Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelor Thesis

Sergei Smekhunov

Emotion recognition in instrumental music using signal processing and machine learning

Fakultät Technik und Informatik
Department Informations- und
Elektrotechnik

Faculty of Engineering and Computer Science
Department of Information and
Electrical Engineering

Sergei Smekhunov

Emotion recognition in instrumental music using signal processing and machine learning

Bachelor Thesisbased on the study regulations
for the Bachelor of Engineering degree programme
Information Engineering
at the Department of Information and Electrical Engineering
of the Faculty of Engineering and Computer Science
of the Hamburg University of Aplied Sciences

Supervising examiner : Prof. Dr. Klaus Jünemann
Second Examiner : Prof. Dr. Robert Heß

Day of delivery 19. März 2018

Sergei Smekhunov

Title of the Bachelor Thesis

Emotion recognition in instrumental music using signal processing and machine learning

Keywords

Machine Learning, Emotion Recognition, Emotion Classification, Music Information Retrieval,Emotion in Instrumental Music, Audio Signal Processing

Abstract

The goal of the present document is to investigate signal processing techniques present in music information retrieval (MIR) in the context of emotion recognition in instrumental music. For this purpose machine learning techniques are employed and classifier is trained. Output of the classifier is used to estimate the efficiency and the contribution of each individual signal feature.

Sergei Smekhunov

Titel der Arbeit

Erkennung der Emotionen in der Instrumentalmusik mittels Signalverarbeitung und maschinellem Lernen

Stichworte

Maschinelles Lernen, Emotionserkennung, Emotionsklassifikation, Musikinformationsabruf, Emotion in der Instrumentalmusik, Audiosignalverarbeitung

Kurzzusammenfassung

Das Ziel des vorliegenden Dokuments ist es, Signalverarbeitungstechniken zu untersuchen, die bei der Musikinformationsgabruf (MIR) im Kontext der Emotionserkennung in der Instrumentalmusik vorhanden sind. Zu diesem Zweck werden maschinelle Lerntechniken eingesetzt und Klassifikator trainiert. Die Ausgabe des Klassifikators wird verwendet, um die Effizienz und den Beitrag jedes einzelnen Signalmerkmals zu schätzen.

Contents

List of Tables

List of Figures

1 Introduction

Since the rise of humanity, music played an important role in everyday life. As human culture developed, music developed alongside it, sprouting various genres, forms, instruments and performance techniques. There are many explanations of the importance of music at different levels of interpretation and one of them is that music conveys strong emotional load. Music has the power to stimulate strong emotions within a listener and, even if the listener is not directly affected, various emotions can be communicated just with a simple music instrument. Mechanisms of this emotional communication were studied for centuries, mainly in the form of music theory. In the 20th century, when speech recognition of a digitized audio signal became a popular research field, a number of techniques were invented that were also applicable to audio signals not containing speech but music. Researchers tried to extract signal features from audio signals and tie it with existing music theory, often going directly to the field of emotion. The combined effort of these researchers led the establishment of a new research field: Music Information Retrieval. In modern day, with the development of machine learning, signal processing, increase in computational power and vast amounts of music available through the internet it became possible to automatically analyze and annotate music using signal audio data. For instance, automatic genre classification was studied for several years (Guaus, 2009) and, while still far from perfect, is already used to automatically detect genre of the music piece. In that context, classification techniques are being developed that allow for emotion detection in music. However, this task is considered much more complex than genre classification, since emotion notation wasn't developing strictly alongside music (as music genres were) and perceived emotion is often considered much more subjective than genre definition. One of the problems, encountered by researchers of emotion-related MIR is that most of the music in a modern day consists of two components, that do not depend on each other directly: vocal and instrumental. Even casting aside the problem of the recognition of a sung text, MIR researchers struggle with identifying mixed emotional signals communicated with the lyrics and the instrumental part separately. To overcome this problem deeper understanding of emotion conveyed by each part (vocal and instrumental) is required. The focus of this thesis is a study of instrumental music since it isn't affected by "emotional noise" caused by the vocal part. Classification of an emotion, definition of an emotion, extraction of representative signal features and selection of an appropriate Machine Learning technique are addressed. Then, on the basis of these parameters classifier is built as a proof of concept in order to estimate the efficiency of the chosen signal features.

2 Requirements

In a general statement, the main requirement of this work is to build and test a classifier that identifies the emotion, perceived in a music piece based on the features extracted from the said music piece via signal processing.

In order to construct such classifier, a set of steps has to be followed:

1. A representation of an emotion identified has to be defined.
2. A set of features extracted from music pieces has to be defined.
3. Machine learning techniques appropriate for the emotion classification have to be chosen.
4. An instrumental music dataset has to be constructed based on the chosen representation of emotions.
5. A software application for feature extraction and classifier training has to be designed and implemented.
6. General efficiency of a said classifier and each particular feature in its context has to be estimated via testing.

2.1 Performace requirements

The expected minimum accuracy of the classifier has to surpass a threshold of 60%.

2.2 Additional requirements

Open source dependencies: To guarantee the availability of a developed software application for extension and/or testing on different datasets, said application has to be built on top of open source libraries and available as an open source.

3 Theoretical background

3.1 Emotion classification

Emotion classification is a contested issue in emotion-related research. Earliest attempts to classify emotions were made by Aristotle in "De Anima", dated approximately 350 BC. Since then, the topic was addressed from different research fields, including biology, psychology and sociology. While there is a wide range of different classification approaches, they can be roughly split into two categories according to two fundamental viewpoints:

- Emotions are discrete and fundamentally different constructs. (Also known as Discrete emotion theory or categorical classification)
- Emotions can be represented in a dimensional model. (Dimensional classification)

3.1.1 Categorical approach

The categorical approach to emotion classification considers that people experience emotions as categories that are distinct from each other. Various concepts were implemented to group and organize emotions into categories.

One of the first psychology papers, that directly focuses on finding and grouping terms connected to emotions was written by Kate Hevner (Hevner, 1936). In her paper, Hevner used a set of 66 adjectives that were presented to the participants of the experiment. Participants listened to music pieces and marked appropriate adjectives. Based on that, adjectives were clustered into eight groups distributed on a circle (Figure 3.1). A relative position of the two groups in this model defines the relation between the two: opposite group is the furthest apart by emotion. Later on, different researchers suggested different sets of adjectives and different groupings.

The categorical approach is often tightly coupled with the concept of basic emotions, introduced by Paul Ekman. It states that there is a limited number of innate and universal emotion categories, that can be used as a basis to derive more broad emotion classes (Ekman, 1992). In his work Ekman suggests six basic emotions, which are:

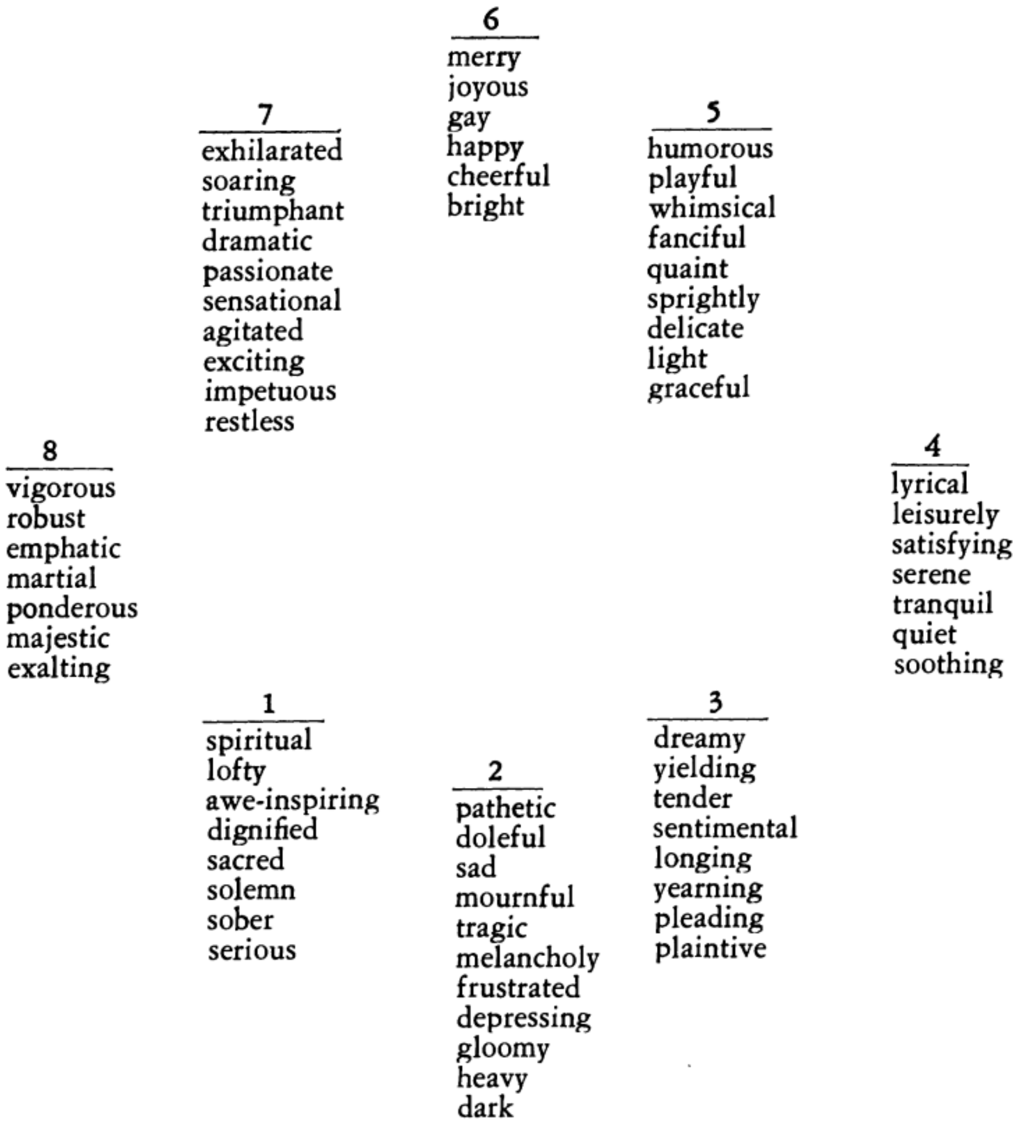


Figure 3.1: Hevner's groups of adjectives (Hevner, 1936)

- happiness,
- sadness,
- fear,
- anger,
- disgust,
- surprise.

Ekman's basic emotion list was initially based on the facial features, and since then various researchers from other fields came up with different sets of basic emotions.

The major drawback of the categorical approach is that number of basic emotion classes is too small in comparison with the range of music emotion perceived by humans, meaning that one class encapsulates a set of emotions that can be quite different. However, increasing the number of basis emotions leads to ambiguity of the definition of an emotional classification unit (Yi-Hsuan Yang, 2011).

The categorical approach by its nature is well-suited for classification algorithms of machine learning. However, careful selection of label set is required, since the same music piece can be assigned multiple labels and high intersection rate would lead to low accuracy of a classifier.

3.1.2 Dimensional approach

In the dimensional approach, emotions are identified on the basis of their location in a space with a small number of emotional dimensions. In this way, the emotion of a music piece is represented as a point or area in an emotion space. Most dimensional models incorporate valence and arousal or intensity dimensions. Dimensional models of emotion suggest that a common and interconnected neurophysiological system is responsible for all affective states. These models contrast theories of basic emotion, which propose that different emotions arise from separate neural systems (Posner u. a., 2005).

One of the first such models was proposed by James Russel (Figure 3.2). This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical axis and valence represents the horizontal axis, while the center of the circle represents a neutral valence and a medium level of arousal.

Due to its nature, the dimensional approach does not suffer from ambiguity as the categorical does. However, it is argued that the dimensional approach obscures important aspects of

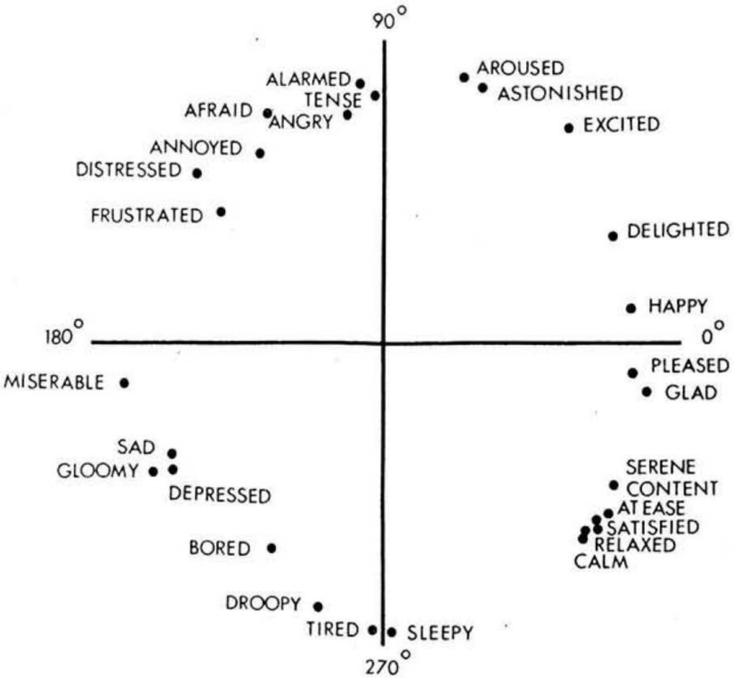


Figure 3.2: Russel's Circumplex model (Russell, 1980)

emotion process. For instance, two emotions that have a significantly different effect on the listener can be placed very closely on the valence-arousal plane.

This issue is usually addressed by creating alternative models or adding dimensions in order to improve classification precision. However, adding dimensions makes data gathering more complicated and causes an increase in error rate.

Because of these limitations, minor modifications and extensions of Russel's model are still commonly used in regression algorithms of machine learning applied to emotion recognition in music.

3.1.3 Perceptual considerations

When performing any measurement of emotion, one must also consider the source of emotion being measured. Many studies, using categorical or dimensional measurements, indicate the important distinction between one's perception of the emotion expressed by music and the emotion induced by music. (Kim u. a., 2010) While both perceived (expressed) and induced emotions are largely dependent on the observer, induced emotion is much more dependent on the observer's background, environment and context of listening, thus has a larger variance (Kallinen und Ravaja, 2006). According to Laukka et al. (Juslin und Laukka,

2004) people tend to agree more on the perceived emotion than on the induced emotion. In this paper, labels, given to music pieces are considered to be emotions perceived by users, as they are more objective.

3.1.4 Summary

Neither of the two approaches provides a full and accurate representation of emotion. However, it is believed that the approaches are complementary and can be used together: in the work of Cowen & Keltner (Cowen und Keltner, 2017) 27 different categories are connected with "gradients", effectively making it a 27-dimensional model of an emotion (Figure 3.3).

After consideration of both approaches, the categorical approach was chosen to form a set of emotion labels tested, since it gives a clearer representation of an emotion to a user and test results can be easily verified. Labels, chosen for the classifier correspond to edge points on Russel's model (Table: 3.1).

Emotion	Valence	Arousal
Sad	Low	Medium
Happy	High	Medium
Epic	Medium	High
Melancholy	Low	Low
Relaxed	High	Low

Table 3.1: Emotion categories of choice and corresponding Russel's Model values

3.2 Feature selection

An important step of audio classification is feature selection. In order to get high accuracy for classification, it is crucial to select a set of meaningful features that can capture both temporal and spectral characteristics of an audio signal. Extensive research in the area of music information retrieval (MIR) led to the discovery of vast amounts of meaningful features that can be extracted from a single music piece. (Laurier und Herrera, 2009) (Herrera u. a., 2005). Although some research has focused on searching for the most informative features for emotion classification, no dominant single feature has emerged.(Kim u. a., 2010) While it is impossible to list all features available, choice of said features for a particular testing set should still be justified.

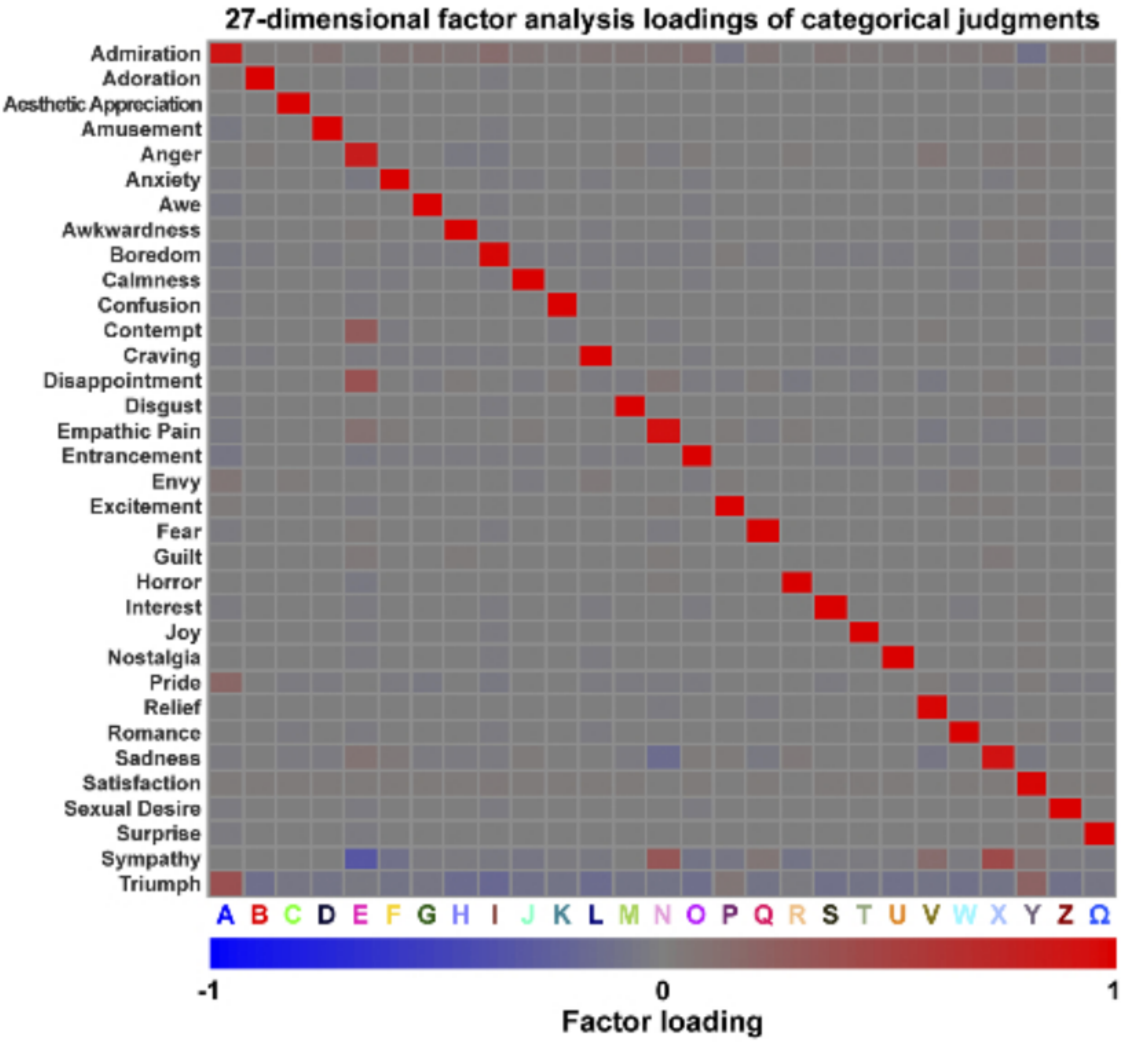


Figure 3.3: The experimental model of Cowen & Keltner (Cowen und Keltner, 2017)

Type	Features
Dynamics	RMS energy
Timbre	MFCCs, spectral shape, spectral contrast
Harmony	Roughness, harmonic change, key clarity, majorness
Register	Chromagram, chroma centroid and deviation
Rhythm	Rhythm strength, regularity, tempo, beat histograms
Articulation	Event density, attack slope, attack time

Table 3.2: Common types of features (Kim u. a., 2010)

3.2.1 Feature selection requirements

In order to select features that will be further used for classification, the following criteria should be met:

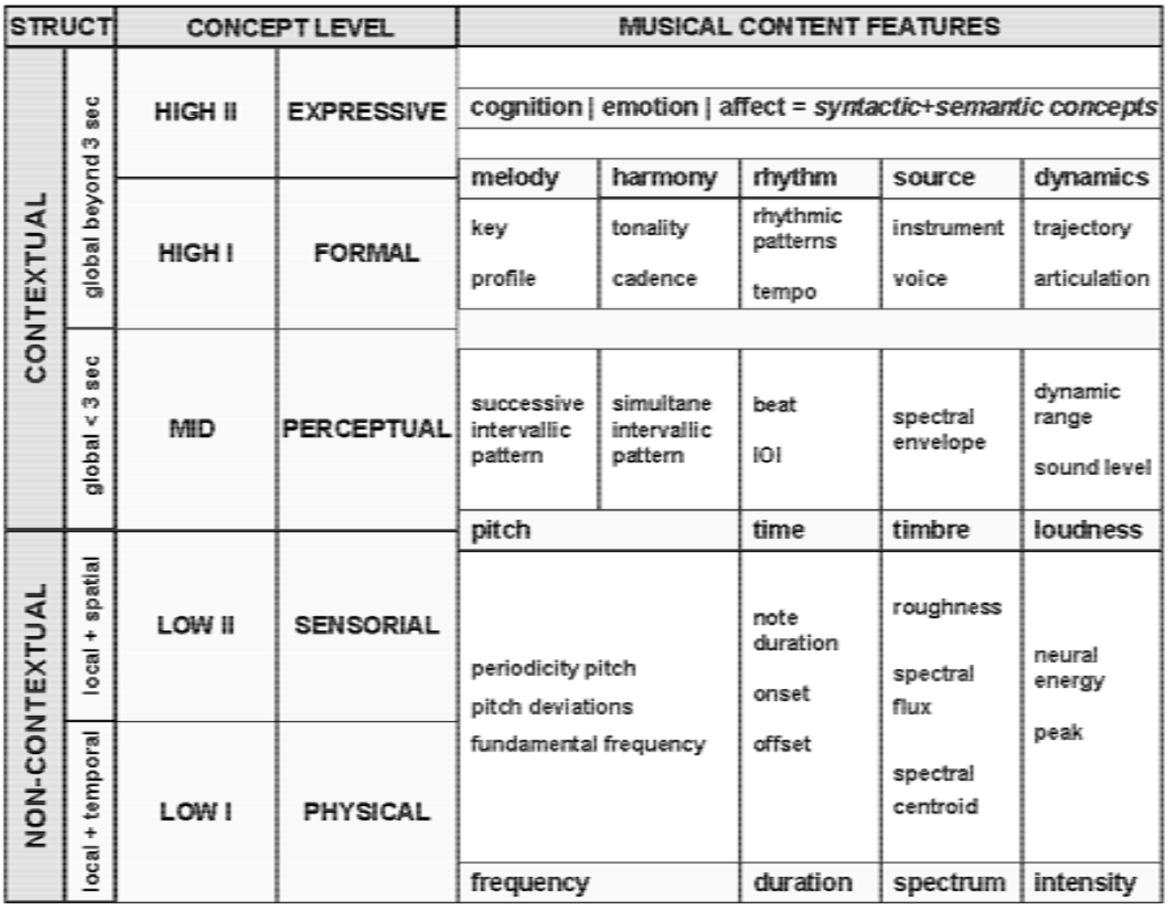
- **Features have to be non-redundant.** Both extraction of additional features and extension of feature-vector for classification significantly increase computational costs for each given entry. To resolve this issue, feature taxonomy (section 3.2.2) is introduced, which allows for a better selection of features.
- **A feature has to hold significant discriminative power**, i.e. values, obtained from feature extraction from differently labeled music pieces should also significantly differ. In order to find such features "most used" (section 3.2.3) and well-tested features are collected from different research papers and combined into a set.

3.2.2 Taxonomy of features

One way of selecting features is to introduce a taxonomy of features and then form a set of features based on their properties. While feature taxonomies still vary based on application, they tend to share common properties and can be simplified to the same level. (Peeters und Rodet, 2004) (Lesaffre u. a., 2003) However, this approach cannot be used on its own, as certain subsets contained within taxonomies would still contain large amounts of features.

While taxonomy of features is rarely addressed directly, researchers tend to group features according to Types (Table 3.2). "Types" here refer to the terminology of the music theory and it's to low-level audio features. Naming conventions and overall usage of types tend to differ from one researcher to another, but timbre, rhythm, dynamics and corresponding features are present in the majority of researches.

A few researchers addressed the issue of feature taxonomy directly, but their works are well-renowned. Lessafre et al. approach an issue of taxonomy (Table 3.3) using a combination of



STRUCT		CONCEPT LEVEL		MUSICAL CONTENT FEATURES				
CONTEXTUAL	global beyond 3 sec	HIGH II	EXPRESSIVE	cognition | emotion | affect = *syntactic+semantic concepts*				
		HIGH I	FORMAL	melody	harmony	rhythm	source	dynamics
				key profile	tonality cadence	rhythmic patterns tempo	instrument voice	trajectory articulation
	global < 3 sec	MID	PERCEPTUAL	successive intervallic pattern	simultane intervallic pattern	beat IOI	spectral envelope	dynamic range sound level
				pitch		time	timbre	loudness
NON-CONTEXTUAL	local + spatial	LOW II	SENSORIAL	periodicity pitch pitch deviations fundamental frequency		note duration onset offset	roughness spectral flux spectral centroid	neural energy peak
	local + temporal	LOW I	PHYSICAL					
				frequency		duration	spectrum	intensity

Table 3.3: Schematic overview of a user-dependent taxonomy for feature extraction (Lesaffre u. a., 2003)

Group	Features
Temporal shape	attack time, temporal increase/decrease, effective duration
Temporal features	autocorrelation coefficients, zero-crossing rate
Spectral shape features	spectral centroid, skewness, kurtosis, slope, MFCC
Harmonic features	harmonic/noise ratio, harmonic deviation, fundamental frequency
Perceptual features	Total/specific loudness, sharpness, loudness spread

Table 3.4: Common types of features (Peeters und Rodet, 2004)

abstraction level, time extent and previously mentioned "Types" to determine the position of a feature in their classification. While Lesaffre et al. base their taxonomy on user perspective (also known as the subject-centered approach), Peeters et al. (Peeters und Rodet, 2004) form a taxonomy using object-centered approach(Gouyon u. a., 2008), focusing on the features of the music piece. Despite approaching the problem from the opposite point of view, Peeters et al. come to a similar set of taxonomy descriptors:

- **The steadiness or dynamicity of the feature**, i.e the fact that the features represent a value extracted from the signal at a given time, or a parameter from a model of the signal behavior along time (mean, standard deviation, derivative or Markov model of a parameter);
- **The time extent of the description provided by the features**, i.e. scope of the extraction of a particular feature (Global, e.g. RMS of a signal or local e.g. attack time).
- **The abstractness of a feature**, i.e. amount of steps required to extract a feature from a signal. Essentially, this descriptor indirectly correlates with concept level, introduced by Lesaffre et al. (Table 3.3).
- **The extraction process of a feature** e.g.:
  - Features that are directly computed on the waveform data (Temporal features),
  - Features that are extracted using frequency domain (spectral features),
  - Features extracted using harmonic modeling
  - Features based on human hearing models (Mel/Bark scale).

  Extraction process loosely corresponds to feature types but is instead focused on the transforms applied to an input signal.

While not using user-dependent groups, Peeters et al. form a set of groups based on the extraction process of the feature extended with the other three descriptors. (Table 3.4)

Based on three taxonomies presented, following requirements can be specified for a feature set:

- Features in a feature set should cover different groups and/or types.
- Features of choice should mainly belong to physical and sensorial concept level (low level of abstraction) to maintain consistency.
- Both global and local features should be used. (Local in conjunction with statistical tools)

3.2.3 Most used features

Another way of resolving the feature selection dilemma would be to choose a subset of features most used in MIR researches involving machine learning (ML). However, to properly determine "most used" features, statistically significant amount of scientific papers should be analyzed, which is itself a separate research. Even if a list of the most used features is taken from a separate source, following issues arise:

- Some of the frequently used features will be inevitably redundant.
- Statistically significant amount of papers has to accumulate, thus a list of frequently used features will never contain recently discovered ones. (Or, in the worst-case scenario, the list itself will be dated).

Despite all the disadvantages listed above, list of the most used features provides a useful reference to techniques commonly used and vastly tested through the years.

Report, presented on Music Information Retrieval Evaluation eXchange (MIREX) by Hu et al. (Hu u. a., 2008) and meta-analysis from following years (Hu und Downie, 2010) as well as recent applications (Speck u. a., 2011) (Panda u. a., 2015) (Imbrasaite, 2015) suggest growing importance (and consequently, usage) of spectral (timbre) and rhythmic features, with MFCC used almost universally. However, a few dynamics features (signal energy/loudness) and pitch features (pitch histogram-based features) are still commonly present (Table:3.5).

3.2.4 Features of choice

After consideration of taxonomy and frequency of usage of features, following set of features was formed:

Group	Features
Energy features	RMS energy, Loudness, Loudness variation, Spectral power
Timbral features	MFCCs, spectral shape, spectral centroid, spectral flatness, spectral flux, zero-crossing rate
Pitch features	Pitch histogram features, highest amplitude, histogram bin summation
Rhythm features	Rhythm strength, regularity, tempo, beat histograms

Table 3.5: Frequently used features

- MFCCs Mel Frequency Cepstral Coefficients (Logan, 2000) are widely recognised by MIR and speech recognition researchers as a very informative feature related that is closely related to timbre perception.(Laurier und Herrera, 2009).
- Zero-crossing rate (ZCR) Being one of the simplest descriptors, ZCR (Kedem, 1986) was actively used by MIR researchers as a measure of the weighted average of the spectral energy distribution. By definition ZCR is easy to compute and provides valuable overview on noiseness of a signal.
- Spectral centroid (SC) Due to limitations of MFCC (one coefficient per frequency subband) SC (Grey und Gordon, 1978) is often used as a complementary feature that allows for more precise spectral shape description. Researhes show that SC correlates to timbral brightness of a music piece (Schubert u. a., 2004).
- RMS energy (RMSE): mean and variation Loudness of a music piece is often represented via RMSE. Combined with the timbral features, RMSE energy becomes quite discriminative for several music categories (Laurier und Herrera, 2007) (Gouyon u. a., 2008).
- Beat histogram (BH) features: values and weights of two highest peaks, a relation of their height. A BH (Tzanetakis u. a., 2001) describes how much periodicity is in the music piece at different tempo levels. In many cases, the most prominent peak of the histogram corresponds to the main tempo of the music piece (Pohle u. a., 2005). Gouyon et al. in their research paper (Gouyon u. a., 2004) suggest that rhythmic descriptors are essential for music classification.

3.3 Preprocessing

In order to get equally precise calculations of the selected features for each music piece, audio segments should be equally long and have the same sampling frequency. Since MIR is a developing research field, no standard approach to segmentation and resampling is established. Moreover, certain tasks might require specifically different sampling rates and segment lengths. However, since this paper does not have any specific requirements, commonly used segmentation/resampling techniques can be employed.

3.3.1 File format

The format of the input file is also a non-standardized property. Since compressed formats are not well-suited for the feature extraction, they should be converted to WAV data format as a part of preprocessing(Guaus, 2009).

3.3.2 Segmentation of a music piece

According to the meta-analysis conducted by Guaus (Guaus, 2009), segmentation of a music piece is considered mandatory when low-level features are being extracted. While the length of a segment varies from 5s to 120s, most commonly used segment length is 30 seconds. To avoid the inclusion of zero padding and intro in the processed segment of a music piece, first 30 seconds of each music piece are discarded. Resulting segments last from 30th second of the original music piece to 60th. Music pieces with insufficient length (less than 60s) are discarded.

3.3.3 Resampling

A sampling rate of each audio segment directly affects the calculation of the most features, thus, in order to provide consistency in feature vectors, single sampling frequency should be chosen. According to the meta-analysis conducted by Guaus (Guaus, 2009), a sampling frequency of 22.05KHz is considered the most suitable for MIR. While human hearing range goes up to 20KHz, requiring at least 40KHz as a sampling frequency, audio signals above 10KHz are considered noise on a perceptual level and can be safely discarded (Rosen und Howell, 2011). Following these guidelines, each music segment is resampled to 22.05 KHz.

3.4 Feature extraction

Here methods of feature extraction from a signal should be thoroughly discussed. The feature set used in this paper consists of local features (MFCC, SC, RMSE), measured in 30ms frames and global features (BH, SCR), measured through the whole music segment. In order to provide a considerably sized feature vector, *mean* and *variance* of local features among the frames are calculated.

3.4.1 Mel Frequency Cepstral Coefficients(MFCCs)

The Cepstrum of an input signal is defined as the Inverse Fourier Transform of the logarithm of the spectrum of the signal(Logan, 2000) :

$$[htbp]C\left[n\right] = \frac{1}{N}\sum_{k=0}^{N-1}\log_{10}\left|X\left[k\right]\right|^{j\frac{2\pi}{N}kn}, 0 < n < N-1 \qquad (3.1)$$

where $X\left[k\right]$ is the spectrum of the input signal and N its length in samples.

Calculation of MFCCs (Sigurdsson u. a., 2006), however, includes a few modifications and is described in the following way:

1. Initial signal is split into short frames, which allows the assumption that signal is stationary through each individual frame.
2. Hamming window is applied to remove edge effects.
3. The spectrum of a frame is found via FFT application.
4. Mel-scaling is applied in a form of a Mel-spaced filterbank.
5. The logarithm of amplitudes of the spectrum is calculated
6. DCT is applied in order to decorrelate overlapping Mel coefficients
7. First 8-15 (depending on the implementation) coefficients are retained, the rest is discarded as insignificant.

The core difference between calculation of Cepstrum as is and MFCCs is the application of Mel Filterbanks. The Mel scale is intended to map the perceived frequency of a tone onto a linear scale that approximates the frequency resolution of human hearing:

$$melfrequency = 2595 \cdot log_{10}[1 + \frac{f}{700}]$$

Based on the center mel-frequencies, Mel-filterbank (Figure 3.4) is created and applied to the signal, resulting in a vector of frequency bins. The number of MFCCs retained varies from researcher to researcher. Generally, latter MFCCs are considered less significant. Guaus in his meta-analysis (Guaus, 2009) suggests usage of 8 MFCCs, as further increase doesn't improve classification accuracy in a general case.

3.4.2 Zero-Crossing Rate (ZCR)

As defined by Kedem (Kedem, 1986) ZCR measures the rate of a waveform changing its sign. For a signal with a length N ZCR is defined as:

$$ZCR = \frac{1}{2}\sum_{n=1}^{N}|sign(x[n]) - sign(x[n-1])|$$

3.4.3 Root Mean Squared Energy (RMSE)

From a mathematical point of view, the time-domain energy of the input signal can be defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=0}^{N}x[n]^2}$$

Where $x[n]$ is the input time-domain data and N is length of $x[n]$ (Guaus, 2009).

To estimate sharpness of the loudness transitions a mean value of the first-order delta of RMSE sample is calculated along with mean and variance:

$$\Delta[k] = E[k] - E[k-1],$$

$$Mean\Delta = \frac{1}{K}\sum_{k=1}^{K}\Delta[k]$$

Where $E[k]$ is a vector of RMSE values for each frame and K is a number of frames.

3.4.4 Spectral Centroid (SC)

From a mathematical point of view, the Spectral Centroid can be calculated as:

$$SC = \frac{\sum_{n=0}^{N-1} f[n]a[n]}{\sum_{n=0}^{N-1} a[n]}$$

Where $f[n]$ is the frequency value of nth bin of the FFT, $a[n]$ is its amplitude and N is a number of bins.(Grey und Gordon, 1978)

3.4.5 Beat Histogram (BH)

The concept of Beat Histogram was initially proposed by Tzanetakis et al. (Tzanetakis u. a., 2001) as a feature for automatic genre classification system. Extraction of the beat histogram (Figure 3.5) is described as follows:

1. The signal is decomposed into a number of octave frequency bands using the DWT.
2. Time domain amplitude envelope of each frequency band is extracted separately as follows:
   a) Full Wave Rectification is applied:
   $$z[n] = abs(y[n])$$
   b) Low-pass (one pole with $\alpha = 0.99$) filter is applied to the rectified wave:
   $$a[n] = (1-\alpha)z[n] - \alpha z[n]$$
   c) The resulting low-frequency signal is downsampled by a factor of k:
   $$b[n] = a[kn]$$
   d) Normalization (mean-removal) is applied to a downsampled signal:
   $$c[n] = b[n] - E[b[n]]$$
3. The envelopes of each band are then summed together and an autocorrelation function is computed:
   $$d[n] = \frac{1}{N}\sum_n c[n]c[n+k]$$
4. The first five peaks of the autocorrelation function are detected and their corresponding periodicities in beats per minute are calculated and added to the histogram

The periodicity corresponding to the most prominent peak of the final histogram is considered to be the tempo in bpm of the audio file. However, since rhythm of the multi-instrumental music pieces is considered quite complex, first two histogram peak values are taken, as well as their weights separately and a relation of weights.

3.4.6 Summary

Resulting feature vector for a signal consists of 27 features:

- Mean + variance for each of eight MFCCs - 16
- Mean + variance for SC - 2
- Mean + variance + mean delta for RMSE - 3
- BH features: "2 highest peaks values, weights and a relation between weights. - 5
- ZCR - 1

3.5 Machine learning

From the practical point of view, ML aims at creating programs that optimize a performance criterion through the analysis of data (Guaus, 2009). Ml algorithms are widely used for approximating models without an obvious relation, creation of adaptive systems and clustering of data.The MIR community has traditionally used ML techniques to classify music. Likewise, the main goal of this thesis is an identification of a relation between emotion and features extracted from a physical signal. Since there is no clear connection that can be drawn berween the two, ML is used to approximate the relation and define influence of different features on it. Traditionally, ML algorythms are split into two groups:

- **Unsupervised learning** - The main property of unsupervised classifiers is that the classification emerges from the data itself, based on objective similarity measures. Multidimensional feature vectors are extracted and the distance between them is calculated in a number of ways. Based on this distance, data entries are combined intro clusters. In this case, no labels are pre-defined, and resulting clusters might have no meaning in a regard of an emotion.
- **Supervised learning** is performed on a previously defined labels and aims to define a relationship between the features extracted from a data entry and the label assigned to it. Then the relation defined is used to identify a label for new data entries. Supervised learning is commonly used in emotion-related MIR research with the combination of hand-labeled datasets.

3.5.1 Supervised learning subtypes

Supervised learning itself can be also split into two groups: regression and classification algorithms. Regression algorithms are fed numerical values as labels of the data and aim to predict a quantitative response. Classification algortighms, on the other hand, deal with categories as labels and aim to predict a catefory that data entry belongs to without any intermediary values. Both approaches are closely connected and use similar algorithms, only the classification approach uses probabilities instead of quantitative response. As described in section 3.1, categorical approach can be considered preferential due to the clarity of labels, thus one of the classification algorithms should be employed.

3.5.2 Decision Trees

One of the definitive algorithms used in classification problems (as well as regression problems) is the decision tree algortihm. The decision tree algorithm splits the training dataset into subsets based on a test attribute value. This process is repeated on each subset in a recursive manner (recursive partitioning). Decision trees classify instances by sorting them down the tree from the root to a lead node which provides the classification of the new instance, and each branch descending from that node is one of the possible values for this attribute. (Mitchell, 1997)

Decision trees for regression and classification hold a number of advantages and disadvantages over the more classical approaches, such as linear regression:

- Trees tend to mimic human perception of a choice.
- Trees are easier to export and analyze when it comes to qualitative predictors.
- Unfortunately, trees generally do not hold the same level of predictive accuracy as some other approaches.
- Trees are very sensetive to errors in input data.

However, by aggregating many decision trees, using methods such as random forests the predictive performance of trees can be substantially improved.

3.5.3 Random forests

The random forest classifier uses several decision trees in order to improve the classification rate. The basic concept behind this algorithm is common to other classifier strategies. It is the idea of combining weak learners (decision tree in this case), to build better models. A set of trees is buit based on the training set and random vector, which allows decorellation of the trees. Once the calssifier is built, input vector is estimated by each decorellated tree separately. The label with highest frequency ratio is then assigned to the input data. (Breiman, 2001)

3.5.4 Summary

Since one of the goals of this work is to test weight of different features in emotion-related decision making, random forests were chosen as a classification algorithm.

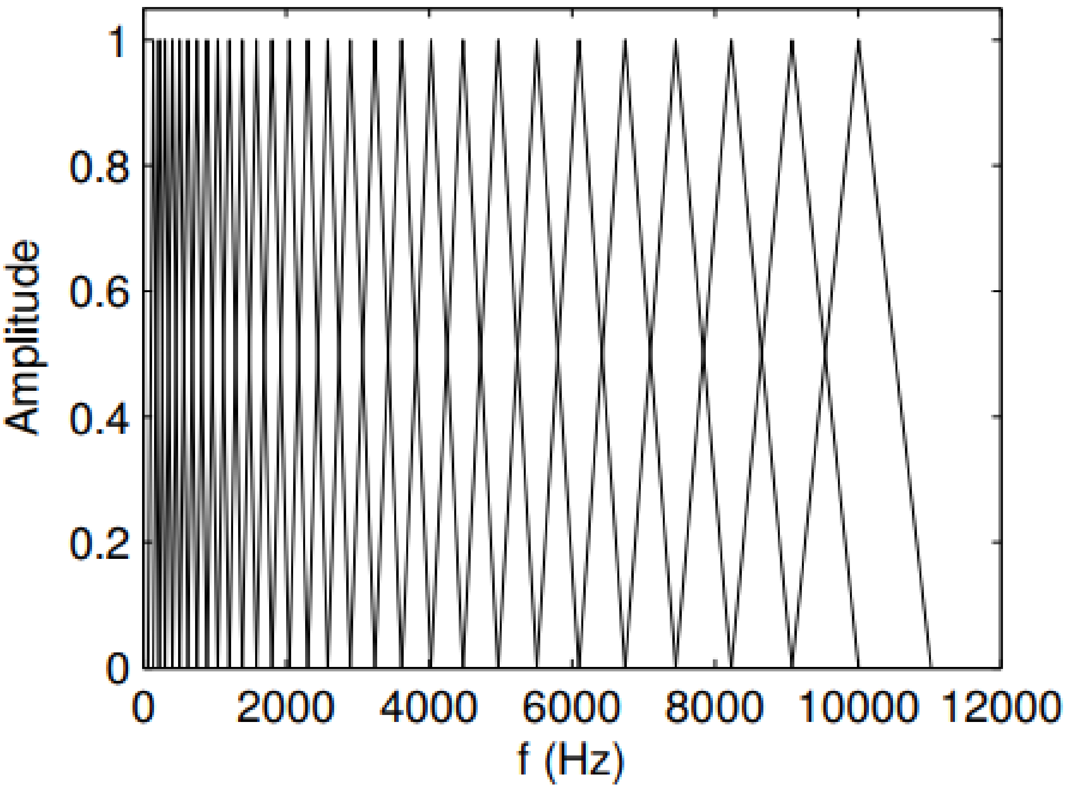


Figure 3.4: Mel-frequency filterbank (Sigurdsson u. a., 2006)

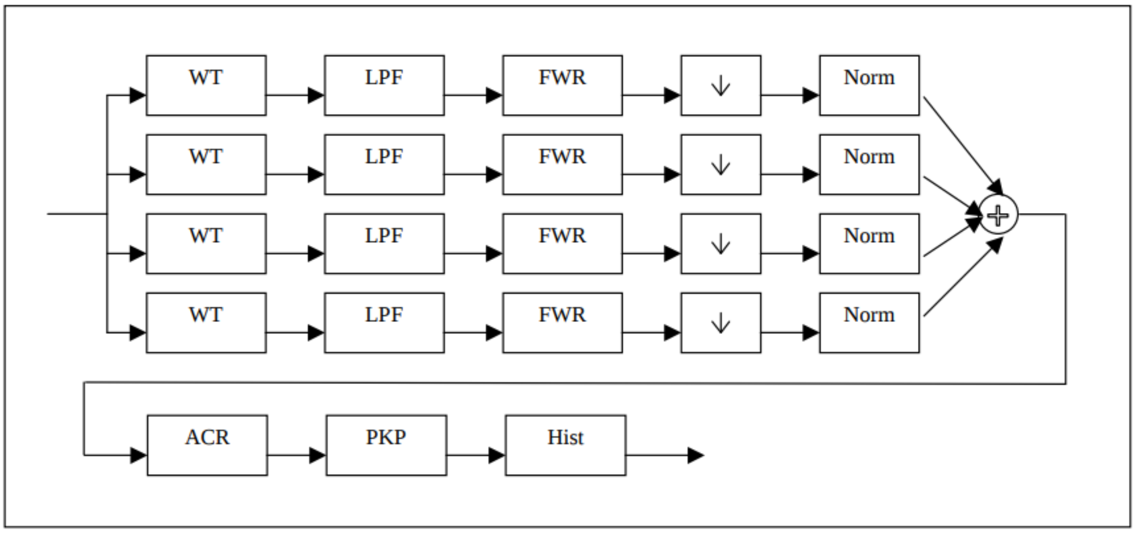


Figure 3.5: Block-diagram of the beat-detection algorithm based on the Discrete Wavelet Transform (Tzanetakis u. a., 2001)

.

4 Software design

4.1 General structure

Classification of audio data is a standard procedure in a field of MIR. General steps of such process are:

1. Labeled audio data gathering.
2. Extraction of metadata (features) from the audio data.
3. Training and testing of a classifier based on extracted metadata.

This chapter describes design decisions made in order to implement this process as a software application.

4.2 Design constraints

While requirements given in the section 2 do not define any particular constraints, leaving a lot of freedom of design and implementation, certain design decisions have to be made to guarantee correct functioning of an application developed. In particular, design of a software application should allow:

- Implementation of feature extraction as defined in the section 3.4,
- Implementation of data gathering and preprocessing described in the section 3.3,
- Implementation of ML techniques chosen in the section 3.5,
- Easy storage of music metadata.
- Structure and code of application should be easy to understand.

4.3 Dataset construction

Since this research focuses specifically on instrumental music, most of publically available MIR datasets cannot be used. It was decided to create a new dataset based on one of the complementary datasets of Million Songs Dataset(MSD) (Bertin-Mahieux u. a., 2011) - Last.fm dataset. Last.fm dataset is provided in a format of SQLite file and contains song names and corresponding tags. After reducing this set to instrumental music via tags, audio tracks can be downloaded from publically available sources(youtube.com).

4.4 Software setup

4.4.1 Selection of the programming language

Theoretically, any high-level programming language can be sufficient for the requirements listed above. However, to achieve ease of implementation and code understanding following requirements have to be met:

- Language should be familiar to the implementing person (author of this thesis).
- Language has to have a variety of signal processing / machine learning libraries.

Considering these requirements, one of four languages can be chosen: Plain C, Java, Matlab, Python. After a careful consideration of advantages and disadvantages of all four languages (Table 4.1), Python was chosen. It should be noted that most computation-heavy python libraries are in fact partially written in C to increase the performance speed. Currently two versions of Python (2.7.x and 3.4.x) are supported. While most libraries of python already have separate version for Python 3, some of them are still under a process of rewriting. Since none of Python 3 advantages can be used in the implementation, it was decided to use Python 2. Additionally, bash is used as a scripting language in order to manage files and SQL is used to manage the database.

4.4.2 Data storage

Since the initial set of songs names and corresponding labels gathered from MSD (Bertin-Mahieux u. a., 2011) was stored as SQLite database, it was decided to keep using this database engine for consistency. Apart from the consistency issue, SQLite has a huge number of advantages over other storage methods:

Language	Advantages	Disadvantages
Plain C	High performance speed Established libraries for ML and signal processing	Questionable ease of use
Java		Low performance speed
Matlab	Is specifically designed for engineers and scientists Based on C, meaning high performance speed	not open-source Not developer-friendly
Python	Easy to use Has dedicated community that provides open-source libraries for signal processing and ML	Slightly lower performance speed than C

Table 4.1: Programming languages considered for software implementation

- SQLite is an SQL database engine which allows easy access to data using SQL queries.
- SQLite is serverless and is stored as a single file, meaning that it can be easily transfered.
- SQLite is a zero-configuration database, meaning that no setup is required.
- SQLite has a very limited set of dependencies, which are already a part of a platform in the majority of cases.

Disadvantages include bad performance in case of multiple simultaneous transactions. However, since application workflow is intended to be strictly sequential, this disadvantage can be completely discarded.

4.4.3 Libraries used

Python is widely known among scientists and engineers for its dedicated community. Multiple libraries for signal processing and machine learning are being developed and updated regularly. Following libraries were chosen for feature extraction and machine learning implementation:

- Essentia (Bogdanov u. a., 2013) - a huge signal-processing library for C with a wrapper for python.

- pyAudioAnalysis (pyAudioAnalysis) - a smaller, but well-designed python library for signal processing.
- Scikit-learn (Pedregosa u. a., 2011) - a machine learning python library developed as a part of SciPy project - a group ofpython libraries designed specifically for engineers and scientists.

4.4.4 Third-party software used

Third-party application youtube-dl (youtube-dl) is used in order to fetch music pieces publically available at youtube.com.

4.5 Application workflow

With the tools selected, general steps defined in the section 4.1 can be further specified and summarized in a flowchart diagram 4.1.

4.5.1 Data gathering

Data gathering is performed in following steps:

1. Last.fm complementary set of MSD is retrieved in a form of SQLite database containing songs and corresponding tags.
2. Set is filtered from all non-instrumental music and split into emotion categories (as defined in 3.1.4) via SQL queries applied to tags.
3. Training (400 track names per category, randomly chosen, 2000 total) and testing (40 track names per emotion category, randomly chosen, 200 total) datasets are formed and stored in the database, containing equal amount of audio tracks from each emotion category.
4. After datasets are defined, youtube-dl is used to fetch corresponding audio data and store it in the file system.

4.5.2 Extraction of metadata

Preprocessing

Audio files are reformatted, segmented and resampled to match the same sampling frequency as defined in section 3.3

Feature extraction

Features, defined in 3.4 are extracted from each track sequentially and stored in the database with corresponing track name and emotion label.

4.5.3 Classifier training and testing

RF-based classifier, as defined in section 3.5 is trained and tested with corresponding training and testing data.

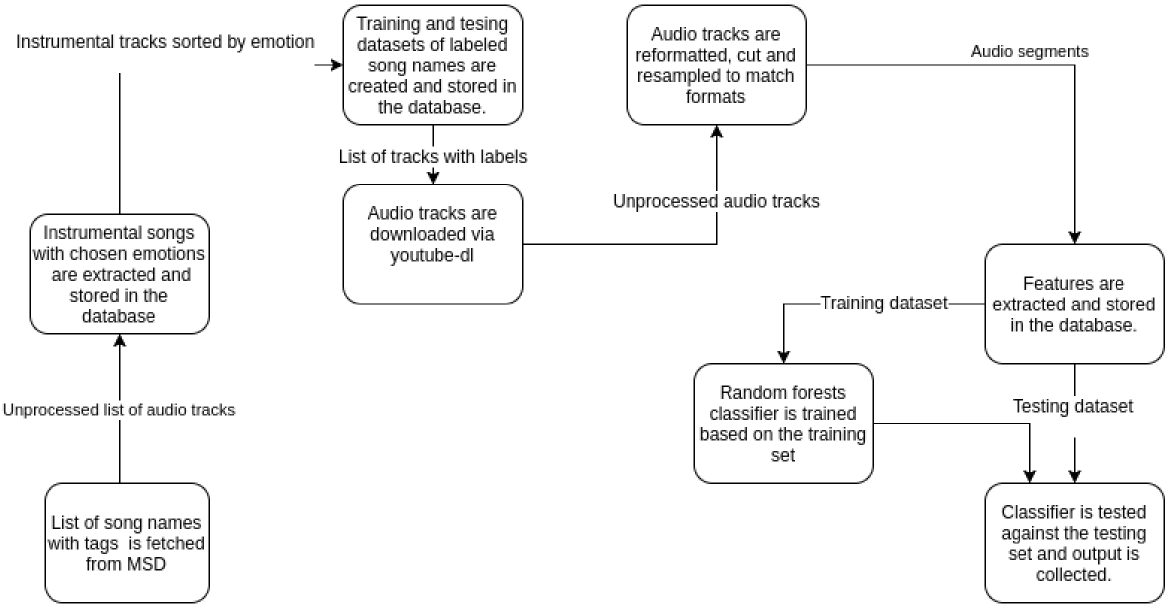

Instrumental tracks sorted by emotion
Training and tesing datasets of labeled song names are created and stored in the database.
Audio tracks are reformatted, cut and resampled to match formats
Audio segments
List of tracks with labels
Audio tracks are downloaded via youtube-dl
Unprocessed audio tracks
Instrumental songs with chosen emotions are extracted and stored in the database
Training dataset
Features are extracted and stored in the database.
Random forests classifier is trained based on the training set
Testing dataset
Unprocessed list of audio tracks
List of song names with tags is fetched from MSD
Classifier is tested against the testing set and output is collected.

Figure 4.1: Detailed application workflow

.

4.6 Application file structure

In order to allow for easier debuging and improve the robustness of an application it is implemented as a set of separated scripts, each script responsible for one of the general steps (data gathering, feature extraction, classifier training). SQLite database is used to persist datasets and corresponding metadata between the steps.

5 Results

5.1 Output of testing

In order to evaluate the efficiency of chosen features and learning methods, application was run 5 times with a randomized training dataset each time. Mean of the results were taken and compiled into the tables.

5.1.1 Classification Accuracy

5.1. It can be observed that accuracy of classification of "Epic" and "Happy" categories is significantly above the rest. To better understand this issue, Table 5.1.1 shows the confusion matrix.

Emotion Label	Sad	Happy	Melancholy	Relaxed	Epic	Overall
Accuracy	62.5%	82.0%	54.6%	68.0%	86.0%	70.62%

Table 5.1: Testing output:Overall accuracy of classification

True label/Predicted Label	Sad	Melancholy	Relaxed	Happy	Epic
Sad	62.5%	23.0%	12.6%	0.2%	1.7%
Melancholy	27.2%	54.6%	18.7%	0.0%	00.0%
Relaxed	8.4%	14.1%	68.0%	8.2%	1.3%
Happy	0.4%	0.1%	9.5%	82.0%	2.0%
Epic	4.1%	2.3%	4.5%	3.1%	86.0%

Table 5.2: Testing output:Confusion matrix

The highest confusion rate can be observed between categories "sad and "melancholy". There are a few possible reasons behind that:

- Semantic issues : "sad" and "melancholic", while defining inherently different emotions, can be sometimes used as synonyms, thus confusion between the tags is inevitable.
- Acoustic issues : both "sad" and "melancholic" are supposedly played in minor, in calm manner. Certain features extracted might identify the same behavior.

Another interesting effect that can be observed is confusion vector for "epic" label. Notably, most of the features do not get falsely assigned "epic" label, yet the opposite happens, in approximately equal parts. The reason behind this might be the very definition of "epic". This label was chosen as a representation of powerful music with high arousal. Assumption can be made that certain melodies with low arousal can be perceived as "epic" in a sense of performance technique, and do not receive any other emotion label.

5.1.2 Feature Impact

Structure of a RF classifier allows for an easy evaluation and representation of weight of each feature. In order to reduce the length of the table 5.3, MFCC and BH features were grouped together.

Feature	MFCC-mean	MFCC-var	SC-mean	SC-var	RMSE-mean	RMSE-var	RMSE-delta	BH	ZCR
Weight	0.314	0.061	0.162	0.014	0.000	0.000	0.110	0.183	0.156

Table 5.3: Testing output:Weights of features

The output of the classifier is mainly defined by the mean MFCC (representing harmonics in timbre) as well as BH features (representing tempo). Notably, ZCR is almost as effective as SC-mean, despite a huge difference in an abstraction level. Most likely this effectiveness of ZCR is caused by "clusterisation" of chosen emotion categories, as "melancholy", "sad" and "relaxed" are expected to have lower ZCR. Both MFCC and BH are a combination of coefficients, so they can be further investigated (Table 5.4, Table 5.5). Expectedly, MFCC1 doesn't hold any value, as it just represent the energy of the frame. Obvious drop in the weight of MFCCs is caused by irrelevance of higher harmonics (since they are essentially noise). Weight of the second peak of BH is surprisingly high, which might be caused by the prevalence of modern multi-instrumental music. Weights of BH peaks are expected to be of value only if BH peak itself is also of value, which causes semi-scaling of peak values and weights.

MFCC coefficient	1	2	3	4	5	6	7	8	9
Weight	0.000	0.122	0.093	0.031	0.031	0.014	0.012	0.005	0.006

Table 5.4: Weights of MFCCs individually

BH Feature	BH-peak1-value	BH-peak2-value	BH-peak1-weight	BH-peak2-weight	BH-peak-relative-weight
Weight	0.090	0.061	0.022	0.010	0.000

Table 5.5: Weights of BH features individually

5.2 Discussion of the results

While threshold specified in the section 2 is met for a mean accuracy, "melancholy" category fell below the thresholds due to the semantic and acoustic errors. It might be that different category would have been a better fit. Another crucial point is a data gathering source. Unfortunately, with multitagging available and a random distribution of the tags, high accuracy might be unachievable. Bright example of this issue is distribution of the tag "epic" into other categories in such proportions.

Feature behaviour was expected, although it might be that different post-processing techniques could enhance the behaviour of MFCCS.

Bibliography

[pyAudioAnalysis] *pyAudioAnalysis - open source python library*. https://github.com/tyiannak/pyAudioAnalysis. – Accessed: 2018-03-08

[youtube-dl] *youtube-dl - Tool for fetching audio from youtube*. https://github.com/rg3/youtube-dl. – Accessed: 2018-03-08

[Bertin-Mahieux u. a. 2011] BERTIN-MAHIEUX, Thierry ; ELLIS, Daniel P. ; WHITMAN, Brian ; LAMERE, Paul: The Million Song Dataset. In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011

[Bogdanov u. a. 2013] BOGDANOV, Dmitry ; WACK, Nicolas ; GÓMEZ, Emilia ; GULATI, Sankalp ; HERRERA, Perfecto ; MAYOR, Oscar ; ROMA, Gerard ; SALAMON, Justin ; ZAPATA, José ; SERRA, Xavier: ESSENTIA: An Open-source Library for Sound and Music Analysis. In: *Proceedings of the 21st ACM International Conference on Multimedia*. New York, NY, USA : ACM, 2013 (MM '13), S. 855–858. – URL http://doi.acm.org/10.1145/2502081.2502229. – ISBN 978-1-4503-2404-5

[Breiman 2001] BREIMAN, Leo: Random Forests. In: *Machine Learning* 45 (2001), Oct, Nr. 1, S. 5–32. – URL https://doi.org/10.1023/A:1010933404324. – ISSN 1573-0565

[Colombetti 2009] COLOMBETTI, Giovanna: From affect programs to dynamical discrete emotions. In: *Philosophical Psychology* 22 (2009), Nr. 4, S. 407–425. – URL https://doi.org/10.1080/09515080903153600

[Cowen und Keltner 2017] COWEN, Alan S. ; KELTNER, Dacher: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. In: *Proceedings of the National Academy of Sciences* 114 (2017), Nr. 38, S. E7900–E7909. – URL http://www.pnas.org/content/114/38/E7900. – ISSN 0027-8424

[Ekman 1992] EKMAN, Paul: An argument for basic emotions. In: *Cognition and Emotion* (1992), S. 169–200

[Gouyon u. a. 2004] GOUYON, F. ; DIXON, S. ; PAMPALK, E. ; WIDMER, G.: Evaluating rhythmic descriptors for musical genre classification. In: *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio*, 2004

[Gouyon u. a. 2008] GOUYON, F. ; HERRERA, P. ; GÓMEZ, E. ; CANO, P. ; BONADA, J. ; LOSCOS, A. ; AMATRIAIN, X. ; SERRA, X.: *Content Processing of Music Audio Signals*. Kap. Sound to Sense, Sense to Sound: A State of the Art in Sound and Music Computing, S. 83–160. In: POLOTTI, P. (Hrsg.) ; ROCCHESSO, D. (Hrsg.): *Sound to Sense, Sense to Sound: A State of the Art in Sound and Music Computing*. Berlin : Logos Verlag Berlin GmbH, 2008. – URL http://smcnetwork.org/http://mtg.upf.edu/files/S2S2BOOK1.pdf. – ISBN 9783832516000

[Grekow 2017] GREKOW, J.: *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces*. Springer International Publishing, 2017 (Studies in Computational Intelligence). – ISBN 9783319706092

[Grey und Gordon 1978] GREY, J. M. ; GORDON, J. W.: Perceptual effects of spectral modifications on musical timbres. In: *Acoustical Society of America Journal* 63 (1978), Mai, S. 1493–1500

[Guaus 2009] GUAUS, E.: *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers*, Universitat Pompeu Fabra, Dissertation, 2009

[Herrera u. a. 2005] HERRERA, P. ; BELLO, J. ; WIDMER, G. ; SANDLER, M. ; CELMA, O. ; VIGNOLI, F. ; PAMPALK, E. ; CANO, P. ; PAUWS, S. ; SERRA, X.: SIMAC: semantic interaction with music audio contents. In: *The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005. (Ref. No. 2005/11099)*, Nov 2005, S. 399–406. – ISSN 0537-9989

[Hevner 1936] HEVNER, Kate: Experimental Studies of the Elements of Expression in Music. In: *The American Journal of Psychology* 48 (1936), Nr. 2, S. 246–268

[Hu u. a. 2008] HU, X. ; DOWNIE, S. J. ; LAURIER, C. ; BAY, M. ; EHMANN, A.: The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: *9th International Conference on Music Information Retrieval*, URL http://mtg.upf.edu/files/publications/Ismir2008.pdf, 14/09/2008 2008

[Hu und Downie 2010] HU, Xiao ; DOWNIE, J. S.: When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In: *ISMIR*, 2010

[Imbrasaite 2015] IMBRASAITE, Vaiva: Continuous dimensional emotion tracking in music, 2015

[Juslin und Laukka 2004] JUSLIN, Patrik N. ; LAUKKA, Petri: Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. In: *Journal of New Music Research* 33 (2004), Nr. 3, S. 217–238. – URL https://doi.org/10.1080/0929821042000317813

[Kallinen und Ravaja 2006] KALLINEN, Kari ; RAVAJA, Niklas: Emotion perceived and emotion felt: Same and different. In: *Musicae Scientiae* 10 (2006), Nr. 2, S. 191–213

[Kedem 1986] KEDEM, B.: Spectral analysis and discrimination by zero-crossings. In: *Proceedings of the IEEE* 74 (1986), Nov, Nr. 11, S. 1477–1493. – ISSN 0018-9219

[Kim u. a. 2010] KIM, Youngmoo ; M SCHMIDT, Erik ; MIGNECO, Raymond ; G MORTON, Brandon ; RICHARDSON, Patrick ; SCOTT, Jeffrey ; A SPECK, Jacquelin ; TURNBULL, Douglas: Music emotion recognition: A state of the art review. (2010), 01, S. 1–12

[Laurier und Herrera 2007] LAURIER, C. ; HERRERA, Perfecto: Audio music mood classification using support vector machine. In: *International Society for Music Information Research Conference (ISMIR)*, URL files/publications/b6c067-ISMIR-MIREX-2007-Laurier-Herrera.pdf, 2007

[Laurier und Herrera 2009] LAURIER, C. ; HERRERA, Perfecto: *Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines*. Kap. 2, S. 9–32. In: VALLVERDU, J. (Hrsg.) ; CASACUBERTA, D. (Hrsg.): *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. Hershey (USA) : IGI Global, 2009, ISSN 9781605663548

[Lesaffre u. a. 2003] LESAFFRE, Micheline ; LEMAN, Marc ; TANGHE, Koen ; BAETS, Bernard D. ; MEYER, Hans D. ; MARTENS, Jean-Pierre: User-Dependent Taxonomy of Musical Features as a Conceptual Framework for Musical Audio-Mining Technology. In: *In: Proc. of the Stockholm Music Acoustics Conference. R. Bresin (Ed*, 2003, S. 635–638

[Logan 2000] LOGAN, B.: Mel Frequency Cepstral Coefficients for Music Modeling. In: *Int. Symposium on Music Information Retrieval*, 2000

[Lu u. a. 2006] LU, Lie ; LIU, D. ; ZHANG, Hong-Jiang: Automatic mood detection and tracking of music audio signals. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14 (2006), Jan, Nr. 1, S. 5–18. – ISSN 1558-7916

[Mitchell 1997] MITCHELL, Tom M.: *Machine Learning*. WCB McGraw-Hill, 1997

[Napiorkowski 2015] NAPIORKOWSKI, Sebastian: Music mood recognition: State of the Art Review. (2015)

[Panda u. a. 2015] PANDA, Renato ; ROCHA, Bruno ; PAIVA, Rui P.: Music Emotion Recognition with Standard and Melodic Audio Features. In: *Applied Artificial Intelligence* 29 (2015), Nr. 4, S. 313–334

[Pedregosa u. a. 2011] PEDREGOSA, Fabian ; VAROQUAUX, Gaël ; GRAMFORT, Alexandre ; MICHEL, Vincent ; THIRION, Bertrand ; GRISEL, Olivier ; BLONDEL, Mathieu ; PRETTENHOFER, Peter ; WEISS, Ron ; DUBOURG, Vincent ; VANDERPLAS, Jake ; PASSOS, Alexandre ; Cournapeau, David ; Brucher, Matthieu ; Perrot, Matthieu ; Duchesnay, Édouard: Scikit-learn: Machine Learning in Python. In: *J. Mach. Learn. Res.* 12 (2011), November, S. 2825–2830. – URL http://dl.acm.org/citation.cfm?id=1953048.2078195. – ISSN 1532-4435

[Peeters und Rodet 2004] Peeters, Geoffroy ; Rodet, Xavier: A large set of audio feature for sound description (similarity and classification) in the CUIDADO project / Ircam, Analysis/Synthesis Team, 1 pl. Igor Stravinsky, 75004 Paris, France. 2004. – Forschungsbericht

[Pohle u. a. 2005] Pohle, Tim ; Pampalk, Elias ; Widmer, Gerhard: Evaluation of Frequently Used Audio Features for Classification of Music into Perceptual Categories. In: *In Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05*, 2005

[Posner u. a. 2005] Posner, Johnathan ; Russel, James A. ; Peterson, Bradley S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. In: *Development and Psychopathology* 17 (2005), Nr. 3, S. 715–734

[Rosen und Howell 2011] Rosen, Stuart ; Howell, Peter: *Signals and systems for speech and hearing*. Bd. 29. Brill, 2011

[Russell 1980] Russell, J.A.: A circumplex model of affect. In: *Journal of personality and social psychology* 39 (1980), Nr. 6, S. 1161–1178. – ISSN 0022-3514

[Schubert u. a. 2004] Schubert, Emery ; Wolfe, Joe ; Tarnopolsky, Alex: Spectral centroid and timbre in complex, multiple instrumental textures. In: *8th International Conference on Music Perception & Cognition (ICMPC)*. Evanston, August 2004

[Sigurdsson u. a. 2006] Sigurdsson, Sigurdur ; Petersen, Kaare B. ; Lehn-Schiøler, Tue: Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. In: *ISMIR*, 2006, S. 286–289

[Speck u. a. 2011] Speck, Jacquelin A. ; Schmidt, Erik M. ; Morton, Brandon G. ; Kim, Youngmoo E.: A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. In: *Proceedings of the 12th International Society for Music Information Retrieval Conference*, October 24-28 2011, S. 549–554. – http://ismir2011.ismir.net/papers/PS4-13.pdf

[Tzanetakis u. a. 2001] Tzanetakis, George ; Essl, Georg ; Cook, Perry: Audio Analysis using the Discrete Wavelet Transform. In: *in Proc. Conf. in Acoustics and Music Theory Applications. WSES*, 2001

[Yi-Hsuan Yang 2011] Yi-Hsuan Yang, Homer H. C.: *Music Emotion Recognition*. CRC Press, 2011. – ISBN 9781439850466

Glossary

MFCC	Mel-Frequency Cepstral Coefficients
MIR	Music Information Retrieval
ML	Machine Learning
MSD	Million Songs Dataset
RF	Random Forests
RMS	Root Mean Square
RMSE	Root Mean Square Energy
ZCR	Zero-Crossing Rate

Declaration

I declare within the meaning of section 25(4) of the Ex-amination and Study Regulations of the International De-gree Course Information Engineering that: this Bachelor report has been completed by myself inde-pendently without outside help and only the defined sources and study aids were used. Sections that reflect the thoughts or works of others are made known through the definition of sources.

Hamburg, March 19, 2018

City, Date

sign