Evaluation of Musical Feature Extraction Tools using Perpetual Ratings

ANTON HEDBLAD

KTH Computer Science
and Communication

Master of Science Thesis
Stockholm, Sweden 2011

Evaluation of Musical Feature Extraction Tools using Perpetual Ratings

A N T O N H E D B L A D

Master's Thesis in Music Acoustics (30 ECTS credits)
at the School of Media Technology
Royal Institute of Technology year 2011
Supervisor at CSC was Anders Friberg
Examiner was Sten Ternström

TRITA-CSC-E 2011:028
ISRN-KTH/CSC/E--11/028--SE
ISSN-1653-5715

Royal Institute of Technology
School of Computer Science and Communication

KTH CSC
SE-100 44 Stockholm, Sweden

URL: www.kth.se/csc

Evaluation of Musical Feature Extraction Tools Using Perceptual Ratings

Abstract

The increasing availability of digital music has created a demand for organizing and retrieving the music. Thus, a new multi-disciplinary research area called music information retrieval, MIR, has emerged. An important part of the content-based field of the research area is to extract musical features, such as tempo or modality, directly from the content, i.e. the audio.

This thesis is an evaluation of available musical feature extraction tools. The evaluation is done by using the extracted musical features as predictors for perceptual ratings in correlation and regression analyses. The 11 perceptual ratings were gathered from a listening test. 22 musical audio features were extracted from the same stimuli, using different systems for musical feature extraction. These were chosen to predict some of the perceptual ratings based on findings in literature.

High inter-subject reliability in the listening test implied a high agreement among the subjects, indicating that 20 subjects were enough. Fairly low inter-correlations between the ratings indicated that they were rated independently from each other. Six out of seven perceptual ratings with *a priori* selected predictors correlated moderately with their corresponding predictor ($r>0.6$). The results from the stepwise regression analyses were also moderate, where the amount of variance explained by the predictors ranged from 29-76%, indicating that there is room for improvements for developing new feature extraction algorithms.

Utvärdering av verktyg för insamling av musikaliska egenskaper med hjälp av perceptuella bedömningar

Sammanfattning

Den ökande tillgängligheten av digital musik har skapat ett behov av att organisera och hitta musiken. På grund av detta har ett nytt tvärvetenskapligt forskningsområde vuxit fram, kallat *music information retrieval* eller *MIR*. En viktig del av den innehållsbaserade grenen av detta forskningsområde är att extrahera musikaliska egenskaper, såsom tempo eller modalitet, direkt från innehållet, dvs. ljudklippet.

Den här uppsatsen är en utvärdering av tillgängliga verktyg för att extrahera musikaliska egenskaper. Utvärderingen är en jämförelse som görs genom att använda dessa extraherade musikaliska egenskaper som prediktorer för perceptuella bedömningar, med hjälp av korrelations- och regressionsanalyser. De elva perceptuella bedömningarna samlades in från ett lyssningsförsök. 22 musikaliska egenskaper extraherades från samma stimuli med hjälp av olika system för att extrahera egenskaper från musik. Dessa musikaliska egenskaper valdes ut för att predicera några av de perceptuella bedömningarna, baserat på litteratur.

Hög validitet mellan försökspersonerna i lyssningsförsöket tyder på enighet mellan bedömningarna, vilket antyder att 20 försökspersoner räckte. Ganska låga korrelationer tyder på att bedömningarna av olika egenskaper var oberoende av varandra. Sex av sju perceptuella bedömningar med på förhand utvalda prediktorer korrelerade måttligt med motsvarande prediktor ($r>0.6$). Resultatet från den stegvisa regressionsanalysen var också måttlig, där 29-76% av variansen kunde förklaras av prediktorerna, vilket visar att det finns rum för förbättringar av algoritmer för att extrahera musikaliska egenskaper.

Table of contents

1 Introduction

This chapter is an introduction to the thesis and includes the goal, the focus and a brief overview of the thesis.

Music information retrieval, MIR, is a new research field, emerged from the need of organizing the growing amount of digital music online and to find ways to retrieve information about the music.

There are two approaches on how to retrieve information from music; metadata-based and content-based MIR. The metadata-based approach retrieves textual metadata to describe the music, whereas the content-based approach retrieves information directly from the content, the music signal. This thesis will focus on the content-based approach. In this approach, the extraction of musical audio features is an important part. These features are extracted directly from the audio signal using different algorithms. Examples of musical audio features are spectral centroid, tempo and modality.

There are many systems for extracting musical audio features available. Many of those have been optimized from "ground truth" data, where music experts have been annotating music excerpts to be used as a reference. In this thesis, a chosen set of extracted musical audio features from different systems are instead compared to perceptual ratings, which are used as ground truth data. The perceptual ratings is data gathered from a listening test with 20 participants and a stimuli set of 100 music excerpts. These two representations are compared using statistical methods, including correlation and regression models.

1.1 Goal

The goal of this thesis is to evaluate how well feature extraction tools for music information retrieval predict perceptual ratings from listening tests.

1.2 Focus & delimitation

This study is a part of the research project SEMIR. The selection of perceptual features was done before the work on this thesis started. Hence, the musical features to be rated are not chosen exclusively for this thesis. The listening test was conducted by me as a part of this thesis.

The stimulus selection was also done before the work on this thesis started and was based on the results from a pilot study, where a listener test was conducted with 5 subjects and a stimulus set of 242 music excerpts.

The musical audio features extracted were chosen to predict perceptually rated features. Some of the rated features did not have any appropriate predictors in the systems for feature extraction used in this thesis. These were harmonic complexity, rhythmic complexity, dynamics and pitch. The function for estimating valence could not be run due to constraints in computing power and was left out as well. Even

though these did not have any predictors mentioned in the literature, they were kept in the test for exploratory reasons.

All of the feature extraction tools have been used with the default values, except when stated differently, see Table 2. All of the features have been extracted from audio files.

The tools for extracting musical audio features used in this thesis are from MIRToolbox, the Mazurka project, QM VAMP-plugins and the VAMP example plugins. It should be noted that there exist more tools for extracting musical audio features, such as Marsyas, jAudio, CLAM, ChucK, M2K, Psysound, IPEM toolbox, MA toolbox (Lartillot, 2008).

1.3 Overview

The thesis starts with a chapter giving a brief background to the research area music information retrieval, along with previous work within the field. Chapter 3 describes the listening test and the perceptual ratings. Chapter 4 describes the musical audio features and how they were extracted. In chapter 5, the results are presented, including correlations and regression analyses. Finally, in chapter 6 the results are discussed and conclusions are drawn from the results.

2 Background

This chapter provides a brief overview of the history of music information retrieval, the state of the art within the field, previous work and applications.

2.1 Music Information Retrieval

As the name implies, music information retrieval, MIR, is about retrieving information about music. MIR, a fairly new area of research, emerged in the late 1990s (Fingerhut, 2004). The research area is multidisciplinary, spanning over several fields of research, including computer science, psychology, statistics, music science and more.

The growing amount of available digital music created a demand for organizing the music. Thus, the emerge of MIR as a research area has been driven the demand of music consumers (Fingerhut, 2004). According to IFPI (2011), an organization representing the interests of the recording business worldwide, there were 13 million tracks licensed by record companies to digital music services worldwide in 2010. This can be compared to 2004, when IFPI (2005) reported "over a million tracks" available online . In six years, the number of available tracks has increased by a factor of approximately 10. This means that the need for methods for retrieving and cataloguing the content has increased.

2.1.1 State of the art

In 2000, ISMIR, the International Society for Music Information Retrieval, was formed (Byrd & Fingerhut, 2002). Since then an annual conference has been held by ISMIR, where proceedings are presented. Coupled to the ISMIR conference, the Music Information Retrieval Evaluation eXchange, MIREX, is organised. MIREX is a competition where algorithms can be tested independently on the same dataset. The aim of MIREX is to help sharing collections, data and code among the ISMIR society, avoiding issues with copyright and intellectual property right infringements (MIREX, 2011). On MIREX, state of the art algorithms are presented. To show how far the field of audio feature extraction has come, the best results from some of the tasks in MIREX 2010 are presented in Table 1. The results range from 0-1, where 1 represents a perfect result.

Task	**Metric**	**Best result**
Audio Onset Detection	Avg. F-measure	0.79
Audio Tempo Estimation	P-Score	0.81
Audio Key Estimation	Weighted Key Score	0.82
Query by Humming/Singing (Task 1a)	Mean reciprocal rank	0.95
Audio Tag Classification - Mood	Class. F-Measure	0.47
Audio Tag Classification - Major Minor	Class. F-Measure	0.48

Table 1. *The best results from some of the tasks in MIREX 2010 (Downie & IMIRSEL, 2010)*

2.1.2 Applications of MIR

There are several commercial systems for retrieving digital music online. One example is Pandora[1] which is based on metadata. The user selects an artist or song and the system plays that artist or song and continues as a radio player with similar music (Casey et. al., 2008). Track and artist similarity measures are derived from detailed metadata. The metadata has been added by hand by music professionals. It is estimated that it takes 20-30 minutes per track to input the metadata. This means it would take approximately 50 person-years to enter metadata for one million songs. For the 13 million songs currently available online, it would take 650 person-years to add metadata to all of them, thus it would be extremely costly and time consuming. A similar system is Last.fm[2], which works in a similar way, but relies on metadata generated by the users instead of by experts. This will keep costs down, but the metadata might not be as reliable as the one added by musical professionals.

Another commercial system for retrieving music online is Shazam[3], where the user records a small piece of the track which is sent to the Shazam server. The server then analyses the music excerpt and gives the user information about the track, such as the name of the song and the artist. This is also known as query-by-example, QBE (Wang, 2006). Shazam is not metadata-based like Pandora, but content-based (Casey et. al., 2008). Thus, Shazam does not need the user to have knowledge about the artist. Instead the user inputs an excerpt of the track the user is looking for. A similar system is Midomi[4], which is a query-by-humming system. Here, the user does not need to record the actual track. Instead the user hums or whistles the song and retrieves information about the song.

There are also other interests in the field of MIR from the music industry. For example, attempts have been made to automatically predict hit songs (Dhanaraj & Logan, 2005). There are some commercial systems for this, for example uPlaya[5], who claim they have achieved a successful system for detecting hits, even though this is contradicted by Pachet & Roy (2008) who claim it is impossible to extract a songs potential popularity automatically with current methods.

Other applications for MIR include systems for automatic genre classification, beat tracking and matching, automatic transcription etc.

2.2 Feature extraction

In content-based MIR-systems, the audio features play an important role. There are a large number of audio features that can be extracted, see for example Peeters (2004) and Burred & Lerch (2004).

[1] http://www.pandora.com, note that this service is restricted outside of the US.

[2] http://www.last.fm/

[3] http://www.shazam.com

[4] http://www.midomi.com

[5] http://uplaya.com/

The extracted features can be divided into three different levels of abstraction. The distinctions are vague, but it is done to show the difference in complexity.

2.2.1 Low level

McKinney & Breebaart (2003) divide the low level features into three groups. The low level signal parameters, MFCC parameters and psychoacoustic features.

The low level signal parameters include RMS level, spectral centroid, bandwidth, zero-crossing rate, spectral roll-off frequency, band energy ratio, delta spectrum magnitude (called spectral flux in this report), pitch, and pitch strength.

MFCC means Mel Frequency Cepstrum Coefficients and is retrieved by taking the discrete cosine transform of the logarithm of the Mel magnitude spectrum (Casey et al, 2008). MFCCs are used to organize sinusoidal modulations of spectral magnitudes, in an array. The first values of the array correspond to long wave spectral modulation. Values in the end of the array correspond to harmonic components in the spectrum. Usually, the first 20 coefficients are used, thus representing formant peaks of the spectrum, which correspond to the timbre of the music.

The psychoacoustic measures mentioned in McKinney & Breebaart (2003) are roughness, loudness and sharpness. Roughness is the perception of frequency modulations of the temporal envelope, ranging from 20-150 Hz and corresponds to musical dissonance. Loudness is the perceived signal strength. Sharpness is a measure of how much high-frequency energy relative to low-frequency energy there is.

Many of those features are defined by their method and can only be computed in one or a few ways. The zero-crossing rate, for example, calculates how many times the signal crosses the zero line and can only be calculated in one way; by counting the number of zero-crossings and divide the sum by a time unit.

The psychoacoustic measures can be calculated in different ways, using different psychoacoustic models. There are numerous ways to calculate the loudness, see for example (Nygren, 2009).

Pitch is an exception, since extracting the pitch from polyphonic music, music with multiple independent melodic voices, is complex. This is because there might be multiple fundamental frequencies present at the same time. Hence, pitch might fit better as a mid level feature.

2.2.2 Mid level

Mid level features are often derived from concepts of music theory, such as modality, tempo and articulation. These features are often established from ground truth data, where music clips have been labelled by professional musicians. These labels are then used as reference when optimizing the algorithms. As opposed to the low level features, it is not as clear how to extract these features. Different algorithms are developed and tested and progress is still being made.

2.2.3 High level

The high level features are often semantic descriptions, such as emotions or other adjectives, like flowing, static or void (Lesaffre et. al., 2006). These features are usu-

ally estimated by a combination of low and mid level features. For example, Eerola et. al. (2009) use 18 low and mid level features to predict emotion descriptions.

2.3 Perceptual ratings of features

Previous literature on the subject of perceptual ratings of musical features is sparse. Lartillot et. al. (2008) conducted a listener experiment to control the validity of a model used to predict pulse clarity. The result from the listener test was used as the dependent variable in a step-wise multiple linear regression model, where best predictors for the model was acquired. This regression model obtained an adjusted R^2 of 0.76, meaning it explained 76% of the variability of the subjects' ratings. 25 subjects participated in a listener test with 100 music excerpts where the pulse clarity was rated. The inter-subject reliability was very good, with a Cronbach's alpha coefficient of 0.971.

Wedin (1972) reports about an experiment where 15 musical experts, each with at least 1.5 years of musical education, rated similar features on forty musical excerpts. The 13 features, or as Wedin calls them, "musico-technical aspects", were:

- Strength or intensity (pp-ff)
- Pitch (bass-treble)
- Rhythm (outstanding-vague rhythm)
- Pulse rate (fast-slow)
- Rhythmic articulation (firm-fluent) and (staccato-legato)
- Tempo (fast-slow)
- Harmony (dissonant, complex versus consonant, simple)
- Tonality (atonal-tonal)
- Modality (major-minor)
- Melody(melodious, singable versus "unmelodious")
- Type of music ("serious"-"popular")
- Style expressed in terms of date and composition

In the study conducted by Wedin, the inter-subject reliability estimated by Ebel's intraclass correlation is good, with values ranging from 0.90-0.99. Some of the similar features were highly correlated, tempo and pulse rate correlated with the correlation coefficient r=0.98 and harmony and tonality with r=0.95.

Wedin uses these ratings as dependent variables in a multiple stepwise regression analysis with three perceptual-emotional dimensions found in previous experiments as dependent variables. Since some features were highly correlated, there were some problems with the results.

3 Listening test

This chapter includes a theory part about the perceptual ratings and a method part about the listening test. The result from the test is presented in the chapter 5.

3.1 Perceptual ratings

In the listening test the subjects rated a total number of eleven features, where nine were perceptual musical features and two were emotional descriptions. The selection of perceptual features to rate was done for the SEMIR project, a series of studies where features from different fields, such as emotion research and ecological perception, are studied. For more information on the selection, see Friberg et. al. (2011).

3.1.1 Perceptual features

Speed (slow-fast)

The perceived speed of the music. No deep analysis regarding the tempo should be done.

Rhythmic clarity (flowing-firm)

The subjects were asked to indicate how well the rhythm is defined, i.e. how emphasized the rhythmic parts of the music are. This is similar to the features *rhythm* (outstanding-vague rhythm) and *rhythmic articulation* (firm-fluent) in Wedin (1972).

Rhythmic complexity (simple-complex)

A rating of how complex the rhythm is. A simple 4/4 rock beat would be rated as simple, whereas a more unconventional time signature such as 5/4 or 7/8 would be rated as a more complex rhythm pattern. Other factors, such as syncopation, would also contribute to a more complex rhythm.

Articulation (legato-staccato)

The overall articulation of the music excerpt, ranging from *legato* to *staccato*. Wedin (1972) uses the same scale, but calls it rhythmic articulation (staccato-legato).

Dynamics (soft-loud)

The overall perceived dynamic level of the music excerpt. This corresponds to pianoforte in music theory. Wedin (1972) uses a similar scale, but calls it *intensity* (pp-ff).

Overall pitch (low-high)

The overall perceived pitch of the whole music clip. This is used by Wedin (1972), with the exception that Wedin used another scale, with bass-treble as extreme values instead of low-high.

Modality (minor-major)

The overall modality over the whole clip on a scale ranging from minor to major. Note that it was rated on a gradual scale ranging from 1-9. This is one of the most important features for recognizing emotions, according to a number of studies (Gabrielsson & Lindström, 2010). Wedin (1972) used the same feature.

Harmonic complexity (simple-complex)

The complexity of the harmony has been used in several studies to predict emotions, i.e. Hevner (1937), *harmony* (simple-complex) and Wedin (1972), *harmony* (dissonant, complex – consonant, simple). This is one of the most demanding features to rate, since the subject needs to have some knowledge about music theory. Thus, there might be some inconsistency between subjects with different amounts of knowledge within the field of music theory.

Brightness (dark-bright)

The overall perceived brightness of the music excerpt. This is the only timbral measure used.

3.1.2 Emotional ratings

In the pilot study, the four quadrants from the energy-valence space, as mentioned by Russell (1980), were used as discrete emotional features. These were labelled *Happiness, Anger, Sadness, and Tenderness*. Due to constraints regarding the number of features to be rated, these four discrete emotional features were reduced to the axes of the aforementioned space, namely *energy* and *valence*.

Energy (low-high)

The overall perceived energy of the music excerpt was to be rated by the subjects.

Valence (negative-positive)

The overall perceived valence of the music excerpt. Valence is a rating of how positive or negative a feeling is. Positive valence is associated with emotions such as joy, love or tenderness, while negative valence is associated with emotions such as anger, fear or sadness.

3.2 Method

3.2.1 Software

The software used to create and run the listening experiment is called *Skatta*[6] and was created as part of a master thesis at KTH (Lindberg, n.d.). The software presents one screen per stimulus. These presentation screens can be customized either by using the built in editor in Skatta or by editing the corresponding XML file. Due to

[6] Skatta is freely available at: http://sourceforge.net/projects/skatta/

the experimental design, this test was made by customizing the XML file since it gave more control and flexibility. Since the eleven features could not fit into the default layout with a single column, a two column layout was made. This caused the alignment to break, but since Skatta is written in Java and utilizes the Swing GUI API, HTML-code could be inserted to align the text, using letters with the same colour as the background.

One crash was experienced, probably caused by storing the results on a remote server (AFS). After the crash, the results were stored locally instead.

3.2.2 Hardware

The test was run on a PC with Microsoft Windows XP Professional using an M-Audio Delta 1010LT sound card. In the listening room the subject sat in front of a computer monitor where the test was displayed and used a mouse and a keyboard as input units.

The subject was placed approximately two meters from a pair of Genelec 1031A active studio monitors. These were connected to the soundcard of the PC with a Behringer Eurorack MX 602A mixer in the signal path. The room used for the experiment was soundproofed and the computer was located in another room to avoid noise in the experiment room.

Due to limited access to the primary listening set-up an additional room with an almost identical set-up was used. The only things that differed in the hardware in the additional room were the soundcard, E-MU 0404 PCI, and the mixer, a Phonic AM220. This room was slightly more reverberant and the computer was in the same room as the subject, adding a bit of noise.

3.2.3 Calibration

To calibrate the sound level of the experiment, an Ono Sokki LA-210 sound level meter was used. During the calibration, a clip of a filtered white noise signal (single-pole low-pass filter, $H(z)=1/(1 - 0.95z^{-1})$ with a spectral slope of -6 dB per octave was played, using the same software as the test was played through in order to use the same signal path as in the experiment. This noise signal was normalized and corresponded roughly to the spectral content of the stimuli. While playing the noise, the output level of the mixer was adjusted so that the sound level corresponded to 75 dB(A) at the listening position, measured with the sound level meter.

3.2.4 Subjects

There were a total number of 20 subjects, 7 female and 13 male, with an average age of 30 years (range 18-55). All subjects answered a questionnaire, including information about themselves and their music experience. The majority of the subjects were students at KTH. All of the subjects had at least some experience of playing a musical instrument, which was a prerequisite for participating. The average years of playing their main instrument was 14 years (range 3-45). The subjects reported that they listened to music 15 hours per week in average (range 3-40). Six of the subject reported hearing impairments in the questionnaire. Since the impairments were all

minor they were not considered critical for the test. Hence all of the subjects were kept in the study. Only one person reported absolute pitch. All but two subjects were Swedish citizens, two of the Swedish citizens reported dual citizenship.

3.2.5 Stimuli

The stimuli consisted of 100 songs selected from a database of 242 ringtones in MIDI-format, mostly instrumental polyphonic versions of popular music. These MIDI files were played through a Roland JV-1010 MIDI synthesizer and recorded as waveform audio. To normalize the audio files, the loudness standard specification *ITU-R BS. 1770* was used (ITU-R, 2006). The tracks were then trimmed to have a constant delay before the first onset. The average length of the stimuli was 30 seconds and stimuli longer than 40 seconds were cropped, using the original MIDI file to make sure the ending sounded natural. All of the 242 stimuli were prepared and used in a pilot study. The results from this pilot study were used in the selection of the 100 stimuli used in this experiment. For more information, see Friberg et. al. (2011).

3.2.6 Procedure

As mentioned above, the subjects answered a questionnaire asking for some personal information and some information regarding the subjects' musical experiences.

The test consisted of 100 stimuli that were presented in a random order for each subject. For each stimulus, the eleven features were to be rated by the test subjects.

All of the features were rated on a 9-point Likert scale, represented by 9 radio buttons. The end positions were labelled and represented the extreme values for each feature, see Figure 1.

The stimuli were presented in a random order. All features had to be rated before the subject could go to the next stimulus. The subjects were able to take breaks freely, and automatic break reminders were inserted into the test in order to encourage the subject to take breaks. After each 26 stimuli, the software reminded the subjects to take a break; hence there were three breaks in the test which contained 100 stimuli. The average time for the test was 1 hour and 44 minutes (range 01:01-02:38). When the test was finished, the results were stored in a predefined folder as a comma separated value-file. As reimbursement, the subjects received two cinema gift cards after the test.

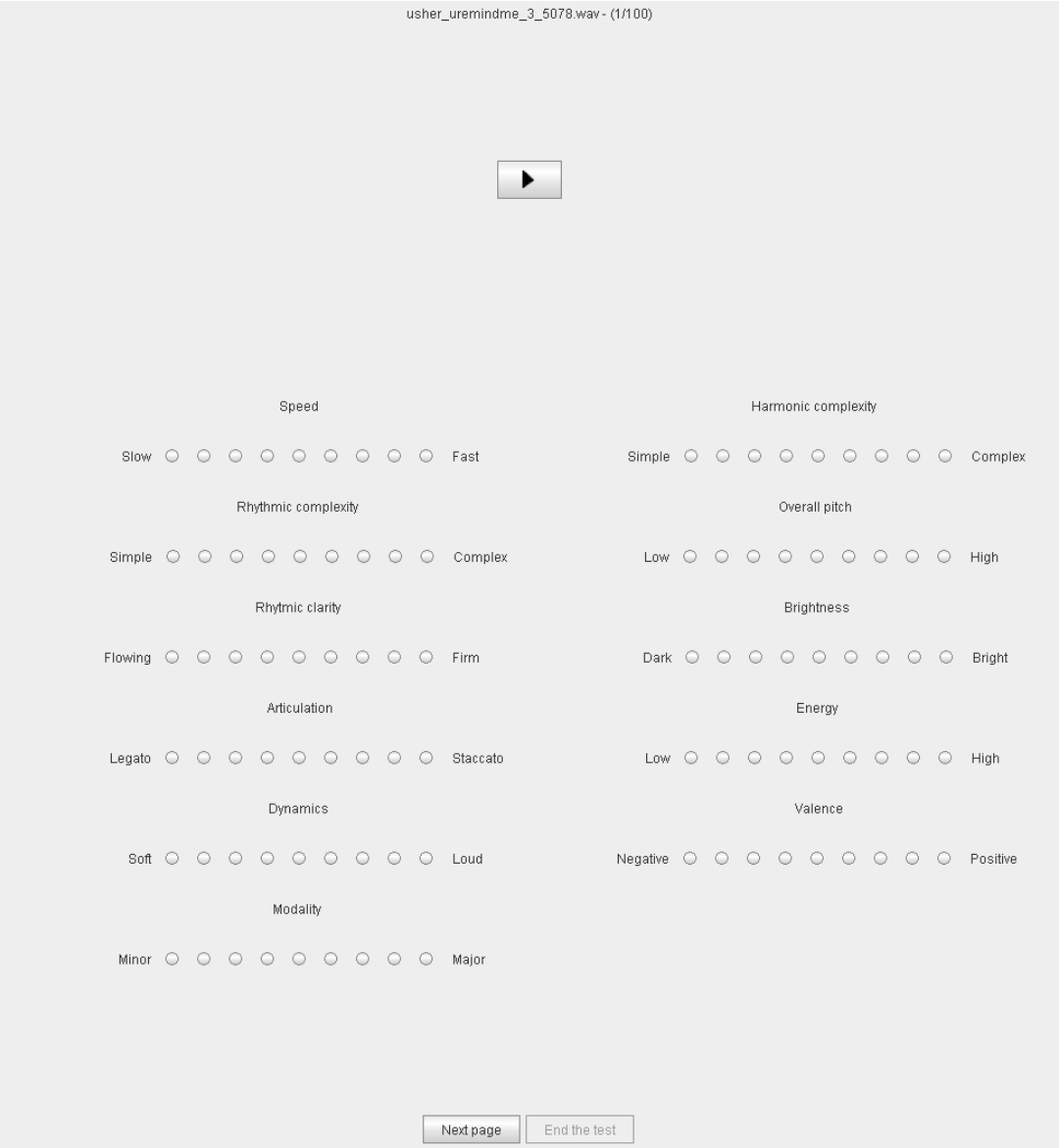

Figure 1. *Screenshot of the listening test screen. The image has been cropped to fit the page.*

4 Extracted audio features

This chapter provides an overview of and details about the extracted features.

Two freely available hosts were used for extracting the audio features; MIRToolbox[7] 1.3.1 and Sonic Annotator[8] 0.5. A number of other systems were considered, but since they could not contribute with more mid level features, they were not used. The low level features they could provide, which were interesting for this study, were already present in MIRToolbox.

MIRToolbox is a toolbox in MATLAB, developed at the University of Jyväskylä. For more information, see Lartillot & Toiviainen (2007). Sonic Annotator is a host program which can run VAMP audio analysis plugins and is developed by Chris Cannam, Mark Levy and Chris Sutton within the omras2 project[9] at the Queen Mary, University of London and Goldsmiths, University of London. The VAMP plugins used were QM VAMP plugins[10], VAMP Example plugins[11] and the SV Mazurka plugins[12], which is a set of functions ported to VAMP from the Mazurka project[13]. All of these plugins are freely available[14].

The features were selected to predict the perceptual ratings. This selection is based on the literature about feature extraction.

The extracted audio features will be referred to with an identifier throughout the report. They can be found in Table 2, where information about the input parameters and the commands used for extracting the features also can be found.

4.1 Low-level audio features

The following part will give a brief explanation of all extracted musical audio features. For each feature, there is a discussion about which perceptual rating the feature might predict.

[7] MIRToolbox is freely available at: https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[8] Sonic Annotator is freely available at: http://www.omras2.org/SonicAnnotator

[9] http://www.omras2.org/

[10] http://www.vamp-plugins.org/plugin-doc/qm-vamp-plugins.html

[11] http://www.vamp-plugins.org/plugin-doc/vamp-example-plugins.html

[12] http://sv.mazurka.org.uk/download/

[13] http://www.mazurka.org.uk/

[14] http://www.vamp-plugins.org/download.html

4.1.1 Spectral Centroid

MT_SC

This is the "centre of mass" of the spectrum, calculated as a weighted mean of the frequencies present in the audio signal. The spectral centroid (SC) has been found to predict the timbral brightness (Schubert et. al., 2008). The SC for each frame *n* is calculated as

$$SC(n) = \frac{\sum_{k=0}^{K-1} f(k)x(k)}{\sum_{k=0}^{K-1} x(k)},$$

where *f(k)* represents the magnitude of bin number *k* and *x(k)* represents the centre frequency of bin number *k* in the frequency domain. K is the number of bins.

4.1.2 Zero Crossing Rate

MT_ZCR

Zero crossing rate, or ZCR, is a measurement of how many times the audio waveform crosses the x-axis, i.e. changing sign from positive to negative, or back. It is calculated as

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} f\{s_t s_{t-1} < 0\},$$

where s is the time signal of length T and the indicator function $f\{A\}$ is equal to 1 if A is true and equal to 0 if A is false.

In a study made by Alluri & Toiviainen (2009) on the perception of timbre in a polyphonic context, eight bipolar timbre scales were reduced to the three dimensions *activity* (soft-hard, strong-weak, high energy-low energy), *brightness* (colourless-colourful, dark-bright) and *fullness* (empty-full) using factor analysis. In this study brightness (dark-bright) was chosen as the only feature related to timbre. A correlation between ZCR and the timbral feature Brightness, with a correlation of r=0.4, was reported.

4.1.3 Brightness

These are all features selected for predicting the rated brightness. Apart from brightness, music with large amounts of high frequency energy might also predict perceptual energy, according to Gabrielsson & Lindström (2010).

MT_Bright_1k, MT_Bright_1.5k & MT_Bright_3k

MIRToolbox estimates brightness by calculating how much spectral energy there is above a threshold frequency value and divides it by the total energy to obtain the ratio. This threshold value has been discussed; the default value is 1500 Hz, while Laukka, Juslin and Bresin (2005) suggest a value of 1000 Hz. A value of 3000 Hz was also suggested by Juslin (2000). For comparison reasons, all three threshold values were used in this study.

Identifier	**Meaning**	**Parameters**	**Host**	**Command**
EX	*Prefix for features extracted with the VAMP Example plugins*			
EX_Onsets	Percussion Onsets	Default	Sonic Annotator	vamp:vamp-example-plugins:percussiononsets:onsets
EX_Tempo	Tempo	Default	Sonic Annotator	vamp:vamp-example-plugins:fixedtempo:tempo
MT	*Prefix for features extracted with MIRToolbox*			
MT_ASR	Average Silence Ratio	Default	MATLAB	mirlowenergy(fname,'ASR')
MT_Bright_1.5k	Brightness	Default	MATLAB	mirbrightness(fname)
MT_Bright_1k	Brightness	Cut-off: 1000 Hz	MATLAB	mirbrightness(fname,'CutOff',1000)
MT_Bright_3k	Brightness	Cut-off: 3000 Hz	MATLAB	mirbrightness(fname,'CutOff',3000)
MT_Event	Event density	Default	MATLAB	mireventdensity(fname)
MT_Mode_Best	Modality	Model: Best	MATLAB	mirmode(fname,'Best')
MT_Mode_Sum	Modality	Model: Sum	MATLAB	mirmode(fname,'Sum')
MT_ Pulse_Clarity_1	Rhythmic Clarity	Model: 1	MATLAB	mirpulseclarity(fname,'Model',1)
MT_ Pulse_Clarity_2	Rhythmic Clarity	Model: 2	MATLAB	mirpulseclarity(fname,'Model',2)
MT_SC	Spectral Centroid	Default	MATLAB	mircentroid(fname)
MT_SF	Spectral Flux	Default	MATLAB	mirflux(fname)
MT_ Tempo_Auto	Tempo	Model: Autocorr	MATLAB	mirtempo(fname)
MT_ Tempo_Both	Tempo	Model: Autocorr & Spectrum	MATLAB	mirtempo(fname,'autocor','spectrum')
MT_ Tempo_Spect	Tempo	Model: Spectrum	MATLAB	mirtempo(fname,'spectrum')
MT_ZCR	Zero Cross-ing Rate	Default	MATLAB	mirzerocross(fname)
MZ	*Prefix for features extracted with VAMP plugins ported from the Mazurka project.*			
MZ_SF_Onsets	Spectral Flux Onsets	Default	Sonic Annotator	vamp:mazurka-plugins: mzspectralflux:spectralfluxonsets
MZ_SRF_ Onsets	Spectral Reflux Onsets	Default	Sonic Annotator	vamp:mazurka-plugins: mzspectralreflux:spectralrefluxonsets
QM	*Prefix for features extracted with VAMP plugins from Queen Mary.*			
QM_Mode	Modality	Default	Sonic Annotator	vamp:qm-vamp-plugins:qm-keydetector:mode
QM_Onsets	Onset detection	Default	Sonic Annotator	vamp:qm-vamp-plugins:qm-onsetdetector:onsets
QM_Tempo	Tempo	Default	Sonic Annotator	vamp:qm-vamp-plugins:qm-tempotracker:tempo

Table 2. *Overview of extracted audio features.*

4.1.4 Spectral flux

MT_SF

This is a measurement of the rate of changes in the power spectrum of a signal. It is usually measured by calculating the Euclidian distance between successive frames of the power spectrum. Each frame n is calculated as

$$SF(n) = \sqrt{\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \left(|X(n,k)| - |X(n-1,k)| \right)^2},$$

where X(k,n) is the discrete Fourier transform of frame n and bin k of the audio clip, using a Hamming window. N is the frame size in frames. The frame length used is the default value in MIRToolbox, which is 0.2 seconds with a hop factor of 1.3, making N=8820 at a sampling rate of f_s=44100 Hz.

Alluri and Toiviainen (2009) reported a correlation between spectral flux and a factor they labelled Activity. It was derived from a factor analysis of the results of a listener test where the subjects rated adjectives. In this study, it may be correlated to the emotional feature energy, which is similar to activity.

Spectral flux has also been used for onset detection (Dixon, 2006; Bello et al, 2005). When using spectral flux for note onsets, it is usually half-wave rectified, looking only at positive numbers.

4.1.5 Average Silence Ratio

MT_ASR

The articulation, ranging from legato to staccato, can be roughly estimated by calculating the average silence ratio, ASR. A higher ASR value means there is relatively more silence in the audio clip, which might indicate staccato. It is calculated as the percentage of frames in the clip where the energy is less than half of the mean square energy of the whole clip (Lartillot, 2010).

4.2 Mid-level audio features

4.2.1 Tempo

These features are selected for predicting the perceptual ratings of speed. According to Gabrielsson & Lindström (2010), speed is an important part of the prediction of the emotional feature energy. Hence, the features estimating tempo might also correlate with the perceptual rating of energy.

MT_Tempo_Auto, MT_Tempo_Both & MT_Tempo_Spect

There are several models for computing the tempo of a music file. With MIRToolbox there are two methods. Both starts by computing the onset detection

curve. The default algorithm, MT_Tempo_Auto, computes an autocorrelation function of the onset detection curve and estimates the tempo from picking the peaks of the autocorrelation curve. The other method, MT_Tempo_Spect, uses a spectral transformation of the onset detection curve, picking the frequency of the highest magnitude within the predefined tempo limits. There is also an option to use both methods, referred to as MT_Tempo_Both in this study, where the autocorrelation function is translated into the frequency spectrum and multiplied to the spectrum curve subsequently.

QM_Tempo

The VAMP plugin QM_Tempo was used with the default beat tracking method, which is called *new*. This is a hybrid between the old method and a new one. The old method uses a two-state beat tracking model, described in Davies & Plumbley (2007). The new method uses a dynamic programming method, described in Ellis (2007). The onset detection algorithm is a complex domain method with a combined energy and phase approach described in Duxbury et. al. (2003)

EX_Tempo

EX_Tempo is a function for detecting tempo in songs with fixed tempo. Since almost all of our stimuli has fixed tempo this function should be applicable. Only a given number of seconds are evaluated, counted from the start of the audio clip. In this experiment, the default value of 10 seconds is used for analysing, meaning that the first 10 seconds of each clip is analysed. The function calculates an overall energy rise function over a series of short frequency-domain frames and then calculates the autocorrelation over them. Then it calculates the tempo from the autocorrelation function using a peak picking algorithm. This method builds on the work by Matthew Davies and is a simplification of a small part of the work described in Davies & Plumbley (2005). It is similar to MT_Tempo_Auto, which also uses an autocorrelation function.

4.2.2 Event Density and Onset Detection algorithms

Event density algorithms estimate the number of note onsets per second. This is done by using an onset detection algorithm. Several algorithms have been used for comparison. These features are selected for predicting the perceptual ratings of speed. They could also be used to predict energy, according to Gabrielsson & Lindström (2010). See also section 4.2.1 Tempo. MT_Event is the only algorithm which calculates the event density directly. For the other functions used, the number of onsets found is divided by the length of the music clip, thus giving an average event density over the whole clip.

MT_Event

This algorithm is based on the work by Klapuri (1999). The signal is divided into 21 non-overlapping frequency bands using a filterbank. For each band, the logarithm of the envelope function of the signal is differentiated, creating a relative distance function. From this function, peaks above a threshold value are picked as onset candi-

dates, creating 21 vectors with onset candidates, one vector for each frequency band. These are combined and onsets below a threshold value are disregarded. The values left in the vector are considered onsets. These can then be counted and the sum is divided by a time unit, resulting in the event density.

QM_Onsets

The VAMP plugin QM_Onsets is used with the complex domain method, an approach that is combining both energy and phase information. The algorithm presented in Duxbury et al (2003).

EX_Onsets

The VAMP plugin EX_Onsets uses a novel percussive feature detection function to separate the percussive elements from the music clip, described in Barry et al (2005). After the percussive elements are separated, an onset detection algorithm using short-time magnitude spectrum and scaling to an estimated time-amplitude function is used.

MZ_SF_Onsets & MZ_SRF_Onsets

MZ_SF_Onsets calculates the onsets from spectral flux, i.e. changes in the spectral energy. This method is described in Dixon (2006).

Information about the other onset detection function from the Mazurka project used, MZ_SRF_Onsets, could not be retrieved.

4.2.3 Rhythmic clarity

MT_Pulse_Clarity_1 & MT_Pulse_Clarity_2

The Pulse clarity estimation in MIRToolbox corresponds closely to the perceptual rating rhythmic clarity in this study. The methods used are the two methods that were found to be optimal by Lartillot et al. (2008). The first method uses a spectral decomposition and the second uses a filterbank to decompose the signal before the onset detection function. The autocorrelation is then calculated, whence the pulse clarity is calculated.

4.2.4 Modality

Obviously, the modality estimating features are selected for predicting the rated modality. According to Gabrielsson & Lindström (2010), the modality can also be used as a part of the prediction of the emotional feature valence.

MT_Mode_Best & MT_Mode_Sum

In the MIRToolbox, the modality is extracted by examining computed key strengths. using two different methods. The default method, MT_Mode_Best, compares the best major and minor key. The other method, MT_Mode_Sum, compares the sum of the key strengths for the minor and major keys.

The key strengths are calculated by taking the cross-correlations of the wrapped and normalized chromagram (Krumhansl, 1990; Gómez, 2006). The chromagram is

calculated by analyzing the spectrum of the 20 highest decibels of the audio file. This spectral energy is then redistributed into frequency bins representing the keys in western music. This representation is called a chromagram. The bins representing the same key in different octaves (multiples of two) are added together, resulting in 12 wrapped bins in the chromagram. This chromagram is then normalized before the key strengths are calculated.

The output for these functions is a decimal value between 0-1, where 0 represents minor and 1 represents major.

QM_Mode

Estimates the key by comparing how a block-by-block chromagram correlates with stored key profiles for every key, as described in Noland & Sandler (2007). The key profiles are taken from analyses of a recording of Book I of the Well Tempered Klavier by J. S. Bach, at A=440 equal temperament.

The results are presented binary for different parts of the file, if the key changes. If the first five seconds of a clip is in minor key and the rest is in major key, the results are presented as blocks:

```
0.000000000,1,"Minor"
5.000000000,0,"Major"
```

In this experiment, a weighted decimal mean is calculated with the durations of each block as weights. If this example clip would be 20 seconds long, this would result in:

$$\frac{(5.0-0.0)*1+(20.0-5.0)*0}{20}=0.25$$

Note that the output from this method, where major=0 and minor=1, is the inverted in relation to the modality estimating functions in MIRToolbox, where major=1 and minor=0.

5 Results

The first part of this chapter includes the results for the perceptual ratings and the extracted musical audio features. The second part is a comparison consisting of correlations and regression analyses.

5.1 Perceptual ratings

5.1.1 Inter-subject reliability

The Cronbach's alpha coefficient, a measure of reliability, was calculated in order to investigate the inter-subject reliability. George and Mallery (2003) provides rules of thumb for the Cronbach's alpha coefficients, considering values greater than 0.9 as "excellent" and values greater than 0.8 as "good" (p. 231). As shown in Table 3 all of the calculated coefficients are greater than 0.8 and a majority is greater than 0.9. This means that the inter-subject reliability can be considered as very good. The perceptual rating with the lowest coefficient was harmonic complexity. This was expected to be relatively low since the task was considered more difficult, assuming knowledge in music theory.

Since the inter-subject reliability was high, implying that the ratings were homogenous, the mean of the ratings can be used for the following analyses, as in Lartillot et. al. (2008).

Feature	**Cronbach's alpha**
Speed	0.98
Rhythmic Complexity	0.89
Rhythmic Clarity	0.90
Articulation	0.93
Dynamics	0.93
Modality	0.93
Harmonic Complexity	0.83
Pitch	0.93
Brightness	0.88
Energy	0.96
Valence	0.94

Table 3. *Inter-subject reliability*

5.1.2 Correlations between perceptual ratings

The pairwise correlation coefficients, denominated with *r* throughout the report, between all perceptual ratings are shown in Table 4. According to Williams (1968) a correlation coefficient between 0.4-0.7 should be considered a substantial relationship and coefficients between 0.7-0.9 should be considered a marked relationship.

This may be used as guidance on how to interpret the correlation coefficients. High correlations reveal dependencies between the extracted audio features, which can result in problems in the following statistic analyses.

The two emotional ratings, energy and valence, are high level features and are expected to depend on combinations of the other ratings. Regression analyses with the emotional ratings as dependent variables and all perceptual ratings as independent variables can be found in Friberg et. al. (2011). In these regression analyses, the emotions could be predicted by the other ratings to a very high degree. 93.1% of the variance of energy and 87.1% of the variance of valence could be explained by combinations of perceptual ratings. Thus, the correlations for the emotional ratings are highlighted with grey background colour in Table 4.

High correlations between the perceptual ratings could be explained by two mechanisms. The first one is that the test subjects could not differentiate between the features. For example songs in a major key might be perceived as brighter than songs in minor keys. Fast songs might be perceived as having higher dynamics.

The other mechanism is covarition of features in the stimuli. Since the stimuli are not controlled in that sense, it might be reasonable to assume that there are relationships between some features in the stimuli. Songs in major key in the stimuli might include brighter instruments, for example.

Less than half of the correlations were significant (p<0.05). Except for the emotional features, all correlations but one are below 0.7 (explaining 49% of the covariance) and can not be interpreted as marked relationships according to Williams (1968). The exception is the correlation between pitch and brightness, which is very high (r=0.90), indicating a high dependency between these features. Thus, the results are moderate since only one correlation can be considered marked, if the emotional ratings are disregarded, meaning the independence is fairly good among the perceptual ratings.

	Speed	Rhythm. complex.	Rhythm. clarity	Articul-ation	Dyna-mics	Modality	Harm. complex	Pitch	Bright-ness	Energy
Rhythmic complexity	-0.09									
Rhythmic clarity	*0.51****	*-0.54****								
Articulation	*0.57****	-0.06	*0.56****							
Dynamics	*0.66****	0.00	0.53***	*0.57****						
Modality	0.19	-0.17	0.01	0.20	0.03					
Harmonic complexity	-0.37***	*0.51****	*-0.63****	*-0.49****	-0.31**	-0.22*				
Pitch	-0.03	-0.04	-0.17	-0.09	0.05	*0.46****	0.21*			
Brightness	0.01	-0.05	-0.16	-0.02	0.12	*0.59****	0.15	***0.90******		
Energy	***0.91******	-0.08	*0.62****	*0.67****	**0.82*****	0.23*	*-0.42****	0.00	0.07	
Valence	0.16	-0.15	-0.05	0.19	-0.05	***0.92******	-0.15	*0.52****	*0.64****	0.18

Table 4. *Inter-correlations for perceptual ratings. N=100, p-values: * < 0.05; ** < 0.01, ***<0.001. Marked relationships, correlations above 0.7, are highlighted with bold font.*

5.2 Extracted audio features

5.2.1 Correlations between extracted features

The pairwise correlations between the extracted features are presented in Table 5 .

The features estimating event density (MT_Event, QM_Onsets, EX_Onsets) correlate with each other with r=0.56-0.87. Interestingly however, is that none of the event density estimating significantly features correlates with any of the tempo estimating features. All have p-values above 0.05.

The features estimating tempo do not correlate well with each other. Four out of ten possible combinations obtained a p-value above 0.05, making the correlations insignificant. The remaining six tempo estimating features correlated with values r=0.23-0.64, where r=0.64 can be interpreted as explaining 41% of the covariance. It should be noted that the best correlation between the tempo estimating features is between MT_Tempo_Spect and MT_Tempo_Both, where MT_Tempo_Both actually is a modification of MT_Tempo_Spect. These low correlation coefficients indicate that they do not estimate the same thing, even though they should. Thus, many of them were not estimating the tempo well.

The three features estimating modality correlated significantly with $|r|$=0.61-0.72, indicating that they all seem to work in the intended way. It is notable that MT_Mode_Best and MT_Mode_Sum correlated negatively, even though they use the same scale for representing the estimated mode. QM_Mode however, uses an inverted scale and is expected to correlate negatively with the two others.

The brightness estimating methods correlated significantly with r=0.66-0.94. The highest correlation, of r=0.94, was between MT_Bright_1k and MT_Bright_1.5k. This is not surprising since the methods are identical, but they use different cut-off-frequencies. Interestingly, MT_SC had strong correlations with the three brightness estimating algorithms from MIRToolbox with very high r-values, especially with MT_Bright_3k (r=0.96). Since both are measures of spectral energy distribution, it might not be so surprising.

The pulse clarity estimations did not correlate significantly with each other. This indicates a problem with one of them, since they don't estimate they same thing.

MT_SF, the calculated spectral flux from MIRToolbox, is significantly correlated to all of the event density estimations (r=0.40-0.59) and to MT_Pulse_Clarity_1 (r=0.79). Since spectral flux is a measure of spectral variations over time, a correlation with the event density estimating extracted audio features might be expected. Also, spectral flux is used in many onset detection algorithms, see e.g. Dixon (2006) or Bello et. al. (2005).

Overall, the correlation analysis indicates that many of the extracted audio features work in the intended way, as indicated by the rather high correlations between similar measures. These were the features estimating event density, brightness and modality in particular. However, the tempo and pulse clarity estimating features showed small inter-correlations indicating some kind of problem in this area.

5.3 Comparison

This section includes a comparison between the perceptual ratings and the extracted audio features. The first part presents and discusses the correlations between them and in the second part regression models are used to see how well the perceptual ratings can be predicted by combinations of extracted features.

5.3.1 Correlations

In Table 6 the pairwise correlations between the perceptual ratings and the extracted audio features are presented.

There are only three features with a marked relationship to perceptual ratings which are in the hypotheses. These are MT_Pulse_Clarity_1 with rhythmic complexity (r=0.73), QM_Onsets with speed and MT_SF with energy (r=0.75). All of the onset detection algorithms correlate significantly with speed, and QM_Onsets correlates with r=0.73, making it a marked correlation and also the best predictor for speed in this evaluation. The correlations between the event density estimating functions and the perceptual ratings of articulation, dynamics and energy were expected, since all of these perceptual ratings correlate with the perceptual rating of speed, see Table 4. It is also interesting that MT_Pulse_Clarity_1 correlates significantly with speed. Another notable observation is that none of the tempo estimating functions is significantly correlated with speed, indicating either that the ratings of speed are totally independent of the tempo of the song or that the tempo estimating algorithms do not estimate the tempo well.

Spectral flux correlates well with speed (r=0.72) as well as with energy (r=0.75). Since the perceptual features speed and energy has a very strong correlation (see Table 4, r=0.91), this is expected. Spectral flux is a measurement of spectral changes over time, making it likely to predict the speed. This also explains the correlations with rhythmic clarity and articulation, since these are ratings of some kind of change in energy over time. A song with a lot of *staccato* will probably change more in energy over time than a song with a lot of *legato*. Rhythmic clarity is reflecting changes in energy over time, since it measures energy spikes in a rhythmic pattern.

The correlation coefficients for the rated brightness are low in general. The only significant correlations are with two of the estimations for modality, MT_Mode_Best and QM_Mode. This is not in the hypotheses and due to the high significance, it cannot be discarded as a coincidence. The correlation is positive, meaning either that major songs sounds brighter or are brighter in the stimuli set. This might be due to the uncontrolled stimuli; songs in the stimuli with major key might have a brighter timbre. The other explanation is that the subjects did not understand how to rate brightness.

The modality is estimated by three different features. All three features correlates significantly with the rated modality, with $|r|$=0.47-0.67, where MT_Mode_Best has the strongest correlation.

Energy is significantly correlated with all of the features estimating event density, MT_SF and all of the brightness estimating features as predicted in the hypothesis. It is however not significantly correlated with the tempo estimating features

	MT_ZCR	MT_Tempo_Spect	MT_Tempo_Auto	MT_Tempo_Both	MT_SF	MT_SC	MT_Mode_Sum	MT_Mode_Best	MT_Bright_3k	MT_Bright_1.5k	MT_Bright_1k	MT_ASR	MT_Pulse_Clarity_2	MT_Pulse_Clarity_1	MT_Event	QM_Tempo	QM_Mode	MZ_SRF_Onsets	MZ_SF_Onsets	EX_Tempo	EX_Onsets
MT_Tempo_Spect	0.07																				
MT_Tempo_Auto	0.13	0.23*																			
MT_Tempo_Both	0.16	0.64***	0.40***																		
MT_SF	0.48***	0.18	0.06	0.02																	
MT_SC	0.64***	0.13	0.16	0.09	0.53***																
MT_Mode_Sum	-0.02	-0.06	0.04	0.01	-0.02	0.04															
MT_Mode_Best	-0.00	-0.02	-0.11	-0.06	-0.12	-0.09	-0.61***														
MT_Bright_3k	0.66***	0.11	0.14	0.06	0.62***	0.96***	0.01	-0.07													
MT_Bright_1.5k	**0.71*** **	0.07	0.07	0.09	0.51***	0.87***	0.03	-0.05	**0.91*** **												
MT_Bright_1k	**0.73*** **	0.07	0.00	0.13	0.42***	0.77***	0.05	-0.04	**0.79*** **	**0.94*** **											
MT_ASR	0.25*	0.05	0.02	0.03	0.51***	0.36***	0.04	-0.22*	0.37***	0.32**	0.24*										
MT_Pulse_Clarity_2	-0.04	-0.21*	0.26**	-0.04	-0.06	0.07	0.07	-0.07	0.05	0.01	-0.05	0.18									
MT_Pulse_Clarity_1	0.41***	0.08	0.11	0.04	0.79***	0.55***	0.01	-0.10	0.61***	0.52***	0.40***	0.62***	0.19								
MT_Event	0.36***	0.11	0.15	0.14	0.59***	0.47***	0.06	-0.07	0.49***	0.41***	0.33***	0.24*	-0.08	0.53***							
QM_Tempo	0.11	-0.13	0.18	-0.08	-0.05	0.04	-0.18	0.26**	0.01	-0.01	-0.04	-0.10	0.07	0.07	0.12						
QM_Mode	0.01	-0.08	0.01	-0.09	0.07	0.06	0.62***	**-0.72*** **	0.05	0.03	0.03	0.15	0.03	0.03	0.02	-0.13					
MZ_SRF_Onsets	0.34***	-0.02	-0.11	-0.09	0.50***	0.34***	0.00	-0.05	0.36***	0.30**	0.25*	0.11	-0.08	0.37***	0.60***	0.05	0.12				
MZ_SF_Onsets	0.33***	-0.05	-0.01	-0.00	0.40***	0.22*	0.03	-0.01	0.20*	0.24*	0.27**	0.02	-0.29**	0.17	0.67***	0.10	0.05	0.60***			
EX_Tempo	0.06	0.16	0.41***	0.30**	0.03	0.01	-0.00	-0.02	-0.01	-0.02	-0.02	-0.23*	0.03	-0.01	0.18	0.38***	-0.02	0.02	0.14		
EX_Onsets	0.25*	0.16	0.11	0.01	0.58***	0.56***	0.04	-0.09	0.54***	0.43***	0.31**	0.31**	-0.03	0.51***	**0.74*** **	0.14	0.07	0.57***	0.56***	0.07	
QM_Onsets	0.45***	0.01	-0.03	0.02	0.54***	0.29**	0.05	-0.05	0.31**	0.30**	0.30**	0.12	-0.18	0.35***	0.69***	-0.00	0.10	**0.74*** **	**0.87*** **	0.06	0.65***

Table 5 *Inter-correlations for the extracted audio features. Correlations above 0.7 are highlighted with bold font. Values expected to predict the same ratings in the hypotheses are highlighted in grey. N=100, p-values: * < 0.05; ** < 0.01, ***<0.001.*

	Speed	Rhythmic complex.	Rhythmic clarity	Articula-tion	Dynamics	Modality	Harmonic complex.	Pitch	Bright-ness	Energy	Valence
MT_Event	0.65***	0.08	0.33***	0.52***	0.47***	-0.01	-0.27**	-0.08	-0.01	0.57***	-0.01
MT_Pulse_cla1	0.61***	-0.22*	**0.73*** **	0.69***	0.56***	0.09	-0.40***	-0.13	-0.07	0.67***	0.03
MT_Pulse_cla2	-0.08	-0.34***	0.16	0.04	-0.12	0.04	-0.11	-0.01	-0.01	-0.07	0.06
MT_ASR	0.21*	-0.03	0.44***	0.62***	0.28**	-0.03	-0.26**	-0.09	-0.13	0.33***	-0.04
MT_Bright_1k	0.26**	-0.04	0.33***	0.18	0.53***	-0.03	-0.19	0.15	0.20*	0.34***	-0.13
MT_Bright_1.5k	0.31**	-0.06	0.42***	0.28**	0.55***	-0.05	-0.22*	0.08	0.16	0.38***	-0.13
MT_Bright_3k	0.37***	-0.07	0.52***	0.40***	0.47***	-0.08	-0.26**	-0.02	0.04	0.41***	-0.15
MT_Mode_Best	0.04	-0.09	-0.11	-0.11	-0.1	0.67***	-0.01	0.41***	0.51***	0	0.69***
MT_Mode_Sum	-0.04	0.09	-0.05	0.04	0.11	-0.47***	0.15	-0.19	-0.25*	-0.03	-0.43***
MT_SC	0.31**	-0.12	0.45***	0.34***	0.34***	-0.1	-0.23*	-0.03	0.03	0.31**	-0.15
MT_SF	**0.72*** **	-0.03	0.66***	0.67***	0.66***	-0.03	-0.39***	-0.15	-0.08	**0.75*** **	-0.07
MT_Tempo_Both	-0.11	0.17	-0.1	-0.06	0.03	-0.21*	0.15	-0.09	-0.08	-0.09	-0.22*
MT_Tempo_Auto	-0.08	0.02	-0.01	0	-0.02	-0.11	0.13	-0.03	0.02	-0.08	-0.13
MT_Tempo_Spect	0.02	0.12	0.04	0.08	0.07	-0.17	0.08	-0.05	-0.05	0.03	-0.16
MT_ZCR	0.43***	0.04	0.27**	0.17	0.53***	-0.02	0.01	0.14	0.15	0.45***	-0.14
QM_Onsets	**0.73*** **	0.24*	0.15	0.38***	0.50***	0	-0.13	-0.06	0	0.62***	-0.01
EX_Onsets	0.55***	0.08	0.36***	0.52***	0.34***	-0.09	-0.24*	-0.13	-0.06	0.45***	-0.06
EX_Tempo	0.15	-0.12	0	-0.05	0.08	-0.05	-0.01	-0.03	-0.04	0.06	-0.1
MZ_SF_Onsets	0.61***	0.17	0.04	0.27**	0.41***	0.06	-0.17	-0.02	0.06	0.51***	0.05
MZ_SRF_Onsets	0.64***	0.15	0.24*	0.32***	0.40***	-0.06	-0.16	-0.05	-0.02	0.55***	-0.01
QM_Mode	0	0.09	0.1	0.08	0.08	-0.58***	0.02	-0.26*	-0.39***	0.02	-0.55***
QM_Tempo	0.09	-0.21*	0.04	0.01	-0.03	0.18	-0.05	0.04	0.05	0.02	0.08

Table 6. *Correlations between perceptual ratings and extracted audio features. Values expected to correlate in the hypotheses are highlighted in grey. Marked relationships (r>0.7) are highlighted with bold font. N=100, p-values: * < 0.05; ** < 0.01, ***<0.001.*

Valence is only significantly correlated with the modality estimating features, which were selected for predicting it.

Out of the seven perceptual and emotional ratings with features selected for predicting them, three had marked relationships with at least one of their corresponding predictors. These were speed, rhythmic clarity and energy. All of the ratings with *a priori* selected predictors correlated significantly (p<0.05) with one of their corresponding predictors and all but one, brightness, obtained correlation coefficients above r=0.6. This could be considered quite moderate results, since r=0.6 can be interpreted as explaining 36% of the covariance. It is also notable that brightness obtained the lowest correlations of all of the features, including the ones lacking a features selected for predicting them.

5.3.2 Regression analysis

To find out how well the perceptual ratings could be predicted by a combination of extracted audio features, a stepwise multiple linear regression analysis was calculated with each perceptual rating as the dependent variable and with the extracted audio

features as independent variables. The dependent variable is the variable predicted by a model with independent variables which act as predictors. The extracted musical audio features were inserted as independent variables.

The R^2 measure estimates how well the regression model fits the actual data. It can be multiplied by 100 to get the percentage of the variance that is described by the model. Adjusted R^2 is modified to take into account how many independent variables are used. The R^2 value can only increase or remain unchanged with added independent variables, whereas the adjusted R^2 may decrease with added independent variables to the model, if they do not contribute enough.

Since there are 22 independent variables and 100 cases in a common multiple linear regression, there are problems with high dimensionality. According to Bryk & Raudenbush (1992) a rule-of-thumb is to have ten observations per predictor. This means a maximum of about ten predictors should be used. To select which independent variables to include, a method for variable entry has to be used. There are various variable selection methods for regression models, but according to Howell (2007), stepwise regression is the best choice. In this study, the step-wise regression model with the largest amount of variables was the model predicting speed with eight independent variables. Thus the number of independent did not exceed ten, meaning there will be no problems with high dimensionality.

The stepwise regression model starts without any independent variables and chooses the best predictors one by one, on the basis of statistical significance. The independent variable with the highest standardized regression coefficient, beta, is added if it is statistically significant. Then the independent variable with the next highest beta is added if it is statistically significant and so on, until no more statistically significant variables are left. Independent variables that have been chosen may also be removed if they loose their statistical significance to the model, when more variables are added. The criteria for variable entry and removal in the stepwise regression are as follows:

Probability-of-F-to-enter <= 0.050
Probability-of-F-to-remove >= 0.100

By using the stepwise regression approach, higher adjusted R^2 values could be achieved, since independent variables which are not contributing are disregarded, resulting in fewer independent variables.

Table 7 gives an overview of the adjusted R^2 values for the stepwise multiple linear regression models along with the number of independent variables selected by the procedure.

The features Rhythmic complexity, Harmonic complexity, Pitch and Valence are not expected to have predictors in the hypothesis, since no specific musical audio features were selected for predicting these ratings. This leaves Brightness as the only feature with an adjusted R^2 below 0.5, explaining less than half of the variance, with musical audio features *a priori* selected for prediction of the rating.

The influences of the independent variables in the stepwise regression models are shown more in detail in Table 8 through Table 19. The variables are in the same or-

der as they were added during the stepwise regression; the one on the top was added first and so on.

In the following models, beta is the standardized regression coefficient and corresponds to how much influence a variable has on the model. Sig. is an abbreviation for the significance, i.e. the chance of the result being random, according to a common rule-of-thumb, values below 0.05 are considered significant. Semipart. is an abbreviation for semipartial correlation and is the correlation between the dependent variable and the independent variable after removing the effects of the other independent variables in the model. This is also related to the change in R^2 when the independent variable is added to the model (PASW, 2009).

A problem in regression models is collinearity, which is a phenomenon where two or more independent variables are highly correlated. Regression models with high collinearity can produce strange results. The collinearity can be measured with the variance inflation factor, VIF. The cut-off for when VIF is implying too serious collinearity is suggested to be 4 according to Miles & Shelvin (2001). The VIF has been calculated for all models and did only exceed 4 in one model, the model with dynamics as independent variable.

Dependent variable	**Adjusted R^2**	**Number of independent variables**
Speed	0.76	8
Rhythmic complexity*	0.14	2
Rhythmic clarity	0.52	1
Articulation	0.62	5
Dynamics*	0.61	6
Modality	0.54	5
Harmonic complexity*	0.23	3
Pitch*	0.16	1
Brightness	0.29	2
Energy	0.68	5
Valence	0.50	2

Table 7. *Overview of regression results. Features marked with * are not expected to have predictors in the hypotheses.*

Speed

As expected in the hypotheses, one of the onset detection algorithms, QM_Onsets, explains most variation in this model. Spectral flux is chosen second and has a quite high Beta coefficient. It is notable that MT_Pulse_Clarity_1 has the second highest Beta coefficient and that there are three features estimating tempo, showing independence between them. Here there were some strange effects in the lower end of the table. The features estimating tempo was included with both positive and negative beta coefficients, indicating the inclusion of these variables was random, or that these features were not dependent of each other.

Independent variable	Beta	Sig.	Semipart.
QM_Onsets	0.468	0.000	0.383
MT_SF	0.311	0.001	0.167
MT_Pulse_Clarity_1	0.399	0.000	0.215
MT_ASR	-0.172	0.011	-0.126
MT_Tempo_Both	-0.128	0.021	-0.115
EX_Tempo	0.165	0.005	0.140
MT_Tempo_Auto	-0.127	0.029	-0.108
MT_Brightness_1.5k	-0.124	0.037	-0.103
Dependent variable: Speed. Adjusted R^2=0.76. N=100.			

Table 8. *Stepwise multiple linear regression coefficients for Speed.*

Rhythmic complexity

Independent variable	Beta	Sig.	Semipart.
MT_Pulse_Clarity_2	-0.329	0.001	-0.328
QM_Tempo	-0.192	0.043	-0.191
Dependent variable: Rhythmic complexity. Adjusted R^2=0.14. N=100.			

Table 9. *Stepwise multiple linear regression coefficients for Rhythmic complexity.*

Rhythmic complexity correlates strongest with MT_Pulse_Clarity_2, as can be seen in Table 6, where it is the only significant correlation for rhythmic complexity. Hence, MT_Pulse_Clarity_2 is expected to influence this model. Note that there were no features that were selected for predicting rhythmic complexity in the hypotheses.

Rhythmic clarity

Independent variable	Beta	Sig.	Semipart.
MT_Pulse_Clarity_1	0.726	0.000	0.726
Dependent variable: Rhythmic clarity. Adjusted R^2=0.52. N=100.			

Table 10. *Stepwise multiple linear regression coefficients for Rhythmic clarity.*

For rhythmic clarity the results were very clear and simple. MT_Pulse_Clarity_1 was the only feature included in the stepwise regression model.

Articulation

MT_Pulse_Clarity_1 was chosen as the first predictor of articulation. However, MT_ASR had the greatest influence in this regression model, since it has the highest Beta coefficient and semipartial correlation. MT_ASR was the only feature that was selected for predicting articulation in the hypotheses.

Independent variable	Beta	Sig.	Semipart.
MT_Pulse_Clarity_1	0.220	0.055	0.120
MT_ASR	0.330	0.000	0.254
MT_Event	0.229	0.005	0.180
MT_ZCR	-0.237	0.001	-0.206
MT_SF	0.306	0.007	0.172
Dependent variable: Articulation. Adjusted R^2=0.62. N=100.			

Table 11. *Stepwise multiple linear regression coefficients for Articulation.*

Dynamics

Independent variable	Beta	Sig.	Semipart.
MT_SF	0.415	0.000	0.273
MT_Bright_1k	0.493	0.000	0.297
MT_SC	-1.365	0.000	-0.347
MT_Bright_3k	1.091	0.000	0.246
QM_Onsets	0.179	0.013	0.148
MT_Mode_Sum	0.123	0.039	0.122
Dependent variable: Dynamics. Adjusted R^2=0.67. N=100.			

Table 12. *Stepwise multiple linear regression coefficients for Dynamics.*

Even though there were no features selected for predicting dynamics, this model explained 67% of the variance (adj. R^2=0.67).

The Beta coefficient for MT_SC is below -1 and for MT_Bright_3k it is above 1. This indicates a problem, since the absolute value of the standardized beta coefficient should not exceed 1.

This indicates problems with collinearity. When running collinearity statistics the VIF was 15.5 for MT_SC and 19.7 for MT_Bright_3k, which clearly exceeds the cut-off of 4, implying that the problem very likely was collinearity. The correlation coefficient between these two features is 0.96, which further implies a problem with collinearity. Both features are measures of the energy distribution in the spectrum, the same thing applies for MT_Bright_1k.

Because of the problems with collinearity in this model, another slightly modified model was calculated (see Table 13). Since the collinearity problems were caused by the strong correlations between the features estimating brightness and spectral centroid, only the two of those features which had the weakest correlation were kept. These features were MT_SC and MT_Bright_1k, which correlated with r=0.77. Thus, MT_Bright_1.5k and MT_Bright_3k were removed from the independent variables available in the stepwise regression process. The results are presented in Table 13. In the modified model only four independent variables were used and the adjusted R^2 decreased to 0.61. The VIF values decreased to 3.0 for MT_SC and 2.5 for MT_Bright_1k and are thus below the cut-off value.

Independent variable	Beta	Sig.	Semipart.
MT_SF	,573	0.000	,431
MT_Bright_1k	,662	0.000	,422
MT_SC	-,554	0.000	-,322
MT_Event	,174	0.032	,137
Dependent variable: Dynamics. Adjusted R^2=0.61. N=100.			

Table 13. *Modified stepwise multiple linear regression coefficients for Dynamics. MT_Bright_1.5k and MT_Bright_3k are not available as independent variables.*

Modality

Independent variable	Beta	Sig.	Semipart.
MT_Mode_Best	0.523	0.000	0.353
MT_Tempo_Both	-0.202	0.005	-0.198
MT_Pulse_Clarity_1	0.273	0.002	0.216
MT_Bright_3k	-0.188	0.032	-0.149
QM_Mode	-0.214	0.038	-0.144
Dependent variable: Modality. Adjusted R^2=0.54. N=100.			

Table 14. *Stepwise multiple linear regression coefficients for Modality.*

MT_Mode_Best was the feature with the most influence in this model. Note that QM_Mode also fits in this model, which shows that there are some independence between MT_Mode_Best and QM_Mode. The negative coefficients for QM_Mode are expected, since the output of this function is inverted.

Harmonic complexity

Independent variable	Beta	Sig.	Semipart.
MT_ZCR	0.484	0.000	0.318
MT_Bright_1k	-0.336	0.011	-0.229
MT_SF	-0.483	0.000	-0.420
Dependent variable: Harmonic complexity. Adjusted R^2=0.23. N=100.			

Table 15. *Stepwise multiple linear regression coefficients for Harmonic complexity.*

There were no features selected for predicting harmonic complexity in the hypotheses. It is notable that there are only low level features in this model and that the explained variance is only 23%.

Pitch

Independent variable	Beta	Sig.	Semipart.
MT_Mode_Best	0.406	0.000	0.406
Dependent variable: Pitch. Adjusted R^2=0.16. N=100.			

Table 16. *Stepwise multiple linear regression coefficients for Pitch.*

MT_Mode_Best was the only feature that was included in the stepwise regression model. There were no features selected for predicting pitch.

Brightness

Independent variable	Beta	Sig.	Semipart.
MT_Mode_Best	0.515	0.000	0.515
MT_Bright_1k	0.221	0.011	0.220
Dependent variable: Brightness. Adjusted R^2=0.29. N=100.			

Table 17. *Stepwise multiple linear regression coefficients for Brightness.*

MT_Mode_Best was clearly the best predictor for brightness in this model, even though it was not a feature selected for predicting brightness. MT_Bright_1k, which only contributes to a small degree, was one of the *a priori* selected predictors for brightness.

Energy

Independent variable	Beta	Sig.	Semipart.
MT_SF	0.374	0.001	0.203
QM_Onsets	0.464	0.000	0.331
MT_Pulse_Clarity_1	0.344	0.000	0.204
EX_Onsets	-0.243	0.004	-0.168
MT_Tempo_Both	-0.121	0.035	-0.120
Dependent variable: Energy. Energy R^2=0.68. N=100.			

Table 18. *Stepwise multiple linear regression coefficients for Energy.*

MT_SF, spectral flux, which was one of the selected features for predicting energy, is chosen first, since it has the highest correlation. In this model QM_Onsets, which was also selected for predicting energy, had the greatest influence. It is notable that EX_Onsets and MT_Tempo_Both receive negative Beta coefficients and semipartial correlations in this model, meaning they contribute in a negative way, while QM_Onsets contribute in a positive way.

Valence

Independent variable	Beta	Sig.	Semipart.
MT_Mode_Best	0.676	0.000	0.675
MT_Tempo_Both	-0.186	0.011	-0.186
Dependent variable: Valence. Adjusted R^2=0.50. N=100.			

Table 19. *Stepwise multiple linear regression coefficients Valence.*

Valence is best predicted by MT_Mode_Best, which was selected for predicting valence. MT_Tempo_Both also predicts valence negatively in the model to some degree.

6 Discussion & conclusions

This chapter discusses the results from the experiments, draws conclusions and suggests improvements and future work.

The high inter-subject reliability (Cronbach's alpha 0.83-0.98) implies that a number of 20 subjects might be enough for this kind of experiment. The inter-correlations between the perceptual ratings show only one marked relationship (between pitch and brightness) if the emotional ratings are excluded. The conclusion is that the gathered data from the listening test was sufficient to use as reference material for this experiment, even though pitch and brightness should be handled with caution.

The perceptual ratings with *a priori* selected predictors, were with the exception of brightness, were significantly correlated with a corresponding expected predictor with values ranging from r=0.62-0.75.

For all of the ratings with an *a priori* selected predictor, except for brightness, regression models could be found who could explain more than 50% of the variance, using step-wise multiple linear regression. The best regression model was the one with speed as dependent variable, where 76% of the variance is explained.

Brightness was the only perceptual rating with an *a priori* selected predictor which could not be predicted with one of the selected predictors using correlation analysis, indicating a problem with brightness. In the regression analyses for brightness there was also an exception, where the adjusted R^2 was significantly lower than for the other perceptual ratings with *a priori* selected predictors. This might be due to the uncontrolled stimuli, but it is more reasonable to think the subjects had difficulties with rating brightness, even though the inter-subject reliability was fairly high (Cronbach's alpha of 0.88). The correlation between the rated pitch and the rated brightness was very high, r=0.90, which indicates that the subjects had difficulties separating these two features, meaning the rated brightness might be a rating of pitch instead. This is hard to control, since there were no pitch estimating feature to compare the rated brightness to.

None of the extracted features estimating tempo was correlated with the perceptual ratings of speed. Neither did they correlate significantly with any of the extracted features estimating event density, which all correlated significantly with the perceptual ratings of speed. This is partly confirmed by Madison & Paulin (2010), who come to the conclusion that there is an inverse relation between relative event density and tempo, making the relative event density counteract the tempo in the perception of speed. They also come to the conclusion that event density contributes substantially to the perceived speed. More surprisingly, the extracted features estimating tempo did not correlate well with each other, which implies they are fairly independent and do not estimate the same thing. In fact, only 60% of the pairwise correlations between the features estimating tempo were significant (p>0.05). The conclusion here is that features estimating tempo do not predict the perceived speed in this experiment.

In the stepwise regression models, there were some notable results. In the model predicting speed, two features estimating tempo were contributing negatively. The

model predicting dynamics had problems with collinearity. In the model predicting energy, EX_Onsets contributes negatively, while QM_Onsets contributes positively. This shows that using stepwise regression is problematic, even though it is considered the best method for variable selection in regression (Howell, 2007), and caution should be taken when using this method.

This study could be extended by extracting more audio features and also features from the MIDI files the audio clips were created from. Since MIDI files are symbolic representations of the music, more precise calculations could be done on most mid level features, such as event density, pitch, modality, dynamics and articulation.

Even though the set of extracted audio features was small and a larger set probably would yield better results, there is still much room for improvements regarding extracted features predicting perceptual ratings. Only three of the seven perceptual ratings with *a priori* selected extracted audio features had a marked relationship with a corresponding extracted audio feature.

Given that perceptual ratings predicted emotions to a high degree in the regression models in Friberg et. al. (2011), feature extraction tools predicting perceptual musical features to a high degree could be used to predict emotions. If the extracted features reflected perceptual ratings in a better way, they could be used to calculate emotional expressions in the music.

7 References

ALLURI, V. & TOIVIAINEN, P. 2009. *In Search of Perceptual and Acoustical Correlates of Polyphonic Timbre.* In: Proceedings of the 7th Triennial Conference of European Society for Cognitive Sciences of Music (ESCOM 2009): 5-10.

BARRY, D., FITZGERALD, D., COYLE, E. & LAWLOR, B. 2005. *Drum Source Separation using Percussive Feature Detection and Spectral Modulation.* In: ISSC 2005, Dublin, September 1 – 2

BELLO, J.P., DAUDET, L., ABDALLAH, S., DUXBURY, S., DAVIES, M., & SANDLER, M. 2005. *A tutorial on onset detection in musical signals.* In: IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035–1047, 2005.

BRYK, A.S., & RAUDENBUSH, S.W. 1992. *Hierarchical linear models.* Newbury Park, CA: Sage.

BURRED, J. J., & LERCH, A. 2004. *Hierarchical Automatic Audio Signal Classification.* In: Journal of the Audio Engineering Society, 52(7/8), (2004), pp. 724-738.

BYRD, D. & FINGERHUT, M. 2002. *The History of ISMIR - A Short Happy Tale.* In: D-Lib Magazine, Vol. 8 No. 11 ISSN: 1082-9873.

CASEY, M. A., VELTKAMP, R., GOTO, M., LEMAN, M., RHODES, C., AND SLANEY, M. 2008. *Content-Based Music Information Retrieval: Current Directions and Future Challenges.* In: Proceedings of the IEEE. Vol. 96, No. 4, April 2008

DAVIES, M. E. P. & PLUMBLEY, M. D. 2005. *Beat Tracking With A Two State Model.* In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2005.

DAVIES, M. E. P. & PLUMBLEY, M. D. 2007. *Context-dependent beat tracking of musical audio.* In: IEEE Transactions on Audio, Speech and Language Processing. Vol. 15, No. 3, pp1009-1020, 2007.

DHANARAJ, R. & LOGAN, B. 2005. *Automatic prediction of hit songs.* In: International Symposium on Music Information Retrieval, pages 488–491, 2005.

DIXON, S. 2006. *Onset Detection Revisited.* In: Proceedings of the 9th International Conference on Digital Audio Effects (DAFx06), Montreal, Canada, 2006.

DOWNIE, J. S. & IMIRSEL. 2010. *MIREX 2010 Evaluation Results.* University of Illinois at Urbana-Champaign.

DUXBURY, C., BELLO, J. P., DAVIES, M., & SANDLER, M. 2003. *Complex domain Onset Detection for Musical Signals.* In: Proceedings of the 6th Conference on Digital Audio Effects (DAFx-03). London, UK. September 2003.

EEROLA, T., LARTILLOT, O., & TOIVIAINEN, P. 2009. *Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models.* In: 10th International Society for Music Information Retrieval Conference (ISMIR 2009).

ELLIS, D. P. W. 2007. *Beat Tracking by Dynamic Programming.* In: Journal of New Music Research. Vol. 37, No. 1, pp51-60, 2007.

FINGERHUT, M. 2004. *Music Information Retrieval, or how to search for (and maybe find) music and do away with incipits.* In: IAML-IASA 2004 Congress, Oslo (Norway), August 8-13, 2004.

FRIBERG, A., SCHOONDERWALDT, E., & HEDBLAD, A. 2011. *Perceptual ratings of musical parameters.* In: H. von Loesch and S. Weinzierl (eds.) Gemessene Interpretation - Computergestützte Aufführungsanalyse im Kreuzverhör der Disziplinen, Mainz: Schott 2011 (Klang und Begriff 4). (in press)

GABRIELSSON, A & LINDSTRÖM, E. 2010. *The role of structure in the musical expression of emotions.* In: P. N. Juslin, & J.A. Sloboda (Eds.), Handbook of music and emotion: Theory, research, applications, New York: Oxford University Press, 2010, pp. 367-400.

GEORGE, D. & MALLERY, P. 2003. *SPSS for Windows step by step: A simple guide and reference.* 11.0 update (4th ed.). Boston: Allyn & Bacon.

GÓMEZ, E. 2006. *Tonal description of music audio signal.* Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

HEVNER, K. 1937. *The affective value of pitch and tempo in music.* In: American Journal of Psychology, 49 (1937), pp. 621-30.

HOWELL, D. C. 2007. *Statistical Methods for Psychology, 6th edition.* Thomson Wadsworth.

IFPI. 2005. *IFPI Digital Music Report 2005.*

IFPI. 2011. *IFPI Digital Music Report 2011.*

ITU-R. 2006. *Rec. ITU-R BS.1770, Algorithms to measure audio programme loudness and true-peak audio level.* International Telecommunications Union.

JUSLIN, P. N. 2000. *Cue utilization in communication of emotion in music performance: relating performance to perception.* In: Journal of Experimental Psychology: Human Perception and Performance, 26(6), 1797-813.

KLAPURI, A. 1999. *Sound onset detection by applying psychoacoustic knowledge.* In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999.

KRUMHANSL, C. 1990. *Cognitive foundations of musical pitch.* Oxford UP.

LARTILLOT, O. 2008. *MIRToolbox 1.1 User's Manual.*

LARTILLOT, O. 2010. *MIRToolbox 1.3 User's Manual.*

LARTILLOT, O. & TOIVIAINEN, P. 2007. *A Matlab Toolbox for Musical Feature Extraction From Audio.* In: International Conference on Digital Audio Effects, Bordeaux, 2007.

LARTILLOT, O., EEROLA, T., TOIVIAINEN, P. & FORNARI, J. 2008. *Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimisation.* In: Proceedings of the 11th Conference on Digital Audio Effects (DAFx-08): 1-4.

LAUKKA, P., JUSLIN, P. & BRESIN, R. 2005. *A dimensional approach to vocal expression of emotion.* In: Cognition & Emotion. 19(5):633-653. URL: http://www.informaworld.com/10.1080/02699930441000445

LESAFFRE, M., LEMAN, M., DE VOOGDT, L., DE BAETS, B., DE MEYER, H., & MARTENS, J-P. 2006. *A user-dependent approach to the perception of high-level semantics of music.* In: Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9), Bologna/Italy, August 22-26 2006.

LINDBERG, J. NO DATE. *A Multiplatform System for the Design and Running of Audio and Visual Perception Tests.* Master Thesis at KTH (in press).

MADISON, G. & PAULIN, J. 2010. *Relation between tempo and perceived speed.* In: J. Acoust. Soc. Am., Vol. 128, No. 5, November 2010, DOI:10.1121/1.3493462

MCKINNEY, M.F. & BREEBART, J. 2003. *Features for Audio and Music Classification.* In: Proceedings of the International Symposium on Music Information Retrieval.

MILES, J. & SHEVLIN, M. 2001. *Applying Regression & Correlation.* Sage publications. P. 130.

MIREX. 2011. *MIREX HOME.* Website. URL: http://www.music-ir.org/mirex/wiki/MIREX_HOME Last modified: 2011-04-02. Retrieved: 2011-04-04.

NOLAND, K. & SANDLER, M. 2007. *Signal Processing Parameters for Tonality Estimation.* In: Proceedings of Audio Engineering Society 122nd Convention, Vienna, 2007.

NYGREN, P. 2009. *Achieving equal loudness between audio files.* Master thesis at KTH.

PACHET, F. & ROY, P. 2008. *Hit song science is not yet a science.* In: 9th International Conference on Music Information Retrieval (ISMIR 2008)

PASW. 2009. *PASW Statistics 18.0.0 for Windows User's guide.* SPSS Inc.

PEETERS, G. 2004. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project.* Tech. rep., IRCAM.

RUSSELL, J. A. 1980. *A circumplex model of affect.* In: Journal of personality and social psychology, vol. 39, pp. 1161 – 1178

SCHUBERT, E., WOLFE, J., & TARNOPOLSKY, A. 2004. *Spectral centroid and timbre in complex, multiple instrumental textures.* In: Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC 04), pages 654–657, North Western University, Illinois, 2004.

WANG, A. 2006. *The Shazam Music Recognition Service.* In: Communications of the ACM. August 2006/Vol. 49, No. 8.

WEDIN, L. 1972. *A Multidimensional Study of Perceptual-Emotional Qualities in Music.* In: Scandinavian Journal of Psychology, Vol. 13, 1972, pp 241-257.

WILLIAMS, F. 1968. *Reasoning With Statistics.* Holt, Rinehart and Winston, New York.

TRITA-CSC-E 2011:028
ISRN-KTH/CSC/E--11/028-SE
ISSN-1653-5715

www.kth.se