frame based manner, by interpolating between each frame data is created under the assumption that there is e.g. a linear relation between each frame, if one uses linear interpolation. This might not be the case, in fact since most algorithms only function with a certain amount of data per frame, due to e.g. change of a transient or change in spectra, there is no way of checking what exists in the underlying structure.

The other issue is when downsampling features, high frequency components will be removed. A spectral analysis was made of each feature in order to see what effect the downsampling had, but by only visual inspection no conclusive results can be made. For this reason a correlation analysis is made to provide an indication of the potential error sources in a later mathematical model. It is obvious that even though an error is found resulting from the resampling, this does not mean that the resulting mathematical model will produce the same error in the output. The effect can potentially be negligible, or prove to be substantial. Nonetheless resampling is used although errors can be introduced due to this method.

Above is a list of the features output for the different feature extractors, meaning that each feature produces an output at e.g. 23 ms intervals etc. It has to be emphasized here that, this is not the direct frame size for the features. One feature can have a frame length of 1 second but have an overlap/hopsize of 10 %. over 2 seconds this feature would output 11 feature samples if the initialization is included. A feature with a frame size of 500 ms with an overlap of 50 % would produce only produce 7 feature samples. So the numbers in table B.12 are the time it takes between each output of the feature extractor.

The issue is finding a sampling frequency that is appropriate. On one hand it should not degrade the data too much, and on the other hand to decrease the amount of data for the mathematical modeling, as there is a limited amount of memory and computational power available. Initially the highest sampling of 885 Hz with a temporal frame size of 1.13 ms was attempted but was not possible, so integer multiples of this of 2 4 8 and 16 was attempted. The minimum was found to be a factor of 8 resulting in a sampling frequency of 110 Hz producing frames of 9 ms. Given excerpts of 15 seconds of duration produces 1654 frames per excerpt. This frame size was then investigated further.

To test the effect of resampling on the features, 10 different excerpts were used where all features that had a higher sampling frequency than 110 Hz were tested.

No. features	Avg. time per. output [ms]
4	1.13
4	1.99
3	2.90
1	5.03
26	5.80
1	7.94
5	9.99
27	11.63
3	19.95
1	20.11
20	23.22
1	81.08
3	99.34
2	117.19
2	182.93
1	312.50
2	357.14
2	375.00
2	714.29
7	750.00
1	937.50

Table B.12: Number of features that produce an output at intervals given in milliseconds. Results are obtained using 15 second excerpts.

The procedure was that all features were computed, then downsampling of the features and subsequently upsampling to the original sampling frequency. The resulting features were compared with the original, using Pearsons squared correlation  $r^2$  as described in section B.2 equation (B.1). Thus the comparison is both of the downsampling and upsampling in one coefficient, the downsampling is assumed here to be the biggest contributor to the error. Upsampling is not tested here, as the effect of the downsampling is seen as the major source of error and using a 100th order polyphase resampling method of *Matlab* it is seen as being negligible. Thus aliasing and other artifacts are not an issue for concern.

### B.3.1 Results

All features that have a higher sampling frequency than the chosen of 110 Hz producing 1654 sampled per 15 second excerpt, and the corresponding correlation coefficients are shown here.

Feature name	$r^2$
Temporal Voicing	0.9385
Fundamental Frequency	0.7187
Fundamental Frequency Order	0.8342
Temporal Voicing	0.9385
Pitch estimate	0.6693
Spectral Bandwidth	0.9702
Spectral Center	0.9765
Spectral Flatness	0.8280
Cepstral 1st Movement	0.9702
Cepstral 2nd Movement	0.9200
Cepstral 3rd Movement	0.8913
Cepstral 4th Movement	0.8732
Cepstral Kurtosis	0.9088
Cepstral Skewness	0.9307
Loudness (CF)	0.9975
Sharpness	0.9606
Loudness (MG)	0.9841
Sharpness (A, MG)	0.9915
Sharpness (Z, MG)	0.9823
Spectral Dissonance (HK)	0.9500
Spectral Dissonance (S)	0.9203
Timbral Width	0.7559
Tonal Dissonance (HK)	0.6982
Tonal Dissonance (S)	0.8081
Volume (S)	0.9574

Table B.13:  $r^2$  coefficients for resampled features



In table B.13  $r^2$  results of features that were downsampled are shown. No features show a serious degradation due to the resampling, where most are above a correlation of 0.9. *Tonal Dissonance (HK)*, *Fundamental Frequency*, *Pitch estimate* and *Timbral Width* are the features that suffer the most but are still not unusable.

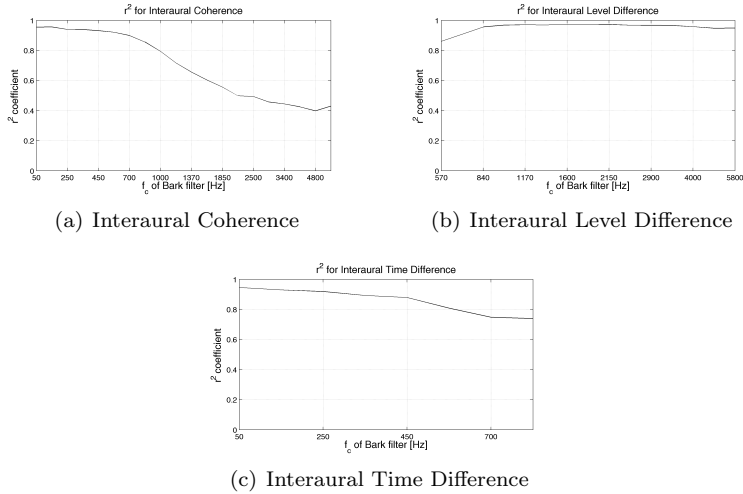
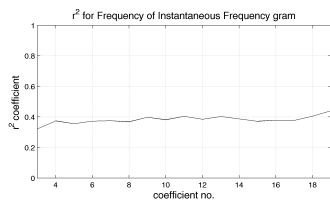


Figure B.10:  $r^2$  correlation of Interaural Differences, the *IC* (a), *ILD* (b) and *ITD* (c). Since the *ITD* and *ILD* was only used in the frequency range where they are physically usable, the center frequency  $f_c$  of the Bark filter for the given Critical Band is used along the abscissa.

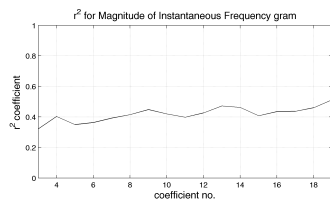
Inspecting figure B.10 the *ILD* does not suffer from the resampling, indicating that the feature does not change rapidly. On figure (a) we see that the *IC* indeed suffers greatly from the resampling, specially in the high frequency coefficients of the *IC* on the *Bark* scale, obtaining a  $r^2$  value of 0.4. The *ITD* also suffers in the coefficients computed for the high frequency content of the audio signal. Implying that the features also contain high frequency content.

On figure B.11 common for all four features is that they do suffer from the resampling a great deal where most obtain a correlation of around 0.4 over all coefficients. Across musical excerpts it is evident that these features contain a great deal of high frequency components across all audio frequency bands.

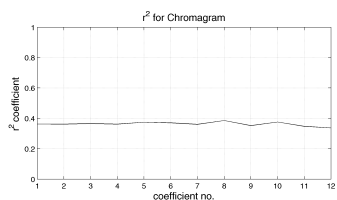
On figure B.12 again a pattern is seen of degradation of the features, for figure B.12(b), B.12(c) and B.12(d), where specially in the low audio frequency area, a great deal of high features frequency content is present and thus suffers from



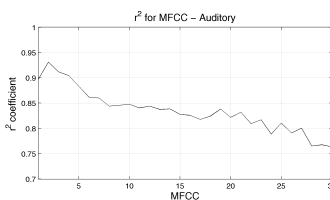
(a) Frequency of instantaneous frequency gram



(b) Magnitude of instantaneous frequency gram



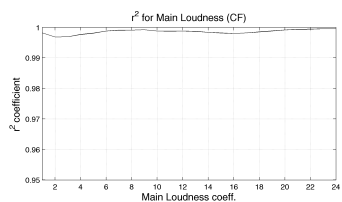
(c) Chromagram based on instantaneous frequency



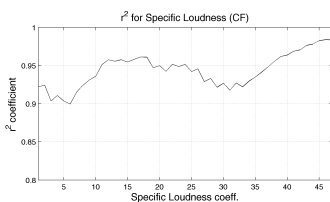
(d) MFCC - Auditory

Figure B.11:  $r^2$  correlation of the Chromagram, instantaneous frequency gram and MFCC.

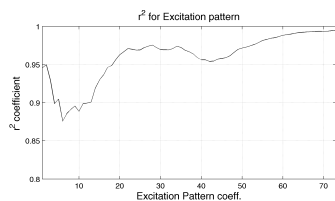
the downsampling. On figure B.12(d) for the Main loudness, still a degradation is seen, not as substantial as others.



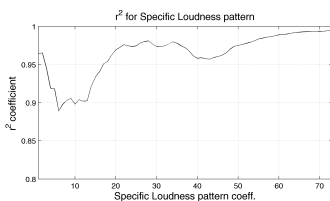
(a) Main Loudness



(b) Specific loudness (CF)



(c) Excitation pattern



(d) Specific loudness

Figure B.12:  $r^2$  correlation of features from the *PSY* toolbox.

### B.3.2 Discussion

One has to remember that the resampling does not imply that the audio signal has changed, but rather if features change e.g. rapidly, creating high frequency content in the features, these could be removed. So a direct comparison of what the feature computes and the degradation due to e.g. downsampling is not trivial. Most of the multi coefficient features, that often contain feature data that is computed across audio bands, seem to suffer the most. As have been mentioned before these errors are not directly transferable to a potential error of a mathematical model. Caution has to be made, when using these features. Another aspect is that when downsampling variance within each feature is removed, thus results of variance and co-variance analysis could be altered due to this.

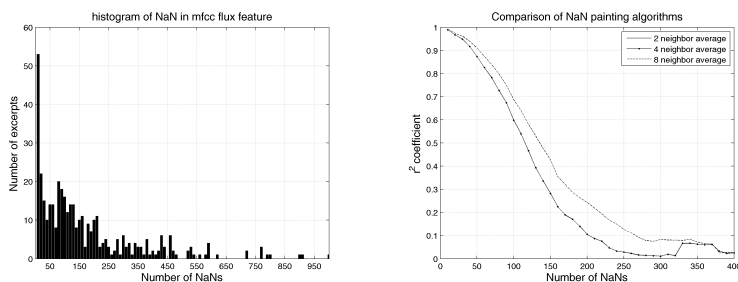
## B.4 Effect of $NaN$ painting

A number of features across all feature packs that were chosen produce few or a lot of  $NaN$  values. Either the choice would be to disregard these features and ignore them completely, but some features merely produce very few and therefore could be corrected using simple linear methods. Three different methods are used in this work, and that is simply making a set of linear equations that include the missing value and the surrounding. They are then solved for the missing value, where the number of values to include is varied. In this work a comparison of 2, 4 and 8 neighboring values are used, where the result of the missing value is a mean of these values. As an example the number of  $NaN$ s are shown on figure B.13(a) in the output of feature MFCC Flux from *MIR* extracted from the 200 excerpts from *pilot2* experiment. It is evident that of over half of the vector in some cases produce  $NaN$ , with a majority below 50. This could simply be errors in the script since a simple euclidean distance between two vector should not be difficult to compute. But this is simple an example and many more have these issues.

To test how well these methods to paint over the missing data perform, and at what level one should disregard the feature all together. 10 feature vectors were used where no  $NaN$  were present, to provide a representative picture. The distribution of the  $NaN$ s across feature vectors that have missing data was estimated, and was used to corrupt the 10 vectors. Incrementally with step of 10 up to 1000 was performed, representing from 0.6% to 60% in increments of 0.6% of the total amount om feature samples in that given vector.

### B.4.1 Results

The results of this approach can be seen on figure B.13(b).



(a) Histogram of 200 excerpts from (b)  $r^2$  results for  $NaN$ -painted MFCC *pilot2* experiment showing the number flux feature vector using an average of of  $NaN$  in the MFCC Flux feature from 10 excerpts.  
*MIR*

### **B.4.2 Discussion**

Up to 50 samples corrupted or 3% of the data the  $r^2$  correlation is above 0.9 for the method using the average of the 8 surrounding values. In general this method shows to have the best performance. Above 400 samples corrupted the method cannot produce results since clusters start to appear where the method cannot produce meaningful results, even using linear extrapolation. Features that have 50 or more missing data values will not be used, and the method of using 8 neighbor values will be used for all features, where 50 data points or below are missing.



# Listening experiment

---

All supplementary notes to the listening experiment section is presented in this appendix.

## C.1 Pilot1 - Graphical interface

The developed graphical Matlab interface that was used for the *pilot1* experiment will be presented here.

### C.1.1 Instructions

The instruction given for both pilot experiments are seen on figure C.1 and C.2. Aside for these instruction a verbal explanation was given of what was written in the text, to ensure that the task at hand was clear. Questions from participants were answered within the boundaries of the test.

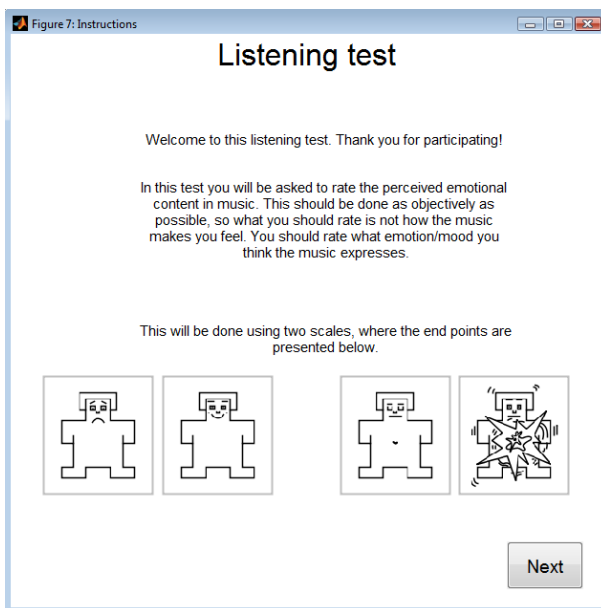


Figure C.1: First screen of instructions to the user, when participating in pilot1 listening experiment.



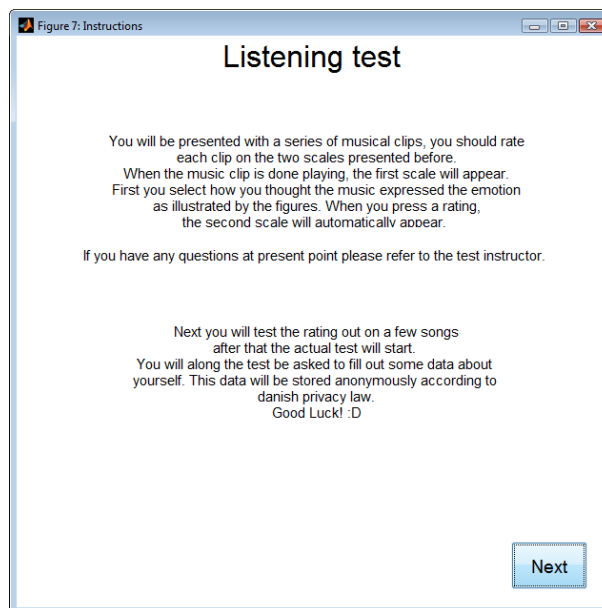
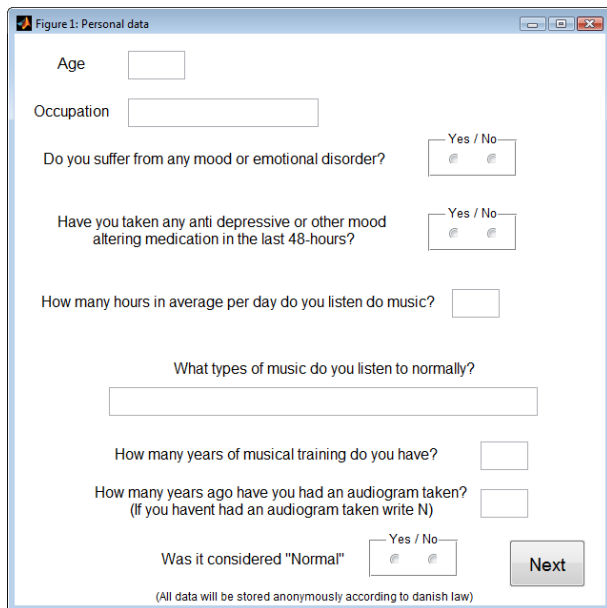


Figure C.2: Second screen of instructions to the user, when participating in pilot1 listening experiment.

### C.1.2 Meta data

To obtain some meta data about each of the participants a questionnaire was presented where the interface can be seen on figure C.3.



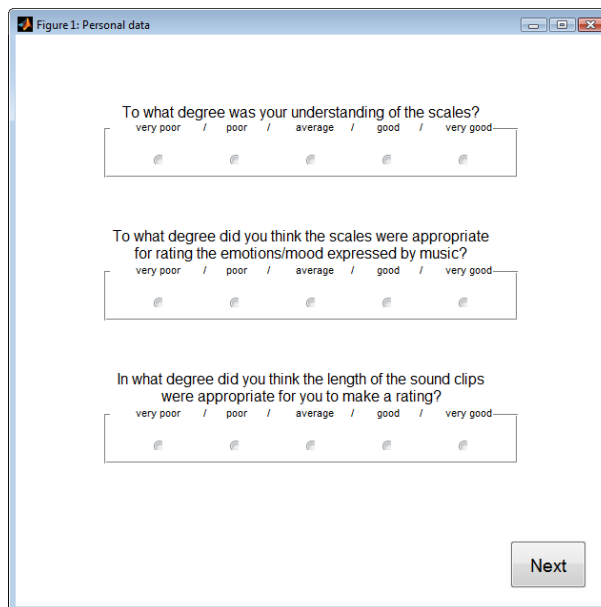
The screenshot shows a window titled "Figure 1: Personal data" with the following fields and questions:

- Age:
- Occupation:
- Do you suffer from any mood or emotional disorder?  Yes /  No
- Have you taken any anti depressive or other mood altering medication in the last 48-hours?  Yes /  No
- How many hours in average per day do you listen do music?
- What types of music do you listen to normally?
- How many years of musical training do you have?
- How many years ago have you had an audiogram taken? (If you havent had an audiogram taken write N)
- Was it considered "Normal"  Yes /  No

At the bottom right is a "Next" button. At the bottom center, there is a note: "(All data will be stored anonymously according to danish law)".

Figure C.3: Questionnaire given to participants prior to the test beginning.

After the ratings of excerpts another questionnaire was presented about the scales and length of each excerpt, which can be seen on figure C.4.



The image shows a window titled "Figure 1: Personal data" with three rating questions. Each question has a five-point scale with radio buttons. The first question is "To what degree was your understanding of the scales?" with options "very poor", "poor", "average", "good", and "very good". The second question is "To what degree did you think the scales were appropriate for rating the emotions/mood expressed by music?" with the same options. The third question is "In what degree did you think the length of the sound clips were appropriate for you to make a rating?" with the same options. A "Next" button is located at the bottom right of the window.

Figure 1: Personal data

To what degree was your understanding of the scales?  
very poor / poor / average / good / very good

To what degree did you think the scales were appropriate for rating the emotions/mood expressed by music?  
very poor / poor / average / good / very good

In what degree did you think the length of the sound clips were appropriate for you to make a rating?  
very poor / poor / average / good / very good

Next

Figure C.4: Questionnaire given to participants after the had finished.

### C.1.3 Prior mood

To see if there was a connection between the mood the person was in, prior to the test, and the resulting ratings, the participants were asked to rate what mood they were in prior to the test, given the two scales. An example of one of the scales are shown in figure C.5.

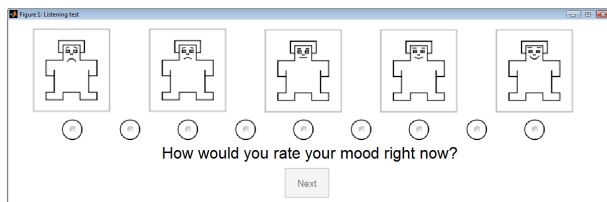
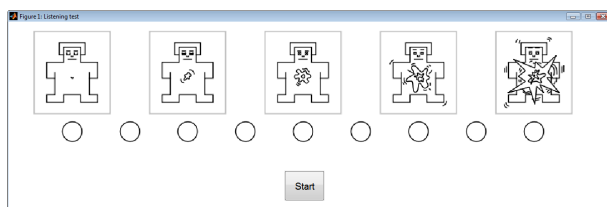


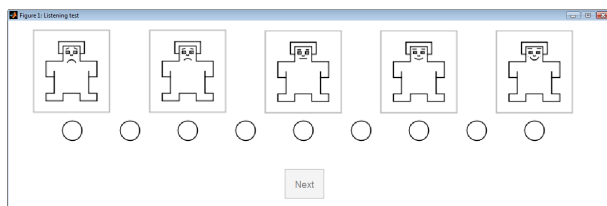
Figure C.5: Rating of mood prior to the test had begun.

### C.1.4 Primary interface

The two scales used to measure the emotional content expressed in music, using the dimensions of valence and arousal, in the form of manikins are shown on figure C.6. The scales were used in both *pilot1* and *pilot2* experiments.



(a) Arousal rating.



(b) Valence rating.

Figure C.6: Manikins used to measure the valence and arousal of emotions expressed in music.

## C.2 Pilot1 - Meta data analysis

All data acquired in *pilot1* except the ratings themselves are presented here. These include the post questioning and the temporal analysis of the test.

### C.2.1 Temporal analysis

As a mean of analyzing how long a potential listening test would take, a timer was used in the rating of each excerpt of each participant. Using this data an analysis can also be made of any excerpts that might be more difficult to rate, e.g. participants did not know what to rate and therefore it took longer time for them to rate. It could also be that given the specific ordering participants in general just need more time, since the cognitive load has increased. On figure C.7 the average time across participants is shown, where no excerpt in particular stand out as being more or less difficult to rate. But comparing the two designs, it seems that using the balanced ordering design, it requires a greater amount of time to rate. In average it took 6 seconds to rate each of the excerpt for the sequential design, whereas 8.7 seconds for the balanced design.

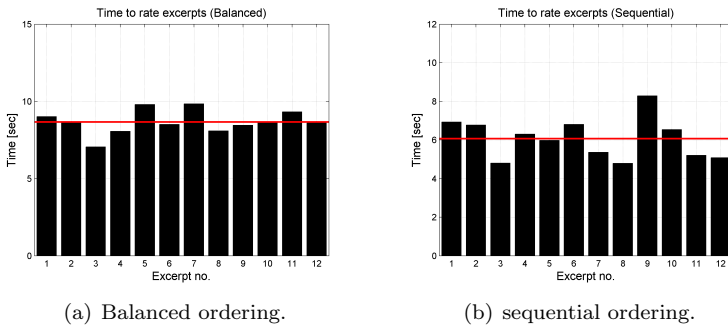


Figure C.7: Average time, over all participants, for each of the ordering designs, it took to rate each excerpt in *pilot1*. (left) Balanced ordering (right) sequential ordering

Comparing the time across test participants, participant 1 in the balanced ordering design, seem to be using quite a lot more time to rate each excerpt.

Using this data acquired, if using the time to rate as an indicator of the cognitive load the participants experience when rating the musical excerpt. Then the balanced design is indeed more difficult than the sequential. This could be an effect of the rather short excerpts of only 7.5 seconds. When participants remember what they just listened to and being familiar with the excerpt they listening to at that given time then, it is easier for them. This tendency is clear

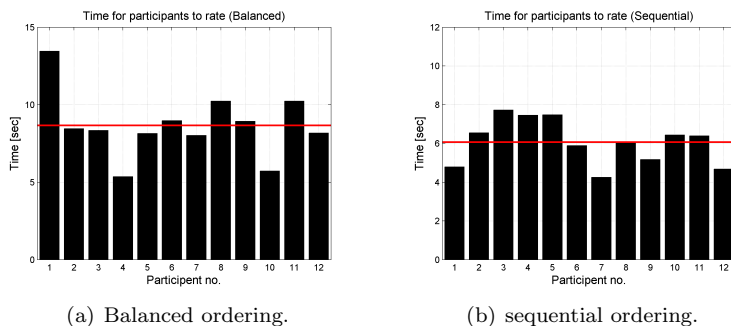


Figure C.8: Average time, over all excerpts, it took each participant in the pilot1 experiment to rate an excerpt. (left) Balanced ordering (right) sequential ordering. The method used to measure the time was the *Matlab* method called *tic, toc* this measure is an approximated method and thus not completely reliable. All times were corrected with the time it took to change after playing an excerpt.

when looking at excerpt 1, 5 and 9 on figure C.7(b) which are the first excerpts of a given clip. These have the highest times to rate of the excerpt from that clip. After that the time to rate them goes down for clip 2 and 3, where clip 1 there is a small deviation at excerpt 4.

### C.2.2 Analysis of scales

The purpose of the questions in *pilot1*, was to establish whether or not participants understood the scales, found them appropriate for rating music and last if the excerpt used had an appropriate length. The results for the 24 participants are shown on figure C.9. It seems that a majority did in fact understand the scales, with a mean little over average. To the question if the participants found the scales appropriate for rating music, the picture is slightly more negative where 25% of the participants found the scales poor.

### C.2.3 Analysis of excerpt length

Using musical excerpts in this test of only 7.5 seconds, the participants were asked if that was appropriate. Judging the results on figure C.9(right) that a majority did not find it so. The average of all ratings was a little below "average".

Post questioning of the participants revealed that they found the excerpt too short. They had to use a high amount of cognitive power, making them very exhausted after the test. Another aspect that became clear was that, the scales given without any anchoring adjectives, made it very difficult for participants

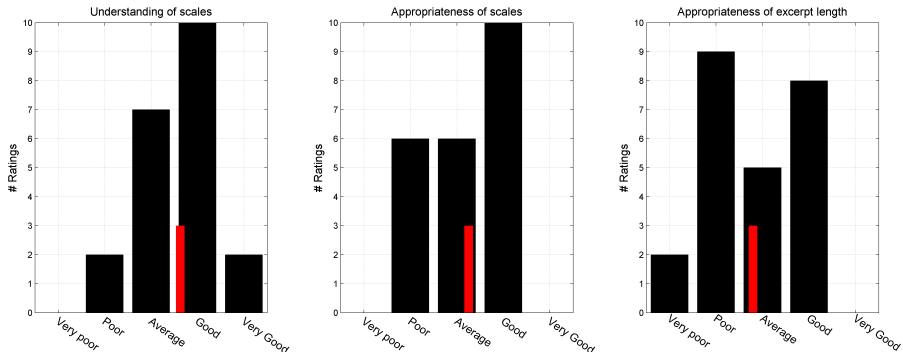


Figure C.9: Results of post questionnaire in the pilot1 experiment. (left) the understanding of the scales (middle) the appropriateness of the scales, (right) the appropriateness of the excerpt length. Red line indicates the mean of ratings.

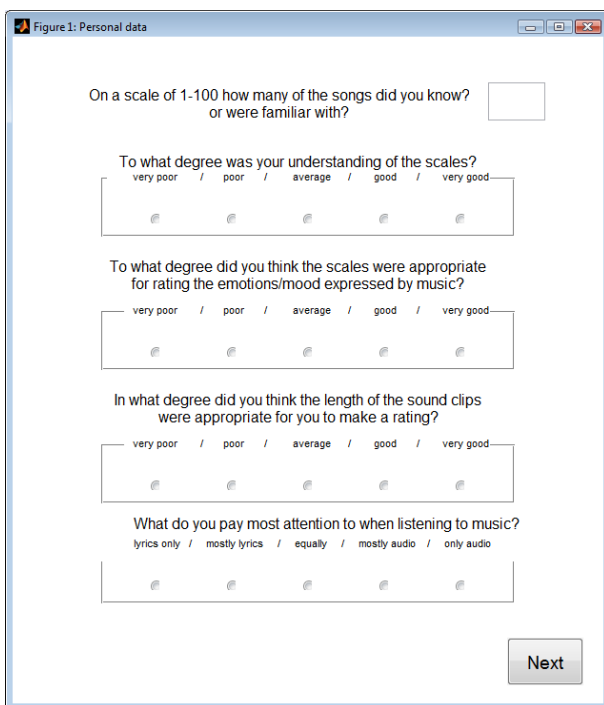
to exactly understand what the arousal scale meant. Due to this fact it is suggested that the scales need to be explained, if that is not enough, adjective end point or markers should be used.

## C.3 Pilot2 - User interface

The interface for the second pilot experiment was nearly identical to that of the first. Two different things were changed in the questions posted prior and post test.

### C.3.1 Meta data

Prior to testing question were asked as was in *pilot1* as was seen on figure C.3, the only extra question that was added, was that of nationality. After the test more questions were added than in the *pilot1* experiment as seen on figure C.10.



The screenshot shows a window titled "Figure 1: Personal data" with a standard Windows-style title bar. The window contains five questions, each with a corresponding input field or radio button options:

- Question 1: "On a scale of 1-100 how many of the songs did you know? or were familiar with?" followed by a text input box.
- Question 2: "To what degree was your understanding of the scales?" with radio button options: "very poor", "poor", "average", "good", "very good".
- Question 3: "To what degree did you think the scales were appropriate for rating the emotions/mood expressed by music?" with radio button options: "very poor", "poor", "average", "good", "very good".
- Question 4: "In what degree did you think the length of the sound clips were appropriate for you to make a rating?" with radio button options: "very poor", "poor", "average", "good", "very good".
- Question 5: "What do you pay most attention to when listening to music?" with radio button options: "lyrics only", "mostly lyrics", "equally", "mostly audio", "only audio".

A "Next" button is located at the bottom right of the window.

Figure C.10: Questionnaire given to participants after the had finished.

## C.4 Pilot2 - Bitrates of musical data

The distribution of bitrates of the musical data used in *pilot2* is shown on figure C.11.



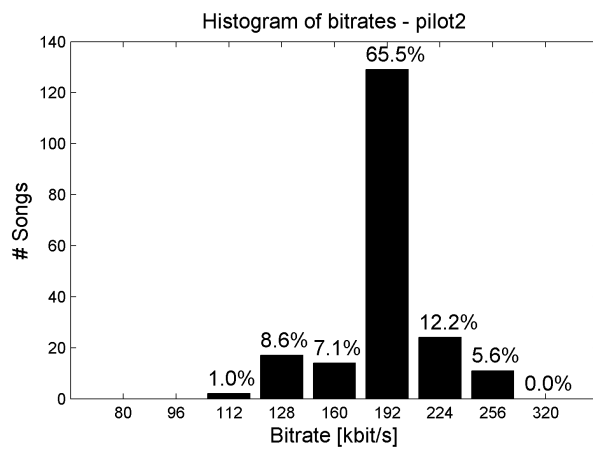


Figure C.11: Histogram of the bitrates of each of the 200 excerpts used in Pilot 2 experiment. Both for CBR and VBR the averages over the whole track was used, where the edges were (28 72,88,104,120,144,176,208,240,280 and 320) kbit. The number above each bar indicates the percentage of the total amount of excerpts

## C.5 Pilot2 - All ratings

### C.5.1 Arousal

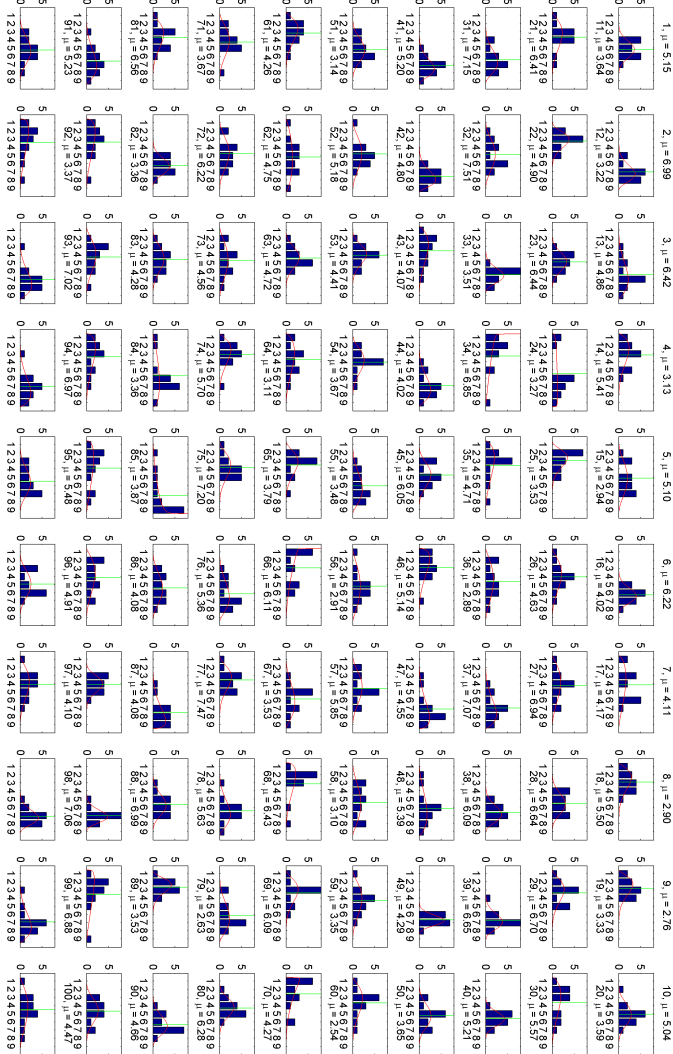


Figure C.12: Histogram of each of the 1-100 rated excerpts *pilot2* experiment rated in the arousal scale. The green line indicates the  $\mu_\beta$  for the individual excerpt. The red line is the beta distribution fitted to the experimental data.

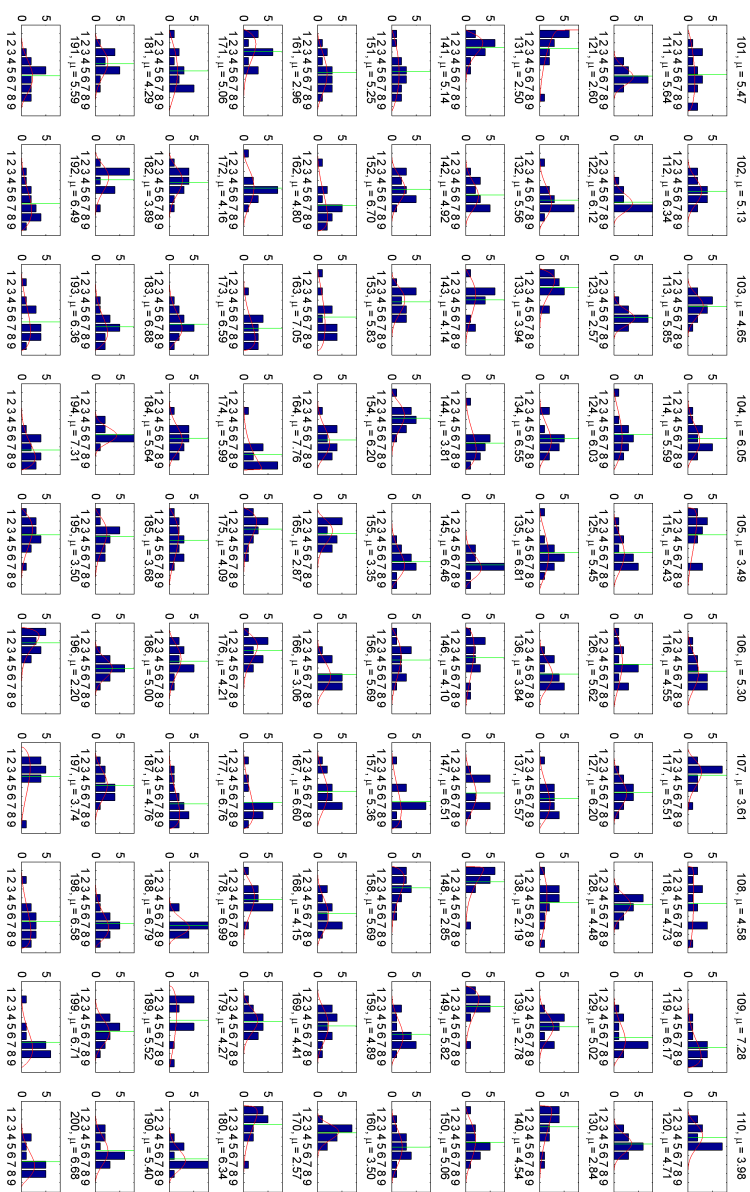


Figure C.13: Histogram of each of the 101-200 rated excerpts *pilot2* experiment rated in the arousal scale. The green line indicates the  $\mu_\beta$  for the individual excerpt. The red line is the beta distribution fitted to the experimental data.

### C.5.2 Valence

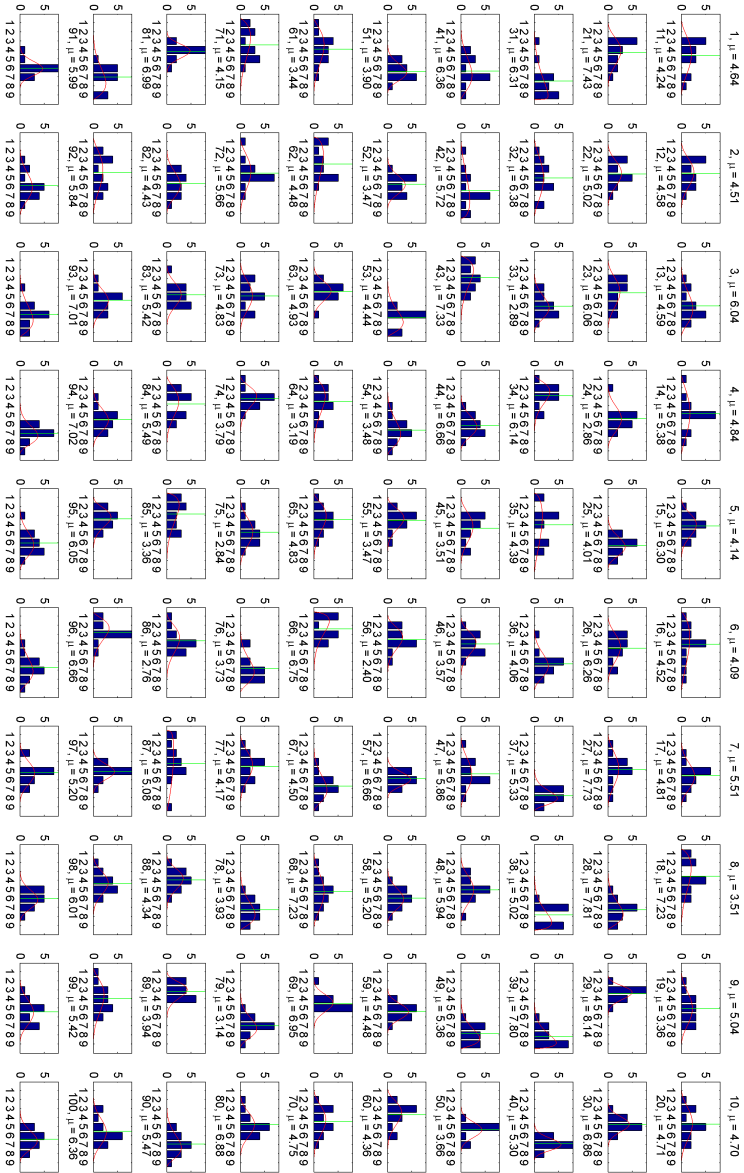


Figure C.14: Histogram of each of the 1-100 rated excerpts *pilot2* experiment rated in the valence scale. The green line indicates the  $\mu_\beta$  for the individual excerpt. The red line is the beta distribution fitted to the experimental data.

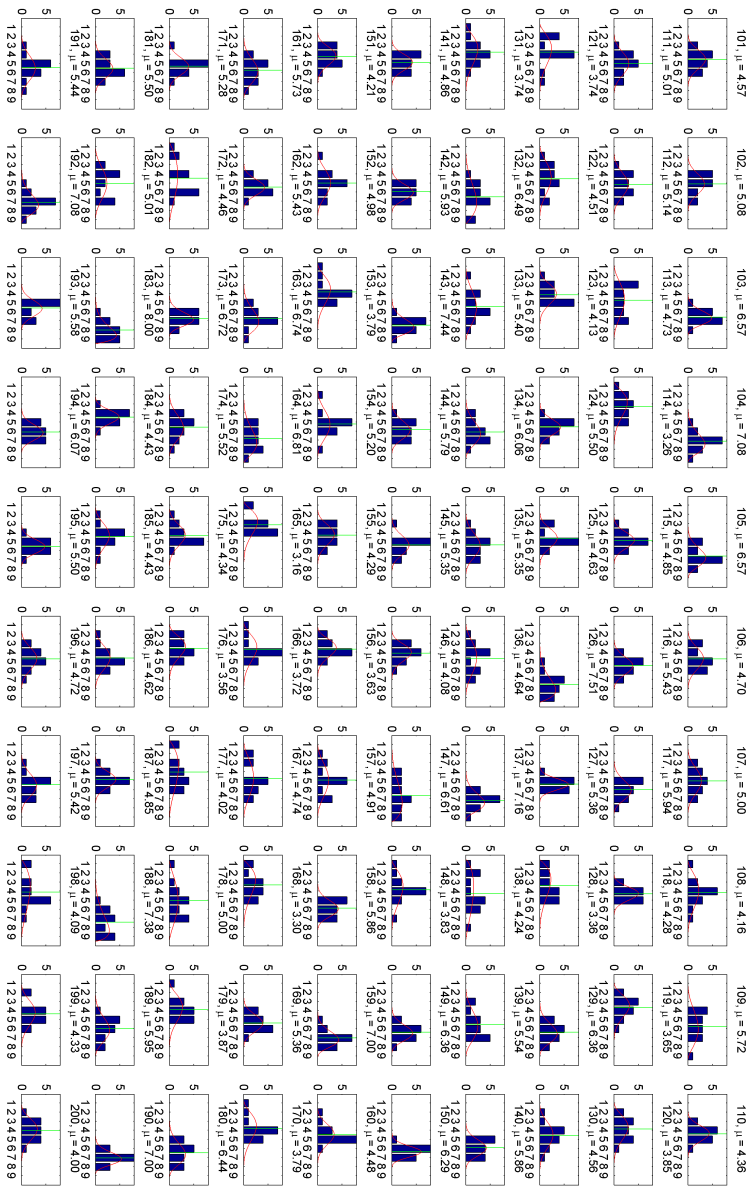


Figure C.15: Histogram of each of the 101-200 rated excerpts *pilot2* experiment rated in the valence scale. The green line indicates the  $\mu_\beta$  for the individual excerpt. The red line is the beta distribution fitted to the experimental data.

## C.6 Pilot2 - Pre-emotional ratings

The histograms of the ratings of each of the participants are presented here.

### C.6.1 Arousal

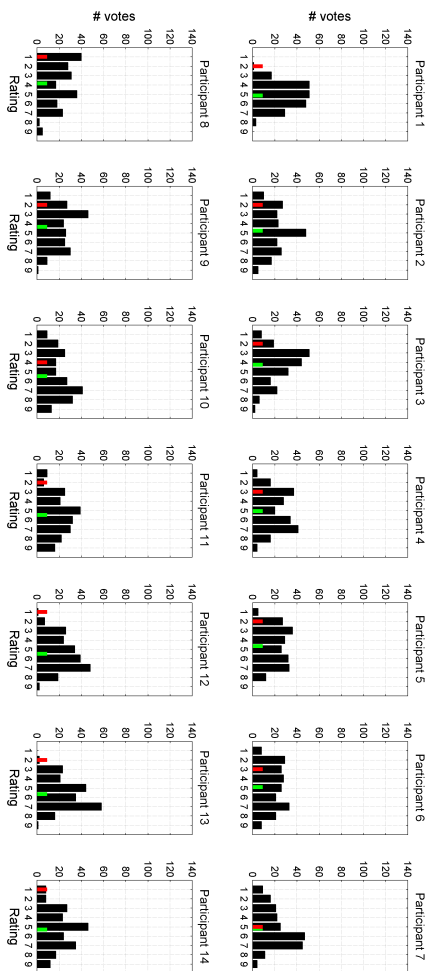


Figure C.16: Histogram of each of the participants arousal ratings across the 200 excerpts in the pilot2 experiment. The green line indicates the mean value of the individual test participants ratings. The red line indicates the arousal rating the participants gave prior to the experiment start.

## C.6.2 Valence

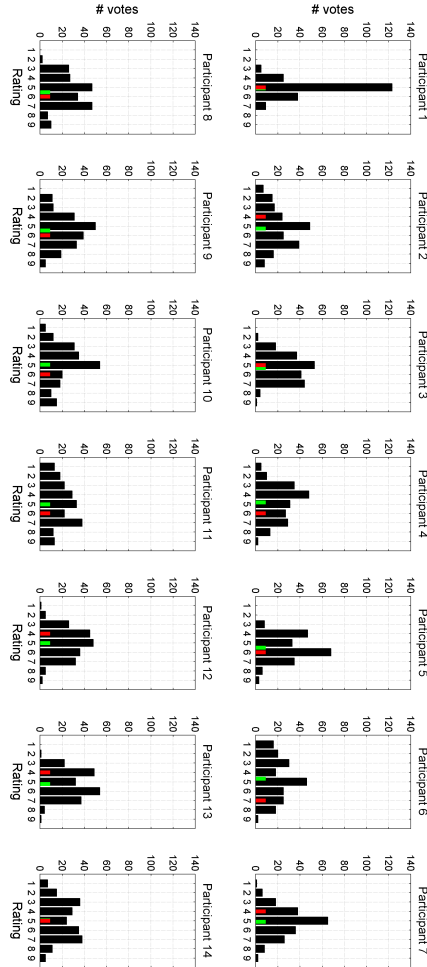


Figure C.17: Histogram of each of the participants valence ratings across the 200 excerpts in the pilot2 experiment. The green line indicates the mean value of the individual test participants ratings. The red line indicates the valence rating the participants gave prior to the experiment start.

## C.7 Pilot2 - Analysis of pre-emotional ratings

All ratings across all excerpts for each participant was shown in section C.6 on figures C.17 and C.16. To see if the emotional rating given by participants prior to the test had an effect on their ratings of all the musical excerpts an analysis is made. The distributions given in section C.6 form the basis of this analysis. An effect could be that given the prior mood of a participant the mean of all ratings would change or the participant would rate with a lower variation. To test this the two histograms are parameterized using the 50 th percentile (i.e the median) and the 25 th percentile. The difference between these two give a very robust measure of the variation, and the median is also a very robust measure of the general tendency of rating provided by a participants.

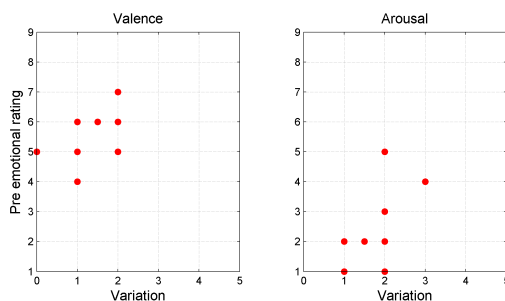


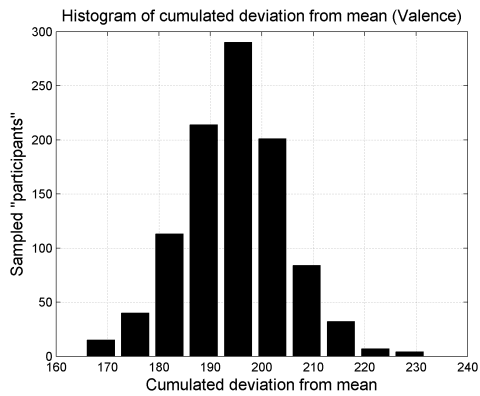
Figure C.18: The emotional ratings provided by participants prior to *pilot2*, representing their mood at that present point. These are compared to the variation of ratings provided by participants across all 200 rated excerpts. The variation is calculates as the difference between the 50 th (i.e the median) and the 25 th percentile of the histogram of all ratings provided by each participants.

Visual inspection of figure C.18 does not show any correlation tendency in the data. To ensure that there is no structure the Pearson's squared correlation coefficient  $r^2$  is computed (see (B.1) in section B.2) for both valence and arousal data. Valence data gives  $r^2$  of 0.218 and arousal gives 0.227, which does not show a strong correlation. The fact that there is no correlation or connection between the two set of emotional ratings gives an indication that participants were good at ignoring their mood, when rating the expressed emotions in music.

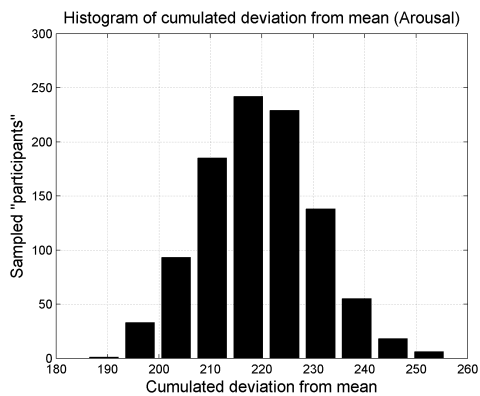
## C.8 Pilot2 - OC1 data foundation

Using the *OC1* on 1.000 sampled ratings from each beta distribution a histogram of the accumulated deviation from mean is shown on figure C.19.





(a) Valence



(b) Arousal

Figure C.19: On (a) and (b) the histogram of the accumulated deviation from mean is plotted, using 1000 ratings sampled from each of the 200 fitted beta distributions, simulating 1000 participants.

## C.9 Pilot2 - Outlier removal

Using *OC2* on the experimental data obtained in *pilot2* the number of outliers removed per excerpt is shown on figure C.20(b) and the number of outliers removed from each participant is shown on figure C.21(b). All figure sum to a total of 85 ratings for valence and 160 for arousal.

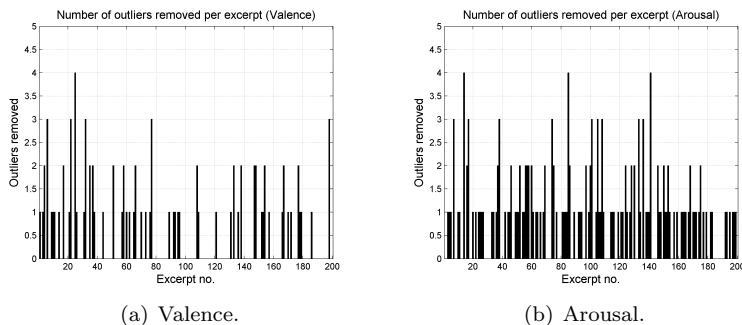


Figure C.20: Number of outliers removed using *OC2* per excerpt.

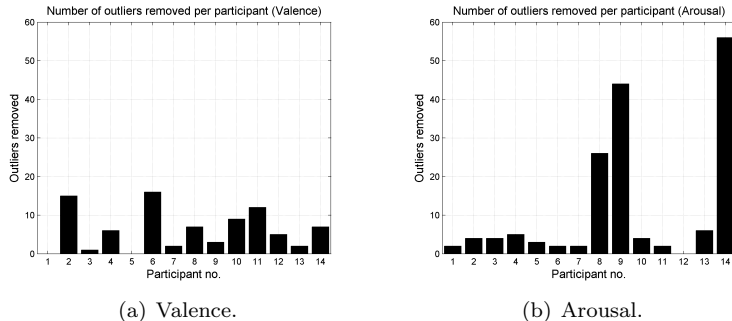


Figure C.21: Number of outliers removed using *OC2* per participant.

On figure C.21(b) it is evident that participant 8, 9 and 14 has clearly the highest amount of ratings removed.

## C.10 Pilot2 - Distribution of beta mean

The distribution of the mean of each beta distribution fitted to the experimental data obtained in *pilot2* is shown on figure C.22 for the arousal scale, together with the distribution of beta mean where outliers have been removed. The data from the valence scale is presented on figure C.23.

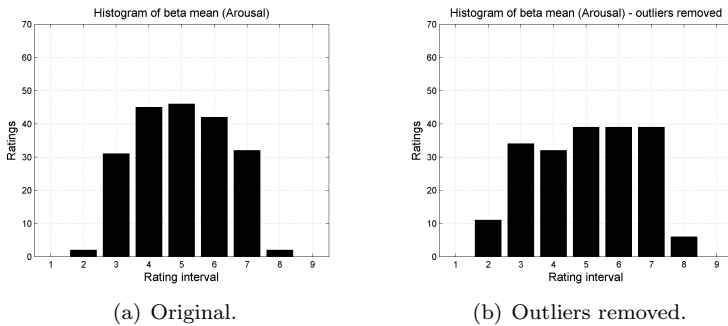


Figure C.22: Distribution of  $\mu\beta$  before and after outlier removal using *OC2* on the arousal scale. The mean and standards deviation for C.22(a) are  $\mu_n = 5.14$  and  $\sigma_n = 1.24$ , and for C.22(b) they are  $\mu_{n,out} = 5.18$  and  $\sigma_{n,out} = 1.25$

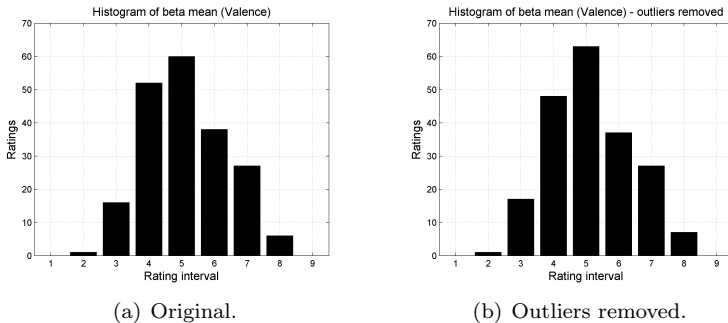


Figure C.23: Distribution of  $\mu\beta$  before and after outlier removal using *OC2* on the valence scale. The mean and standards deviation for C.23(a) are  $\mu_n = 4.96$  and  $\sigma_n = 1.36$ , and for C.23(b) they are  $\mu_{n,out} = 4.98$  and  $\sigma_{n,out} = 1.58$

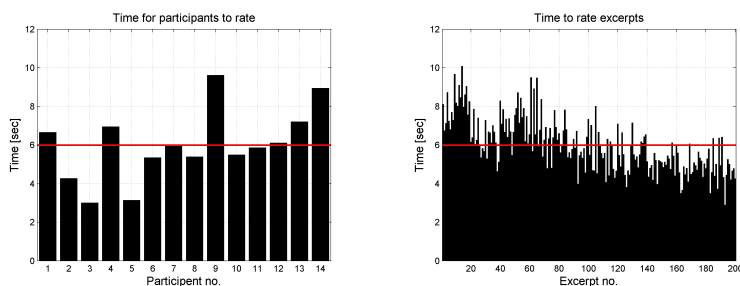
Looking at the difference in standard deviation of the distribution of beta mean, there is a clear widening of the distribution as an effect of the outlier removal, when looking at the arousal scale results.

## C.11 Pilot2 - Meta data analysis

All the data obtained in the pilot2 experiment other than the valence and arousal ratings will be analyzed and discussed here.

### C.11.1 Temporal analysis

As a mean of analyzing how long a potential listening test would take, a timer was used in the rating of each excerpt of each participant. Using the same method and analysis as was used in *pilot1* where the results can be seen in section C.2.1. An extended analysis could be made of participants, either if they are fast or slow at rating, they could potentially have more or less deviation from other participants ratings. On figure C.24(a) the time it took each participants to rate all 200 excerpts in average is shown.



(a) Average time, over all excerpts, it took each participant in the pilot2 experiment to rate an excerpt. (b) Average time, over all participants, it took to rate each excerpt in the pilot2 experiment.

Figure C.24: Temporal analysis of the *pilot2* experiment. Red line indicates the average time to rate one excerpt over all participants. Since breaks within the testing was allowed, durations of over 60 seconds per excerpt was removed. The ratings was set to the mean of all other participants for that excerpt. The method used to measure the time was the *Matlab* method called *tic*, *toc* this measure is an approximated method and thus not completely reliable. All times were corrected with the time it took to change after playing an excerpt.

Participant 9 and 14 seem to take longer time than the average across all excerpts, with a time close to 9 and 10 seconds respectively. Participants 2, 3 and 5 seem to be very fast in their ratings averaging around 4 and 3 seconds respectively. It is difficult to draw a direct line between the cognitive load of the participants due to these temporal measurements of ratings, but being 3-4 seconds slower per excerpt could indicate that those particular participants found the task difficult.

Looking across excerpts on figure C.24(b) there is a great deal of variance in the time it takes to rate each excerpt. *The Chemical Brothers - Chemical Beats* was rated in average in 2.88 seconds, whereas *Amon Tobin - Cosmo Retro Intro Outro* was rated in average in 10.1 seconds. Admittedly the track is difficult to interpret and might be outside the mainstream musical repertoire. Using these results excerpts could be excluded from other tests due to the cognitive load they strain people with. Conversely one might, by using that approach, exclude excerpts that give a great deal of information, in the mathematical modeling. Comparing the results of the temporal analysis of *pilot1* (see section C.2.1) and the present, it shows that the time to rate even though using a balanced ordering design, the times have gone down. This could either be by training, since there are more excerpts to rate here, and thereby people become faster, or it is a result of the increased excerpt length. Likely it is a combination of the two.

### C.11.2 Rating of scales

The two manikin based scales were rated by the participants after *pilot2* was finished, see figure C.10. The results for the 14 participants are shown on figure C.25 (left), where they were asked whether or not they understood the scales, i.e. the manikins. A majority of participants rated either average or good, with a mean rating of the scales, between average and good. Post verbal questioning showed that the scale that was the most difficult to understand was the scale of Arousal. People intuitively did not understand that the manikin was excited or not excited as was instructed.

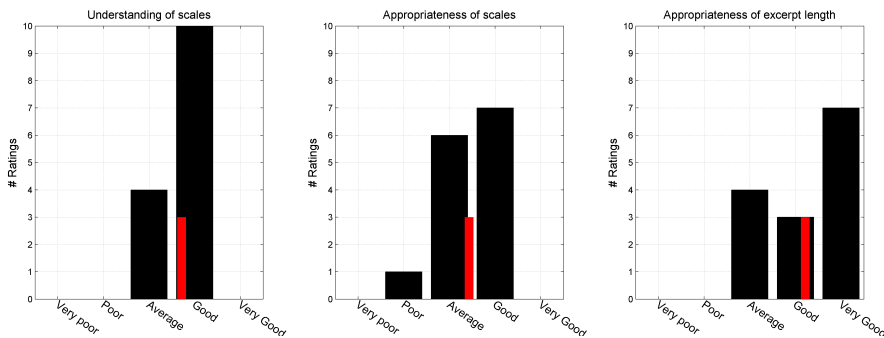


Figure C.25: Likert ratings for the (left) understanding of the scales/manikins used. (Middle) Appropriateness of the scales for use with music. (Right) Appropriateness of the excerpt length.

The second question was whether or not the participants found that the ratings were appropriate for the rating of music. The results can be seen in the middle on figure C.25. The ratings here lean more towards average with one participant

finding it poor. By post verbal questioning it showed again that the dimension of arousal was hard to transfer to music, again this might be influenced by the fact that some participants did not understand the scale completely, even though thorough instructions were given.

The last question was regarding the appropriateness of the length of each excerpt. In *pilot2* each excerpt was 15 seconds long, where *pilot1* the excerpt were only 7.5 seconds. In *pilot1* the average rating of the appropriateness of excerpt length was a little under average whereas in *pilot2* it is above good. Showing that the increase in excerpt length had a very beneficial effect.

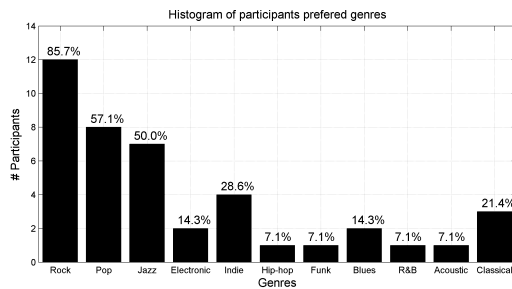
### C.11.3 Musical background

In this section a series of question to each participants were asked, in order to obtain some knowledge about their preferences and musical experience.

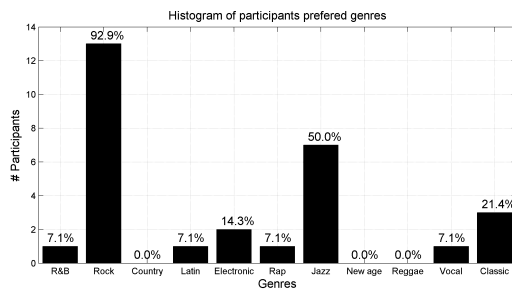
#### Musical preference by participants

Given that a great deal of different genres were tested in *pilot2*, see figure 4.4 in section 4.6. Potentially participants that normally would listen to music that is much different to that of the test, would be biased and thereby rate much different than other participants. On figure C.26 a histogram of the genres as was stated by each participant is shown.

In figure C.26(b) the genres were converted using the *AMG* genres to compare with the genres of the music that was in the test. The conversion itself is biased in the form of the selection that is present in *AMG*. The fact that pop and rock are put in the same category can be debated. Nonetheless a majority of the participants indeed have a preference for pop/rock which is also heavily weighted in *pilot2*. Jazz is not that used neither is Latin or blues. Since the preferred music by the participants are similar to the broad genres of the test, given the broad genres, participants could tend to give a different ratings, than compared to other participants that would reproduce this test.



(a) Preferred genres as stated by the participants in the test.

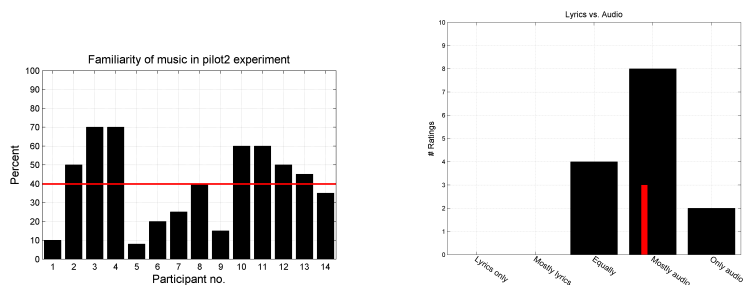


(b) Genres were converted using the same method of the AMG.

Figure C.26: Pre-test questioning of test participants in pilot2 listening experiment, regarding their preferred genres. The participants were allowed to name multiple, thus % are not summing to 100.

### Familiarity

One thing is whether or not participants like the genre, but if they can recognize a great deal of the songs, they might tend to be bias. Another aspect of the fact that they know a lot of songs is that it might be difficult for them to only rate the excerpt and the contents in them. They might tend to rate the song as a whole since they know that excerpt within a context. On figure C.27(a) the percentage of the excerpts that each participant was familiar with is shown. An average of 40% is known by the participants, with a minimum of 8% and a maximum of 70%. The goal was to choose tracks that were not particularly popular, but given the data it seems that for some participants they were very familiar with the data. Being familiar can be interpreted to know the artist, that specific track or that genre in general. A comparison between ratings of participants and their preference can be looked into.



(a) How familiar participants were with the excerpts they rated, rating from 1-100. (b) What participants pay attention to when listening to music.

Figure C.27

### Lyrics vs. Audio

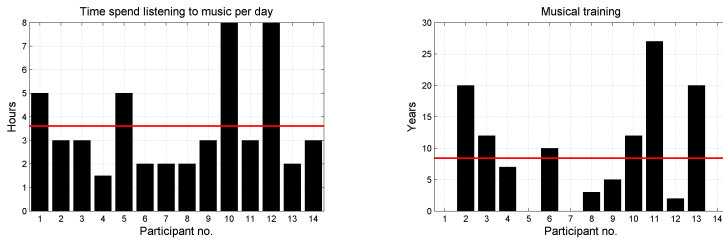
Given that the choice of features used within the test, it was relevant to obtain some indication amongst the test participants of what they paid most attention to, lyrics or the audio of music. On figure C.27(b) the results of the questionnaire is presented, which shows that for the participants within the experiment, there is a clear tendency towards listening to mostly audio and only audio. This result is also influenced by the genres that people prefer, where some genres are more vocally and lyrics orientated and other e.g. electronic and jazz can be mostly musical.

### Musical experience

Two different measures to indicate the musical experience of participants were made. The first being how much time a person spends on listening to music. The results are presented on figure C.28(a), which shows an average of around



3.5 hours a day, with a maximum of 8 hours and a minimum of 1.5 hours. Which type of listening situation here is not known, e.g. is it background music in a workplace or is it intense listening situation, where the music is the center of attention. Nonetheless it gives an indication if people have a musical interest.



(a) Amount of time each participant spend listening to music each day. (b) The years of training each participant received of musical training.

Figure C.28

The other measure goes towards their musical skill, which is measured using the number of years they have had musical training. Between participants there is a very high variation, with an average of around 8 years and a maximum of 28 years, which is a few years under that participants age. Comparing figure C.28(a) and C.28(b) there seem to be a connection that if the participant has extensive musical training, the participant does not listen to music that much every day.

## C.12 Pilot2 - Meta data influence on ratings

A great deal of information about each participants was collecting in *pilot2* e.g. musical experience, preference of lyrics or audio, familiarity of tested music and their ratings of the scales used, etc. The results of these were all presented in section C.11. The purpose of this section is to investigate if there is any connection between any of these variables and the subsequent ratings provided by each participant. The approach to investigate this is the same as was done in section C.7, where the distribution of ratings provided by each participants for all excerpts is used. A variation measure is calculated by the different between the 50 th and the 25 th percentile, this is compared to the meta data. By visual inspection if there is any correlation between the two, there is indication of a connection. Furthermore the Pearson's squared correlation coefficient  $r^2$  is computed (see (B.1) in section B.2) for both valence and arousal data. Two examples of analysis are given on figure C.29 and C.30, where visual inspection does not show any particular tendency towards in structure or relation between the familiarity of excerpt or the time spend listening to music every day and the

participants ratings.

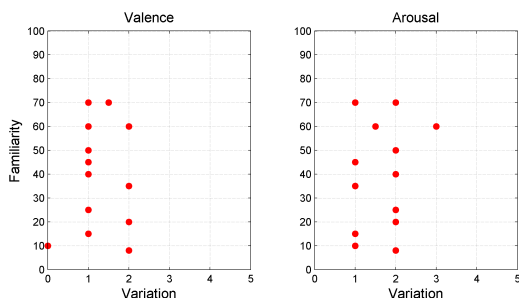


Figure C.29: The familiarity of excerpts in the *pilot2* experiment (1-100) are compared to the variation of ratings provided by participants across all 200 rated excerpts. The variation is calculates as the difference between the 50 th (i.e the median) and the 25 th percentile of the histogram of all ratings provided by each participants.

The correlation show the same picture with  $r^2$  for familiarity of 0.003 for valence and 0.041 for arousal. Same with hours spend listening to much each day which results in  $r^2$  to be 0.058 for valence and 0.160 for arousal.

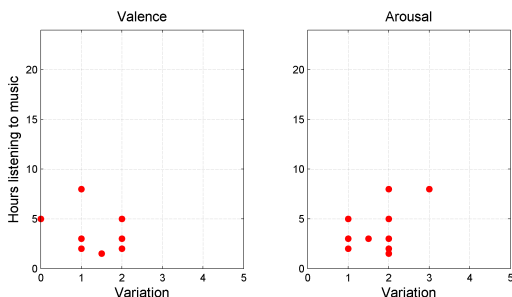


Figure C.30: The time spend listening to music each day by participants they participated in the *pilot2* experiment (0-24) are compared to the variation of ratings provided by participants across all 200 rated excerpts. The variation is calculates as the difference between the 50 th (i.e the median) and the 25 th percentile of the histogram of all ratings provided by each participants.

This test procedure was repeated for all the meta data parameters obtained in *pilot2* and similar results were found for all. There was no measurable connection between meta data provided by participants and their ratings.

## APPENDIX D

# Mathematical modeling

---

All supplementary notes to the mathematical modeling section is presented in this appendix.

### D.1 Features selected by *SFS*

No.	Feature name	Feature pack
1	MFCC - AUD (2/30)	ISP
2	Pulse Clarity	MIR
3	Main Loudness (15/24)	PSY
4	Pulse Clarity - Gammatone	MIR
5	Spectral Flatness per. band (10/19)	YAAFE
6	MFCC VB (3/30)	ISP
7	Spectral Flatness per. band (14/19)	YAAFE
8	Frequency of maximum energy in modulation (10-40Hz range)	YAAFE
9	Inharmonicity	MIR
10	<i>CENS</i> (2/12)	CM
11	Interaural Coherence (17/20)	ID
12	Envelope klapuri (324/410)	MIR
13	<i>CENS</i> (3/12)	CM
14	<i>CENS</i> (12/12)	CM
15	Excitation pattern (70/73)	PSY
16	Specific Loudness pattern (70/73)	PSY
17	Main loudness (17/24)	PSY
18	<i>CENS</i> (1/12)	CM
19	Envelope shape statistics (variance)	YAAFE
20	Fluctuations (15/15)	MIR
21	Envelope klapuri (375/410)	MIR
22	<i>CENS</i> (7/12)	CM
23	MFCC (21/40)	MA
24	Spectral Flatness per. band (13/19)	YAAFE
25	<i>CENS</i> (9/12)	CM
26	Tempo related to the highest autocorr.	MIR
27	Cepstral std. dev.	PSY
28	Cepstral centroid	PSY
29	Tempo	PSY
30	Tempo	MIR
31	<i>ILD</i> (15/15)	ID
32	Envelope klapuri (293/410)	MIR
33	Envelope klapuri (13/410)	MIR
34	Sonogram (23/24)	AM
35	<i>ILD</i> (5/15)	ID
36	<i>CENS</i> (11/12)	CM
37	<i>CENS</i> (5/12)	CM
38	Pitch (38/88)	CM
39	OBSIR (1/8)	YAAFE
40	<i>CENS</i> (10/12)	CM
41	Fluctuations (1/15)	MIR
42	Pitch (22/88)	CM
43	Loudness (3/24)	YAAFE
44	Loudness level	PSY
45	Energy difference between mean energy in range and energy at max freq . (10-40Hz range)	YAAFE
46	MFCC - VB (15/30)	ISP
47	Pitch (51/88)	CM
48	<i>CENS</i> (6/12)	CM
49	Pitch (84/88)	CM
50	Spectral Centroid	ISP
51	MFCC (2/40)	AM
52	Envelope klapuri (119/410)	MIR
53	Envelope klapuri (2/410)	MIR
54	Pitch (34/88)	CM
55	Envelope klapuri (306/410)	MIR
56	Pitch (15/88)	CM
57	MFCC (12/20)	YAAFE

Table D.1: Names of the 57 features selected by *Sequential Feature Selection*.

## D.2 Features selected by *LARS*

Here is an overview of the number of features selected by *LARS*.

Feature pack	Valence			Arousal			Original
	$\alpha$	$\beta$	$\mu_\beta$	$\alpha$	$\beta$	$\mu_\beta$	
CT	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	124
ISP	$0 \pm 0$	$0 \pm 0$	$11 \pm 38$	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	188
MA	$3 \pm 1$	$4 \pm 10$	$5 \pm 13$	$3 \pm 5$	$2 \pm 0$	$4 \pm 9$	64
MIR	$0 \pm 0$	$0 \pm 0$	$2 \pm 0$	$0 \pm 0$	$0 \pm 0$	$4 \pm 0$	510
PSY	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	$0 \pm 0$	$235 \pm 92$	302
YAAFE	$18 \pm 43$	$13 \pm 37$	$23 \pm 50$	$17 \pm 41$	$23 \pm 48$	$25 \pm 47$	141
ID	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	43
<i>ALL</i>	$2 \pm 1$	$0 \pm 0$	$0 \pm 0$	$2 \pm 0$	$2 \pm 1$	$0 \pm 0$	1373
<i>MIN</i>	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$3 \pm 0$	65
<i>PCA050</i>	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$2 \pm 1$	50
<i>FFS</i>	$2 \pm 0$	$2 \pm 1$	$2 \pm 0$	$2 \pm 0$	$2 \pm 0$	$9 \pm 1$	47

Table D.2: Number of features chosen for 7 different acoustical feature packs trained using *LARS* with a 50-fold *CV* method on two coefficients of the beta distributions denoted  $\alpha$  and  $\beta$ , and the beta mean  $\mu_\beta$ . Training was performed on each channel of the excerpts separately.

## D.3 Features selected by *stepwise*

Here is an overview of the number of features selected by *LARS*.

Feature pack	Valence			Arousal			Original
	$\alpha$	$\beta$	$\mu_\beta$	$\alpha$	$\beta$	$\mu_\beta$	
CT	116 $\pm$ 1	119 $\pm$ 2	117 $\pm$ 2	116 $\pm$ 1	119 $\pm$ 2	114 $\pm$ 1	124
ISP	159 $\pm$ 1	157 $\pm$ 1	161 $\pm$ 1	158 $\pm$ 2	157 $\pm$ 1	159 $\pm$ 2	188
MA	61 $\pm$ 1	62 $\pm$ 1	62 $\pm$ 1	61 $\pm$ 0	62 $\pm$ 0	61 $\pm$ 0	64
MIR	509 $\pm$ 0	509 $\pm$ 2	508 $\pm$ 4	508 $\pm$ 3	509 $\pm$ 0	509 $\pm$ 2	510
PSY	301 $\pm$ 0	301 $\pm$ 0	301 $\pm$ 0	301 $\pm$ 0	301 $\pm$ 0	301 $\pm$ 0	302
YAAFE	135 $\pm$ 1	135 $\pm$ 1	136 $\pm$ 1	132 $\pm$ 1	133 $\pm$ 2	134 $\pm$ 1	141
ID	43 $\pm$ 0	43 $\pm$ 0	43 $\pm$ 0	43 $\pm$ 0	43 $\pm$ 0	43 $\pm$ 0	43
ALL	0 $\pm$ 0	0 $\pm$ 0	0 $\pm$ 0	0 $\pm$ 0	0 $\pm$ 0	0 $\pm$ 0	1373
MIN	65 $\pm$ 0	65 $\pm$ 0	65 $\pm$ 0	65 $\pm$ 0	65 $\pm$ 0	65 $\pm$ 0	65
PCA050	49 $\pm$ 0	49 $\pm$ 0	49 $\pm$ 0	49 $\pm$ 0	49 $\pm$ 0	49 $\pm$ 0	50
FFS	47 $\pm$ 0	47 $\pm$ 0	47 $\pm$ 0	47 $\pm$ 0	47 $\pm$ 0	47 $\pm$ 0	47

Table D.3: Number of features chosen for 7 different acoustical feature packs trained using *stepwise* with a 50-fold *CV* method on two coefficients of the beta distributions denoted  $\alpha$  and  $\beta$ , and the beta mean  $\mu_\beta$ . Training was performed on each channel of the excerpts separately.

## D.4 Emotional ratings and audio features

To compare audio features selected by *SFS* and the emotional ratings participants provided in *pilot2* an analysis of some of these features is made. To simplify the emotional ratings for comparison the emotional data is divided into four quadrants of the two dimensional emotional model. Furthermore to reduce the influence of middle ratings, e.g. potential "dont know" ratings the following boundaries are used .

1. Valence < 4.5, Arousal > 5.5 (Black) - Negative-Excited
2. Valence > 5.5, Arousal > 5.5 (Blue) - Positive-Excited
3. Valence < 4.5, Arousal < 4.5 (Red) - Negative-Not excited
4. Valence > 5.5, Arousal < 4.5 (Magenta) - Positive-Not excited

Given the boundaries for 4 different quadrants of the valence-arousal space the grouping of the musical excerpt can be seen on figure D.1. The data obtained in *pilot2* is not evenly distributed across the quadrants resulting in a difference between the number of excerpts in each group. Group 4 has very few excerpt where group 3 has a high amount of excerpt. A reduction of excerpts to make the amount equal in each quadrant is not made since the data foundation would be very small to perform a comparison of emotional ratings and audio features.

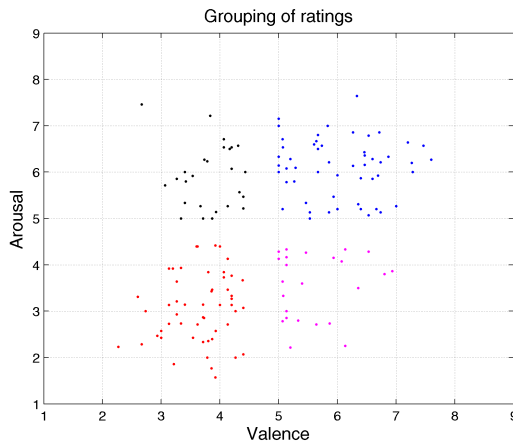


Figure D.1: Grouping of valence and arousal data, where 108 excerpt were chosen for the analysis.

### D.4.1 Results

*Boxplots* of selected features are shown in figure D.2 and D.3. The red line in each blue box indicated the median of the mean feature vector for each grouping. The blue line around each box indicates the 25 th and the 75 th percentile. The so-called whiskers or bars on each box indicate the highest and lowest values for that given feature. Outliers are not shown on these plots.

Common for all features shown in figure D.2, *CENS*, inharmonicity, average loudness and average of the coefficient 2 and 3 of *MFCCs* is that they separate groups 1 and 2 from 3 and 4.

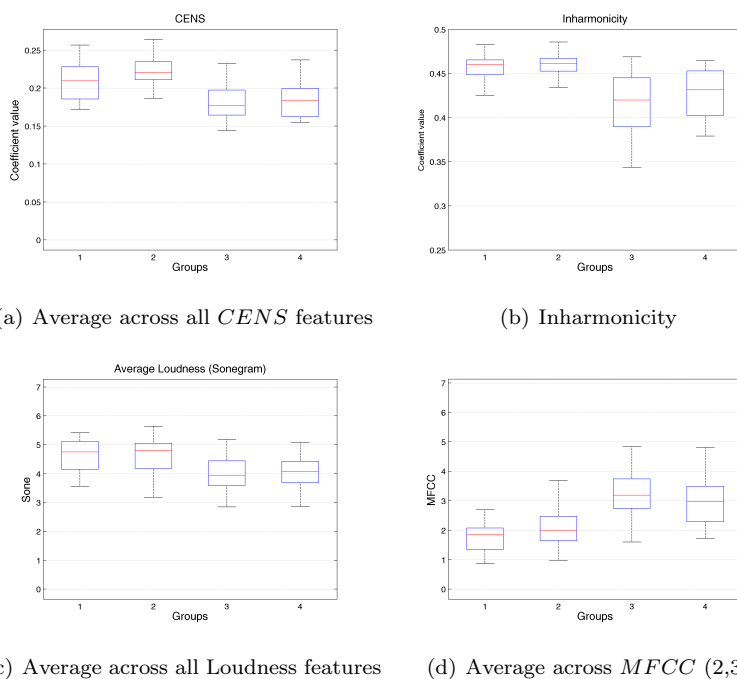


Figure D.2: Comparison of features of excerpts grouped in 4 quadrants based on the emotional ratings obtained for those excerpts.

On figure D.3 showing tempo and pulse clarity, surprisingly the tempo features only show slightly higher values only for group 2 (Positive-Excited). Groups 1, 3 and 4 have very similar tempo across all excerpts. Looking at pulse clarity ([Lartillot et al., 2008]) a clear distinction of group 2 the Positive-Excited grouping is seen. Both features show a separation across the valence dimensions. Figure D.3(b) which shows the average of all flatness per band coefficients show as the features displayed on figure D.2 that it separates the arousal dimension.



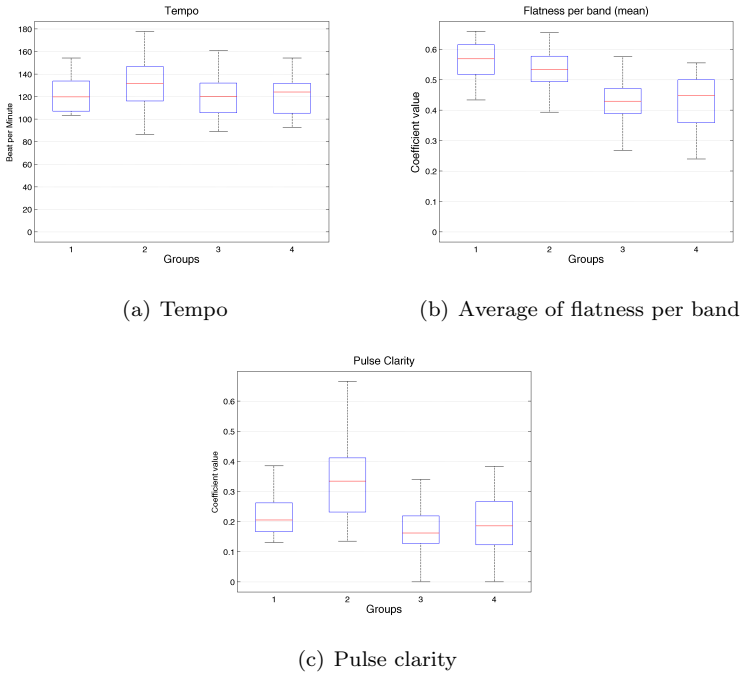


Figure D.3: Comparison of features of excerpts grouped in 4 quadrants based on the emotional ratings obtained for those excerpts.

## D.4.2 Discussion

6 different features are compared using grouping of emotional ratings into 4 quadrants. 5 out of 7 features, *CENS*, inharmonicity, Average loudness and average of *MFCCs* and flatness show a good separation of the arousal dimension. Whether the values are high or low there is a separation. For the valence dimension only pulse clarity and tempo show a difference for the Positive-Excited quadrant. Thus indicating that this quadrant contains music that has a higher tempo and a higher pulse clarity. Similar to the results found in this investigation for average loudness in [Laurier et al., 2009a] they obtained high values with low variation for averaged loudness with emotional categories anger, happy and tenderness which are associated with the dimension of arousal.

## D.5 Temporal emotional modeling

Using the predictions made by the regression models designed in chapter 5 a clustering of these predictions are attempted using non-supervised learning. The results for one excerpt for each of the valence and arousal dimensions are shown on figure D.4.

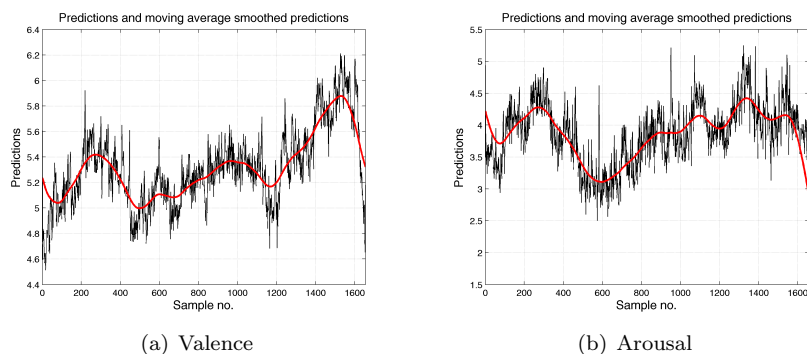


Figure D.4: Predictions of valence and arousal beta means by *stepwise* on *SFS*. Predictions were smoothed using a 200 tab moving average filter and down-sampled to 200 samples.

Using *K-Nearest Neighbor* (KNN) and *Gaussian Mixture Model* (GMM) an attempt in made to cluster the data. Each vector of emotional predictions of 200 sampled is used to represent that given excerpt. Using this data vector no *GMM* model could be found that would converge. Instead the naive *KNN* was used and the average of within-cluster sums of point-to-centroid distance averaging over 20 runs was used to find the optimum  $K$ . Using both this distance measure and visual inspection 5 was chosen to be appropriate for arousal and 4 for valence.

### D.5.1 Results

The resulting groupings for valence are shown on figure D.5. The initial 5 samples of predictions show a relative lower value compared to the rest of the temporal curve. This is due to an artifact of the audio features extraction and subsequently the predictions made by the model, thus it should be disregarded.

#### Valence

The 4 groups show a small variation of valence predictions, where cluster 1 seem to start high and dip in valence and to throughout the excerpt increase in valence. Cluster 2 and 4 are similar in progression but differ in the amount of fluctuations, where the 4 th cluster peaks in the beginning and drops off in the

end. Cluster 2 starts low, peaks in the middle and drops off at the end. The opposite of cluster 4 is 3 that starts low and builds up until the end.

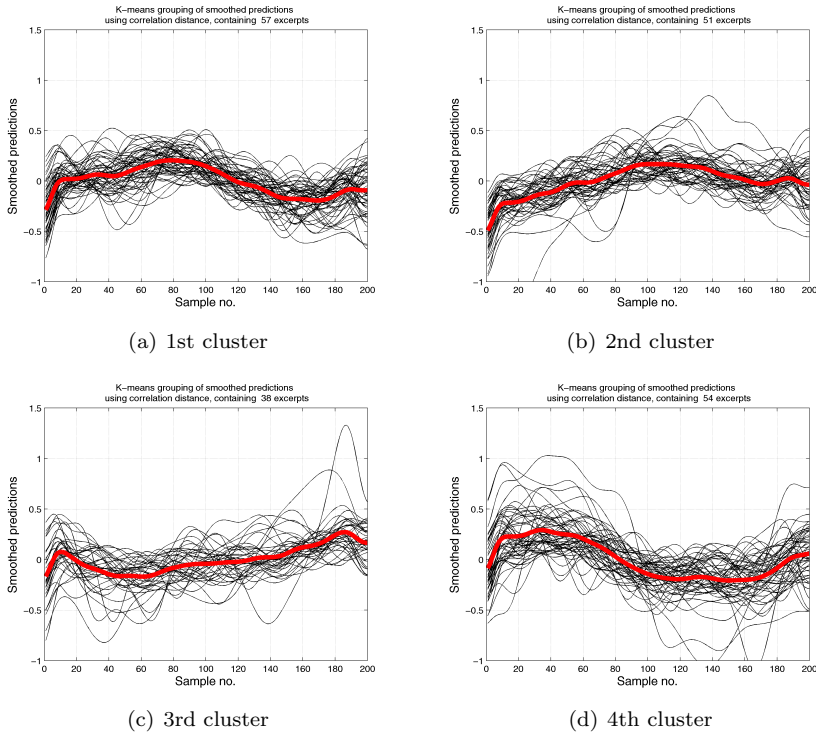


Figure D.5: 4 clusterings of emotional predictions using *KNN* on moving average smoothed temporal predictions using a *stepwise* regression model trained on *SFS* predicting beta mean coefficients of valence. Red line indicates the mean of all excerpt within the individual clusters.

### Arousal

The clustering of valence temporal predictions are shown on figure D.6. Clusters 1 and 2 are opposites where the first one starts up excited and drops off at the end, where cluster 2 starts low and builds up at the end. Cluster 3 seem to have a cyclic structure with peaks every 5 seconds. Cluster 4 seem to start at a medium level of excitement, then to drop down and rise until the end. Cluster 5 is the most extreme of the 5 clusters with a change of 1 rating from start and then drop down at the end.

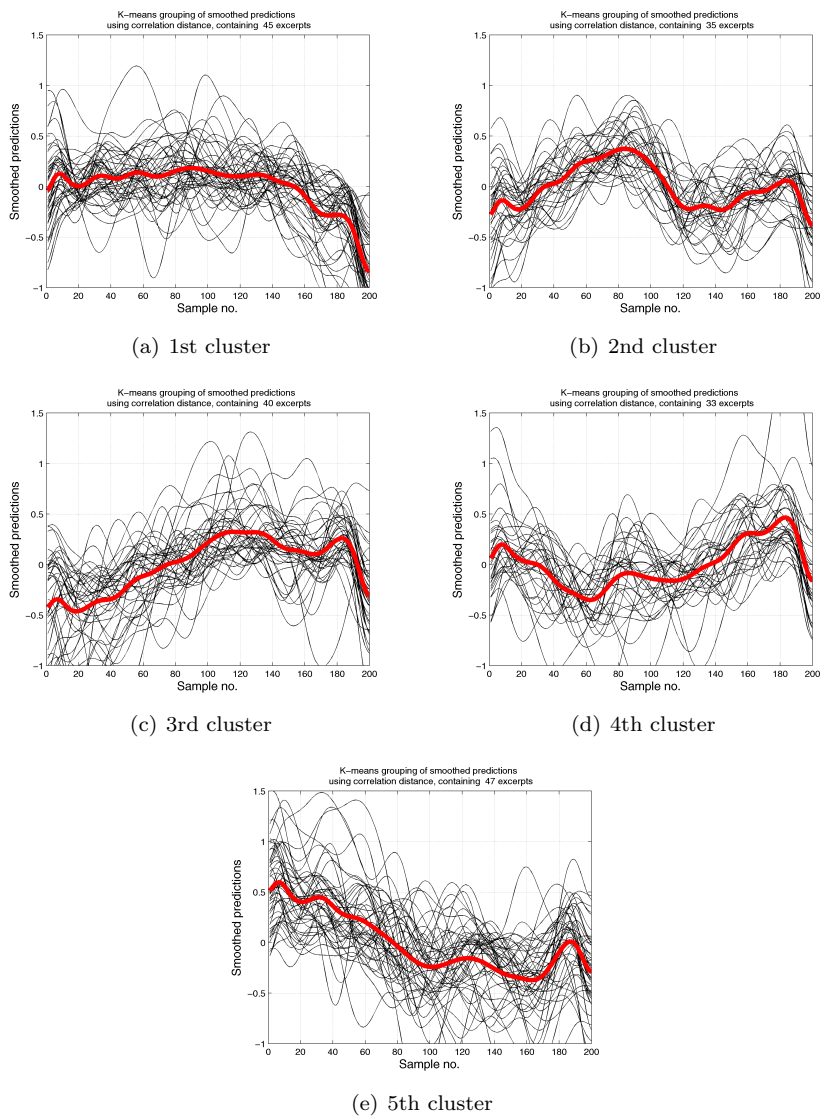


Figure D.6: 5 clusterings of emotional predictions using *KNN* on moving average smoothed temporal predictions using a *stepwise* regression model trained on *SFS* predicting beta mean coefficients of arousal. Red line indicates the mean of all excerpt within the individual clusters.

### D.5.2 Discussion

A simple method was used to group the emotional predictions exploratorially made by the *stepwise* regression model trained on *SFS* predicting beta mean coefficients for both valence and arousal, using unsupervised machine learning by *KNN*. 4 and 5 clusters were used for valence and arousal respectively where some structure could be found in the structure. It has to be said that some tendencies was found in the data, but differences between the maximum and minimum of each averaged curve for each clusters is relatively small. Whether this is the case in the real world, that music within 15 seconds changes in average 1 rating for some music could be. If the predictions are correct some excerpt change in arousal a great deal more of 3-5 ratings, where for valence the changes are not that great. This would also correspond to the general low variance predictions that was obtained in section 5.8.4. For the sake of grouping musical excerpts based on the temporal changes of emotional content expressed in music, using post-rating data, there was clearly some structure. Future research could look into verifying these results and look into what categorize these musical excerpts.



# Bibliography

---

- [Ali and Peynirciogly, 2006] Ali, S. O. and Peynirciogly, Z. (2006). Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music*, 34(4):511–534.
- [Barrett and Bliss-Moreau, 2009] Barrett, L. F. and Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, 41:167–218.
- [Barrett and Russell, 1999] Barrett, L. F. and Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, 8(1):10–14.
- [Bigand et al., 2005a] Bigand, E., Filipic, S., and Lalitte, P. (2005a). The time course of emotional response to music. *Annals New York Academy of Sciences*, 1060:429–437.
- [Bigand et al., 2005b] Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A. (2005b). Multidimensional scaling of emotional response to music: The effect of musical expertise and of the duration of excerpts. *Cognition and Emotion*, 19(8):1113–1139.
- [Bonnell et al., 2001] Bonnell, A., Faita, F., Peretz, I., and Besson, M. (2001). Divided attention between lyrics and tunes of operatic songs: Evidence for independent processing. *Perception and Psychophysics*, 63(7):1201–1213.
- [Bradley and Lang, 1994] Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25(1):49–59.
- [Chalupper and Fastl, 2002] Chalupper, J. and Fastl, H. (2002). Dynamic loudness model (dln) for normal and hearing-impaired listeners. *Acta Acustica United with Acustica*, 88:378–386.

- [Cohen, 2010] Cohen, A. J. (2010). *Music and Emotion: theory, research, applications*, chapter 31 - Music as a source of emotions in film, pages 879–908. Oxford; New York: Oxford University Press, 2010 edition.
- [Cross, 2009] Cross, I. (2009). *The Oxford Handbook of Music Psychology*, chapter 1 - The nature of music and its evolution, pages 3–13. Oxford University Press.
- [Duke and Colprit, 2001] Duke, R. A. and Colprit, E. J. (2001). Summarizing listener perceptions over time. *Journal of Research in Music Education*, 49(330):330–342.
- [Duxbury et al., 2003] Duxbury, C., Bello, J. P., Davies, M., and Sandler, M. (2003). Complex domain onset detection for musical signals. *Proceedings of the 6th International Digital Audio Effects (DAFx-03), September, 8-11, UK, London*.
- [Eerola et al., 2009] Eerola, T., Lartillot, O., and Toivianen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression. *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 612–626.
- [Efron et al., 2004] Efron, B., Hastie, T., and Johnstone, I. (2004). Least angle regression. *The annuals of Statistics*, 32(2):407–499.
- [Encyclopedia of Britannica, 2008] Encyclopedia of Britannica (2008). Encyclopedia of Britannica.
- [Evans and Schubert, 2006] Evans, P. and Schubert, E. (2006). Quantification of gabrielson’s relationship between felt and expressed emotions in music. *International Conference on Music and Perception and Cognition (ICMPC)*, 9th:446–454.
- [Gabrielsson, 2001] Gabrielsson, A. (2001). *Music and Emotion: theory and research*, chapter 19 - Emotions in strong experiences with music, pages 431–451. Oxford; New York: Oxford University Press, 2001 edition.
- [Gabrielsson, 2002] Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different? *Music Scientiae*, Special issue 2001-2002:123–147.
- [Gabrielsson and Juslin, 1996] Gabrielsson, A. and Juslin, P. (1996). Emotional expression in music performance: Between the performer’s intention and the listener’s experience. *Psychology of Music*, 24(1):68–91.
- [Hallam et al., 2009] Hallam, S., Cross, I., and Thaut, M. (2009). *The Oxford Handbook of Music Psychology*. Oxford University Press.



- [Harte et al., 2006] Harte, C., M.Sandler, and Gasser, M. (2006). Detecting harmonic change in musical audio. *1st ACM Workshop on Audio and Music Computing for Multimedia, AMCMM'06*, pages 21–26.
- [Herbert et al., 2008] Herbert, C., Junghofer, M., and Kissler, J. (2008). Event related potentials to emotional adjectives during reading. *Psychophysiology*, 45:487–498.
- [Hevner, 1936] Hevner, K. (1936). Experimental studies of the elements of expression in music. *American journal of Psychology*, 48(2):246–268.
- [Hu and Downie, 2007] Hu, X. and Downie, J. S. (2007). Exploring mood metadata: Relationship with genre, artist and usage metadata. *8th International Society for Music Information Retrieval Conference (ISMIR 2007)*.
- [Hu and Downie, 2010] Hu, X. and Downie, J. S. (2010). When lyrics outperform audio for music mood classification: A feature analysis. *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 619–624.
- [Hu et al., 2008] Hu, X., Downie, J. S., Laurier, C., Bay, M., and Ehmann, A. F. (2008). The 2007 mirex audio mood classification task: Lesson learned. *ISMIR 2008 - Session 4a - Data Exchange, Archiving and Evaluation*.
- [Juslin and Laukka, 2003] Juslin, P. and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129:770–814.
- [Juslin and Västfäll, 2008] Juslin, P. and Västfäll, D. (2008). Emotional response to music: The need to consider underlying mechanism. *Behavioral and Brain Sciences*, 31:559–621.
- [Kim et al., 2010] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnball, D. (2010). Music emotion recognition: A state of the art review. *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 255–266.
- [Kissler et al., 2007] Kissler, J., Herbert, C., Peyk, P., and Junghofer, M. (2007). Early cortical response to emotional words during reading. *Psychological Science*, 18(6):475–480.
- [Lartillot et al., 2008] Lartillot, O., Eerola, T., T., P., and Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation, and optimization. *International Conference on Music Information Retrieval, Philadelphia*.
- [Lartillot et al., 2010] Lartillot, O., Eerola, T., Toiviainen, P., and Fornari, J. (2010). Multi-feature modeling of pulse clarity: Design, validation and optimization. *ISMIR 2008 (Session 4c) Automatic Music Analysis and Transcription*, pages 521–526.

- [Laurier et al., 2009a] Laurier, C., Lartillot, O., Eerola, T., and Toiviainen, P. (2009a). Exploring relationships between audio features and emotion in music. *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009)*, pages 260–264.
- [Laurier et al., 2009b] Laurier, C., Meyers, O., Serrá, J., Blech, M., and Herrera, P. (2009b). Music mood annotator design and integration. *Seventh International Workshop on Content-Based Multimedia Indexing*, pages 156–161.
- [Laurier et al., 2009c] Laurier, C., Sordo, M., Serrá, J., and Herrera, P. (2009c). Music mood representation from social tags. *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 381–386.
- [Leman et al., 2005] Leman, M., Vermeulen, V., Voogdt, L. D., Moelants, D., and Lesaffre, M. (2005). Prediction of musical affect using a combination of acoustical structural cues. *Journal of New Music Research*, 34:39–67.
- [Lu et al., 2010] Lu, Q., Chen, X., Yang, D., and Wang, J. (2010). Boosting for multi-modal music emotion classification. *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 105–110.
- [Mathworks, 2010] Mathworks (2010). *Matrix Laboratory*. Mathworks, 2010b edition.
- [McCraty et al., 1998] McCraty, R., Barrios-Choplin, B., Atkinson, M., and Tomasino, D. (1998). The effects of different types of music on mood, tension and mental clarity. *Alternative therapies*, 4(1).
- [Meng et al., 2005] Meng, A., Ahrendt, P., and Larsen, J. (2005). Improving music genre classification by short-time feature integration. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5:497–500.
- [Müller, 2007] Müller, M. (2007). *Information Retrieval for Music and Motion*. Monograph, Springer.
- [Moore, 2004] Moore, B. C. J. (2004). *An introduction to the Psychology of Hearing*. Elsevier, 5th edition.
- [Moore et al., 1997] Moore, B. C. J., Glasberg, B. R., and Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc.*, 45(4):224–240.
- [Moore. et al., 1997] Moore., B. C. J., Glasberg, B. R., and Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240.

- [Mørup et al., 2008] Mørup, M., Madsen, K. H., and Hansen, L. K. (2008). Approximate 10 constrained non-negative matrix and tensor factorization. In *Accepted ISCAS 2008 special session on Non-negative Matrix and Tensor Factorization and Related Problems*.
- [Müller et al., 2009] Müller, M., Ewert, S., and Kreuzer, S. (2009). Making chroma features more robust to timbre changes. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872.
- [Müller et al., 2005] Müller, M., Kurth, F., and Clausen, M. (2005). Audio matching via chroma-based statistical features. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 288–295.
- [Nichols et al., 2009] Nichols, E., Morris, D., Basu, S., and Raphael, C. (2009). Relationship between lyrics and melody in popular music. *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*.
- [Palmer, 1997] Palmer, C. (1997). Music performance. *Annu. Rev. Psychol.*, 48:115–38.
- [Pampalk et al., 2002] Pampalk, E., Rauber, A., and Merkl, D. (2002). Content-based organization and visualization of music archives. *International Multimedia Conference: Proceedings of the tenth ACM international conference on Multimedia, December 01-06*.
- [Peeters, 2004] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. version: 1.0 (23. april 2004).
- [Posner et al., 2009] Posner, J., Russell, J., Gerber, A., Gorman, D., Colibazzi, T., S. Yo, Z. W., Kangarlu, A., Zhu, H., and Peterson, B. S. (2009). The neurophysiological bases of emotion: An fmri study of the affective circumplex using emotion-denoting words. *Human Brain Mapping*, 30(3):883–895.
- [Russell, 1980] Russell, J. A. (1980). A circumflex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- [Russell and Barret, 1998] Russell, J. A. and Barret, L. F. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, 74(4):967–984.
- [Russell and Barret, 1999] Russell, J. A. and Barret, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotions, dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819.

- [Russell and Carroll, 1999] Russell, J. A. and Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, 125(1):3–30.
- [Scherer and Oshinsky, 1977] Scherer, K. and Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1(4):331–346.
- [Scherer and Zentner, 2001] Scherer, K. L. and Zentner, M. R. (2001). *Music and Emotion: theory and research*, chapter 16 - Emotional effect of music: Production rules, pages 361–392. Oxford; New York: Oxford University Press, 2001 edition.
- [Scherer, 2003] Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256.
- [Schmidt and Kim, 2010] Schmidt, E. M. and Kim, Y. E. (2010). Prediction of time-varying musical mood distributions from audio. *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 465–470.
- [Schmidt, 2005] Schmidt, M. (2005). Least squared optimization with  $l_1$ -norm regularization. *CS542B Project report*.
- [Schubert, 1996] Schubert, E. (1996). Enjoyment of negative emotions in music: An associate network explanation. *Psychology of Music*, 24:18–28.
- [Schubert, 1999a] Schubert, E. (1999a). Measurement and time series analysis of emotion in music. PHD Thesis.
- [Schubert, 1999b] Schubert, E. (1999b). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, 51(3):154–165.
- [Schubert, 2010] Schubert, E. (2010). *Music and Emotion: theory, research, applications*, chapter 9 - Continuous Self-report methods, pages 223–253. Oxford; New York: Oxford University Press, 2010 edition.
- [Schubert et al., 2006] Schubert, E., Evans, P., and Rink, J. (2006). Emotion in real and imagined music: Same or different? *International Conference on Music and Perception and Cognition (ICMPC)*, 9th:810–814.
- [Sigurdsson et al., 2006] Sigurdsson, S., Petersen, K. B., and Lehn-Schiøler, T. (2006). Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. *7th International Conference on Music Information Retrieval (ISMIR 2006)*.
- [Sjöstrand, 2005] Sjöstrand, K. (2005). Matlab implementation of LASSO, LARS, the elastic net and SPCA. Version 2.0.

- [Sousou, 1997] Sousou, S. D. (1997). Effect of melody and lyrics on mood and memory. *Perceptual and Motor Skills*, 85(1):31–40.
- [Stevens and Byron, 2009] Stevens, C. and Byron, T. (2009). *The Oxford Handbook of Music Psychology*, chapter 2 - Universals in music processing, pages 14–23. Oxford University Press.
- [Turnbull et al., 2007] Turnbull, D., Barrington, L., and Langkriet, G. (2007). Five approaches to collecting tags for music. *ISMIR 2008 - Session 2c - Knowledge Representation, Tags, Metadata*, pages 225–230.
- [Tzanetakis and Cook, 2002] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5).
- [Yang et al., 2007] Yang, Y., Su, Y., Lin, Y., and Chen, H. (2007). Music emotion recognition: The role of individuality. *Proceedings of the international workshop on Human-centered multimedia*, pages 13–20.
- [Yang et al., 2008] Yang, Y.-H., Lin, Y. C., Su, Y.-F., and Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech and Language*, 16(2).
- [Zacharov and Bech, 2006] Zacharov, N. and Bech, S. (2006). *Perceptual Audio, Evaluation-Theory, Method and Application*. John Wiley and Sons.
- [Zentner and Eerola, 2010] Zentner, M. and Eerola, T. (2010). *Music and Emotion: theory, research, applications*, chapter 8 - Self-report measures and models, pages 187–222. Oxford; New York: Oxford University Press, 2010 edition.