technische universität
dortmund

Bachelorarbeit

Relevanz cepstraler Merkmale für Vorhersagen im Arousal-Valence Modell auf Musiksignaldaten

Philipp Kramer
April 2016

Gutachter:
Prof. Dr. Günter Rudolph
Dr. Igor Vatolkin

Technische Universität Dortmund
Fakultät für Informatik
Algorithm Engineering (11)
https://ls11-www.cs.uni-dortmund.de

Inhaltsverzeichnis

Kapitel 1

Einleitung

Musik ist für viele Menschen ein wesentlicher Bestandteil des alltäglichen Lebens. Das Gebiet der *Music Recommendation* übernimmt dabei die Aufgabe, Musiktitel anhand gewählter Kriterien zu finden und anschließend vorzuschlagen. Dazu können zum einen Metadaten, wie Künstler, Album oder Genre verwendet werden, jedoch sind diese nicht immer verfügbar. Es liegt daher nahe, nur die Audioinformationen der Songs selbst zu analysieren und zu vergleichen. Die *Music Emotion Recognition* beschäftigt sich zu diesem Zweck mit der Vorhersage von Emotionen. Üblicherweise werden durch verschiedene Algorithmen Merkmale extrahiert, die einzelne Aspekte eines Audiosignals wiedergeben. Dadurch ist es möglich, auf komplexere Eigenschaften, wie Emotionen, zu schließen. Bisher wurde keine eindeutige Menge von Merkmalen gefunden, die diese Aufgabe optimal löst. In dieser Arbeit sollen nicht-cepstrale Merkmale mit Merkmalen des Cepstrums für die Vorhersage von Emotionen im Arousal-Valence Modell verglichen werden. Zudem soll der Fragestellung, ob die Hinzunahme cepstraler Merkmale eine relevante Verbesserung erziehlt, nachgegangen werden. Dabei wird sich zeigen, dass die Vorhersage des Valence-Wertes schwieriger ist, jedoch durch das Merkmal CMRARE deutlich an Genauigkeit gewinnen kann. Die durchgeführten Studien zeigen keine Verschlechterung der Vorhersagen, wenn cepstrale Merkmale zusätzlich verwendet wurden.

1.1 Motivation

Radio, Fernsehen, Internet und Smartphones machen es möglich, in immer mehr Alltagssituationen Musik zu genießen. Dabei stehen wir vor der Wahl zur eigenen Musiksammlung zu greifen, bei der zuvor Lieblingstitel oder Alben ausgewählt wurden, oder eine Radiostation zu hören, mit nicht oder gering beeinflussbaren Playlisten. Letzteres entspricht allerdings auf lange Sicht nicht der eigenen Stimmung und es beginnt die Suche nach einem neuen Sender. Eigene Playlists hingegen haben den Nachteil von Hand erstellt werden zu müssen, was entweder viel Aufwand bedeutet oder aufgrund von wenigen Songs zu

repetitiven Zusammenstellungen führt. Zu wissen, welche Emotionen ein einzelner Song erzeugt, erlaubt es, mit vergleichsweise geringem Aufwand Musiktitel vorzuschlagen, die sich in ihrer Stimmung ähneln, um so ein positives Hörerlebnis zu unterstützen. Titel, Interpret, Album, Erscheinungsjahr etc. sind die oft verwendeten Kriterien für *Music Recommendation*. Im Bezug auf Emotionen sind diese Metadaten hingegen unzureichend, da nicht vorausgesetzt werden kann, dass Titel eines Künstlers oder eines Albums ähnliche Emotionen hervorrufen. Doch die kontinuierlich steigende Zahl macht eine manuelle Annotation aller und zukünftig folgender Musiktitel aufwändig. Die Verwendung bestehender Daten ist damit ein wichtiges Kriterium der *Music Emotion Recognition*, kurz MER [24]. Ein Bereich der MER beschäftigt sich mit der Analyse kontextbezogener Informationen (*Metadaten*) in Form von Texten auf Webseiten, Liedtexten oder Stichworte (*Social Tags*) zu Liedern, wie bei dem Online-Musikdienst *Last.fm*[1]. Diese Herangehensweise setzt ein ausreichendes Vorhandensein von Informationsquellen voraus und kann gerade bei z.B. unbekannten Künstlern keine oder nur schlechte Ergebnisse liefern.

Der andere Zweig der MER, die inhaltsbasierte Audioanalyse, verwendet nur das Audiosignal selbst zur Vorhersage von Emotionen. Dieser Ansatz kann damit begründet werden, dass es einem Menschen genügt Musik zu hören, um die übermittelten Emotionen zu beschreiben. Zudem ist es auf diese Weise möglich, nicht katalogisierte Musik, wie z.B. aus *Jam Sessions*, automatisiert einzuordnen. Im Verlauf dieser Arbeit soll eine Vorgehensweise zur Emotionsvorhersage erläutert und getestet werden.

1.2 Literaturhinweise

Mit steigender Verfügbarkeit leistungsfähiger Rechner innerhalb der letzten Jahrzehnte wurde es möglich, zeiteffizient digitale Audiosignale hinsichtlich ihres Inhalts zu analysieren. Zu jährlich stattfindenden Konferenzen der „International Society of Music Information Retreival“[2], kurz ISMIR, werden eine Vielzahl, für die digitale Musikanalyse interessante Ausarbeitungen, eingereicht. 5 Jahre nach der ersten ISMIR Konferenz wurde 2005 die „Music Information Retrieval Evaluation eXchange“[3] (MIREX) gegründet. Sie hat als Ziel, in Form eines jährlichen Wettbewerbs, state-of-the-art Algorithmen zu vergleichen. Seit 2007 ist dort auch der Bereich der Emotionsvorhersage unter dem Titel „Mood Classification“[1] zu finden.

Mit der Entstehung des zweidimensionalen Arousal-Valence Modells im Jahr 1980 nach Russell [49] wurde eine Grundlage für die Darstellung von Emotionen geschaffen. Eine Emotion wird dort durch ihre erzeugte Erregung (Arousal) und Wertigkeit (Valence) beschrieben. Obwohl dieses Modell die kontinuierliche Platzierung ermöglicht, wurde es zunächst oft

[1] Last.fm: http://www.last.fm

[2] ISMIR: `http://ismir.net`, aufgerufen am 26.3.2015

[3] MIREX: `http://www.music-ir.org/mirex`, aufgerufen am 26.3.2016

für die Klassifizierung von Emotionen angewandt. Liu et al. [29] nahmen hierzu das Modell nach Thayer [54], welches den AV-Raum in die vier Quadranten als Emotionsklassen einteilt. Mit 4-Facher Kreuzvalidierung erreichten die Autoren eine Klassifikationsgenauigkeit von 85%. Xiao et al. [62] konzentrierten sich auf die Auswirkung der Klassifikationslängen. 4 s, 8 s, 16 s und 32 s wurden dort getestet, wobei 16 s eine Genauigkeit von 88.46% brachte. Beide Teams analysierten für ihre Tests eine Sammlung aus 60 klassischen Musikstücken. Mit Emotionsdaten aus „Moodswings“, einem Online-Spiel zur Annotationsgewinnung, wurde die Einteilung in vier Klassen auch bei Schmidt et al. [51] vorgenommen. Die Musikstücke sind bei der dort verwendeten Datenbank größtenteils dem Genre Pop zuzuordnen. Mit den cepstralen Merkmalen *Mel-Frequency Cepstral Coefficients* (MFCC) und (*Octave-Based*) *Spectral Contrast* (OBSC) wurde eine Trefferrate von 50.18% bei der Klassifikation erreicht. Auch nennen die Autoren das Problem der Einteilung in Klassen und wechselten daher zur Vorhersage durch Regression. Die Verwendung der Support Vector Regression (SVR) verwirft die Kontinuität des Modells während der Berechnung, wie es andere Klassifikationsansätze machen würden, nicht. Support Vector Regressen findet bei Han et al. [17] für eine 11-Klassen Vorhersage auf einem modifizierten Modell nach Thayer Anwendung. Die Klassifikationsgenauigkeit wird dort mit 94.55% beziffert.

In [52] betrachten Schmidt et al. die Abhängigkeit von Tonart und Tempo eines Musikstücks zu Arousal und Valence. Sie bestätigen eine Korrelation von Songs in Dur zu positiven Emotionen, sowie hohes Tempo zu positiven Valence und Arousal Werten. Weiterhin werden dort Untersuchungen zu verschiedenen cepstralen und nicht-cepstralen Merkmalen, bezüglich ihrer Relevanz für die genannten Zusammenhänge, unternommen.

In vielen Arbeiten, die ein neues Merkmal für die Anwendung im Gebiet des MIR vorstellen, werden Tests zum Vergleich mit bisherigen Merkmalen gemacht. In [32] wird das Merkmal *Cepstral Modulation Ratio Regression* (CMRARE) eingeführt und seine Aussagekraft durch ein Klassifikationsproblem mit Sprache, Musik und Geräuschen verdeutlicht. Hierbei wurde der Vorteil von CMRARE gegenüber statischer und dynamischer MFCC's festgestellt. Ergebnisse der Genre Klassifizierung z.B. in [21] und [3], mit *Octave-* und *Shape-Based Spectral Contrast* als beschriebenes Merkmal, können für Emotionsvorhersagen ebenso relevant sein, da ein Zusammenhang von Emotionen zu Musikstücken bestimmter Genres besteht. In [21] wird klassische Musik aus Barock, Romantik, sowie Pop, Jazz und Rock mit dem OBSC Feature klassifiziert. Eine Klassifikationsgenauigkeit von 82.3% wurde so erreicht, mit MFCC's lag sie bei nur 74.1%. Das sieben Jahre später in [3] vorgestellte Merkmal *Shape-Based Spectral Contrast* zeigt für die Klassifikation von Blues, Klassik, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae und Rock, eine, gegenüber dem als Grundlage genommenen *Octave-Based Spectral Contrast Feature*, überlegene Trefferrate. Zudem wird in dieser Arbeit die Auswirkung von verlustbehafteter Kompression durch das MP3-Format näher untersucht. Diesbezüglich erweisen sich MFCC's als robustes Merkmal mit einer erhöhten Fehlklassifikation von nur 0.7%.

In [50, p. 496] werden einzelne Merkmale zur Bestimmung von Wut, Angst, Freude und Trauer, sowie zur Vorhersage von Arousal und Valence durch lineare Regression betrachtet. Viele der dort verwendeten Merkmale finden auch in dieser Arbeit Anwendung, da AMUSE (Advanced Music Explorer) [60] als Framework zur Extraktion verwendet wird. Die Anzahl der RMS Peaks (siehe Abschnitt 3.1.1) zeigt sich sowohl für Arousal, als auch für Valence hilfreich. Auch wurde die „1000 Songs Database“, mit Arousal-Valence Annotationen zu 744 frei erhältlichen Songs für die durchgeführten Tests benutzt. Soleymani et al. [53] vergleichen in dem Paper zur Datenbank verschiedene cepstrale Merkmale, wie MFCC, OBSC und Chroma. Die Vorhersage von Valence schneidet dort, wie auch in anderen genannten Arbeiten mit Bezug auf das AV-Modell, deutlich schlechter gegenüber Arousal ab. Die Datenbank bietet neben statischen, für je einen 45 Sekunden langen Musikclip, auch dynamische Annotationen in einem 500 ms Intervall. Der durchschnittliche Fehler der Vorhersagen fällt für diese Art der Annotationen minimal geringer aus. Die hier in Kapitel 5 gemachten Studien sind denen von Rötter und Vatolkin [50] im Ansatz sehr ähnlich, da sich viele Merkmale durch Extraktion mittels AMUSE gleichen und die 744 Musikstücke der „1000 Songs Database“ verwendet wurden. Darüber hinaus soll jedoch die Aussagekraft einzelner Merkmalsgruppen wie *Energy*, *Timbre*, *Harmony and Melody*, *Tempo and Rhythm*, sowie die Gruppe der cepstralen Merkmale, in dieser Arbeit untersucht werden.

1.3 Aufbau der Arbeit

Nach der Motivation dieser Arbeit in Kapitel 1, soll zunächst im 2. Kapitel das den Untersuchungen zugrunde liegende Modell der Emotionsdarstellung erläutert werden. Abschnitt 2.2 beschreibt im Anschluss die „1000 Songs“-Datenbank, aus der die für Studien in Kapitel 5 verwendeten Musikstücke, sowie Emotionsdaten stammen. Kapitel 3 erläutert in Abschnitt 3.1 vier mögliche Gruppen von nicht-cepstralen Merkmalen. Zu jeder Gruppe wird dort beispielhaft die Berechnung ausgewählter Merkmale beschrieben. Diese Arbeit soll die Relevanz drei cepstraler Merkmale für Emotionsvorhersagen genauer untersuchen. Hierzu werden im nachfolgenden Abschnitt 3.2 *Mel-Frequency Cepstral Coefficients*, *Octave-Based Spectral Contrast*, sowie das Merkmal der *Cepstral Modulation Ratio Regression* erklärt. Abschnitt 3.3 beschreibt, wie diese Merkmale im Rahmen der Arbeit extrahiert wurden. Die zum Verständnis der durchgeführten Studien notwendigen Grundlagen werden in Kapitel 4 beschrieben. Dazu gehört die *Multiple Lineare Regression* (Abschnitt 4.1), welche die Methode der eigentlichen Vorhersage darstellt, *Kreuzvalidierung* (Abschnitt 4.2), welche die Daten für Test und Training einteilt, sowie ein Verfahren zur Selektion von Merkmalen nach dem Prinzip der *Minimum Redundancy — Maximum Relevance* (Abschnitt 4.3). Kapitel 5 erläutert zunächst, wie die folgenden Ergebnisse gewonnen wurden. Einen Überblick über gemachte Studien gibt der Testablauf in Abschnitt 5.1. Nachdem in 5.2, 5.3

und 5.4 die Parameter der drei cepstralen Merkmale unabhängig voneinander optimiert wurden, beschreibt Abschnitt 5.5 die Ergebnisse der in 4.3 erklärten Methode zur Auswahl relevanter, gering redundanter Merkmale. Die Fragestellung nach der Relevanz cepstraler Merkmale wird in Abschnitt 5.6 beantwortet. Kapitel 6 fasst gewonnene Ergebnisse und Aussagen der Arbeit abschließend zusammen.

Kapitel 2

Emotionen

In der psychoakustischen Forschung geht es um die Beziehung zwischen Anregung (*Stimulus*) und Reaktion (*Response*) [47]. Der Mensch (Subjekt) wird als „Blackbox“ zwischen der Musik als Anregung und der Emotion als Reaktion betrachtet, die es hinsichtlich sensorischer Prozesse zu untersuchen gilt. Wenn von Emotionen im Zusammenhang mit Musik die Rede ist wird zwischen ausgedrückter, wahrgenommener und induzierter Emotionen unterschieden. Die vom Künstler selbst ausgedrückte Emotion (*Expressed Emotion*) muss dabei nicht zwingend mit der vom Hörer wahrgenommenen Emotion (*Perceived Emotion*) übereinstimmen. Aus wahrgenommener Emotion folgt als drittes die durch Umwelt und persönliche Faktoren beeinflusste induzierte oder gefühlte Emotion (*Induced / Felt Emotion*) [13, 15, 23, 65, 50]. Da ausgedrückte und induzierte Emotion nicht nur von der Musik selbst abhängig ist, soll der Fokus im Folgenden auf der wahrgenommenen Emotion liegen. Die MER konzentriert sich dabei auf den Prozess zwischen dem Low-Level Audiosignal und der Wahrnehmung des Menschen (High-Level) [65] . Diese „Blackbox“ wird dabei meist als statistisches Modell angenommen [47]. Um dieses trainieren und anschließend testen zu können, wird eine Menge von manuell annotierten Musikstücken benötigt, welche die *Ground Truth* für alle Untersuchungen bilden. Das Sammeln dieser Daten ist ein wiederum eigenständiges Gebiet, bei dem verschiedene Ansätze Anwendung finden.

Naheliegend ist es, eine ausgewählte Menge von Musiktiteln aus möglichst verschiedenen Musikrichtungen von Experten kennzeichnen zu lassen. Doch die Wahrnehmung von Emotionen in der Musik geschieht eher subjektiv und eine Beeinflussung durch persönliche Erfahrungen ist in Einzelfällen möglich. Viel mehr soll die Emotionswahrnehmung alltäglich gehörter Musik erfasst werden. Das heißt Pop statt klassischer Musik, die für frühere Studien bevorzugt Verwendung fand, sowie Menschen ohne besondere Vorkenntnisse statt Experten für die Erhebung von Emotionsdaten [63]. Um allgemeingültige Aussagen zu erhalten, besteht die Notwendigkeit, die Daten über mehrere Tester zu mitteln. Üblich ist es daher bezahlte Umfragen einzusetzen, bei denen Experten zuvor eine geringere An-

zahl von Musikstücken zur Qualitätskontrolle annotieren. Teilnehmer müssen sich zuvor qualifizieren, um so für Umfragen zugelassen zu werden [53].

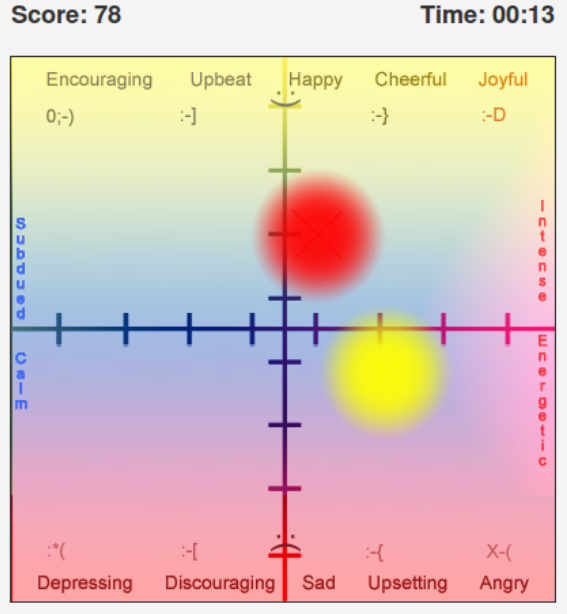


Abbildung 2.1: MoodSwings [2]

Eine andere Möglichkeit stellen die *Games with a purpose* dar [61]. Der Anreiz soll dabei nicht das Geld, wie es bei Umfragen meist der Fall ist, sondern der Spassfaktor sein. „MajorMiner“[1], “ListenGame“, „TagATune“ und „Herd It“[2] sammeln Daten zu meist 30 Sekunden langen Songausschnitten in Form von Stichworten (*Tags*) [31, 56, 28, 6]. Diese Stichworte beschreiben Kategorien von Emotionen, welche sich teilweise sehr ähneln. Fröhlich (*glad*) und Zufrieden (*pleased*) sind Begriffe, die eine hohe emotionale Ähnlichkeit aufweisen. Wie wäre demnach eine Zuordnung in nur eine der beiden Kategorien zu beurteilen? Wie im Modell nach Russell in Abbildung 2.2 zu sehen, lassen sich diese Stichworte in einem zweidimensionalen Raum platzieren. Nach diesem Prinzip arbeitet „MoodSwings“[3]. Der Nutzer wird aufgefordert, die momentan wahrgenommene Emotion in einem Koordinatensystem einzuordnen (Abbildung 2.1). Ein Großteil von Emotionsstichworten lässt sich in diesem Arousal-Valence Modell [49] eindeutig platzieren. Durch Überführung von kategorischen Annotationen in eine kontinuierliche Beschreibung durch Arousal und Valence wird deren Ähnlichkeit bei der Vorhersage beachtet. Werden anschließend Emotionensbeschreibungen durch Stichworte benötigt, können diese im AV-Modell klassifiziert werden.

[1] MajorMiner: `http://majorminer.org`, aufgerufen am 8.2.2016
[2] Herd It: `http://herdit.org`, aufgerufen am 8.2.2016
[3] MoodSwings: `http://music.ece.drexel.edu/mssp/`, aufgerufen am 8.2.2016

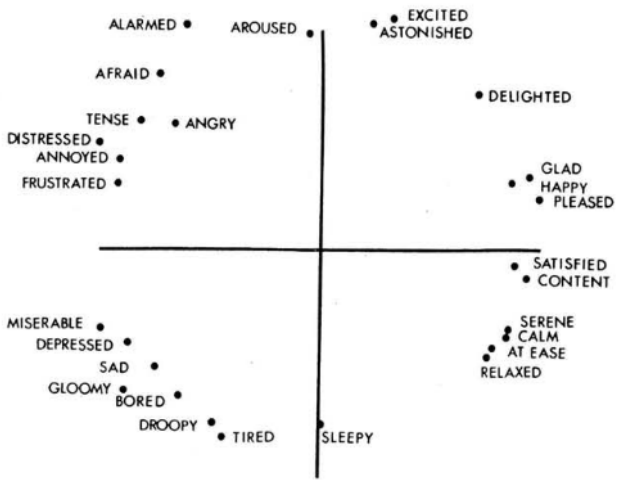


Abbildung 2.2: Multidimensionale Anordnung von Emotionen im Arousal-Valence Modell nach Russell [49]

2.1 Arousal-Valence Modell

Die Erregung oder Stärke der Emotion (*Arousal*) wird im Arousal-Valence Modell auf der horizontalen Achse dargestellt. *Valence*, die Wertigkeit, beschreibt, ob das Gefühl eher positiv (*Happy*) oder negativ (*Sad*) ist. Durch die Kontinuität des Modells ist es möglich, ein Gefühl feiner abzustufen und minimale Veränderungen zu erkennen. Die Eingabe durch den Nutzer erfolgt durch einfaches Platzieren des Mauszeigers im AV-Raum. Hohe Abtastraten sind hierdurch ohne Mehraufwand zu realisieren. Um den Anwender zu sinnvollen Eingaben zu motivieren, werden Punkte nach Übereinstimmung mit anderen Spielern vergeben, was einen geringen Administrationsaufwand bedeutet und zugleich eine hohe Qualität der gewonnenen Daten gewährleistet. Ein Großteil von Emotionen kann in diesem 2D-Raum zuverlässig und logisch auch von ungeübten Personen eingeordnet werden. Arousal und Valence sind somit als die zwei grundlegenden Emotionsdimensionen anzusehen [49]. Nicht alle Emotionen haben innerhalb des AV-Modells einen eindeutigen Platz. Wut und Angst zum Beispiel liegen nah beieinander (hoher Arousal Wert, geringe Valence) [64, p. 20]. Das Hinzunehmen einer dritten Dimension wird in der Literatur teilweise vorgeschlagen [9], würde im Gegenzug die praktische Anwendung erschweren. Im Rahmen dieser Arbeit soll ebenfalls das Arousal-Valence Modell zur parametrischen Darstellung von Emotionen verwendet werden.

2.2 „1000 Songs Database“

Ein aktueller Datensatz zur Emotionsanalyse im Arousal-Valence Modell ist die „1000 Songs Database“ [53]. Durch Crowdsourcing wurden mithilfe von Amazon Mechanical Turk[4], einer Platform für bezahlte Umfragen, Arousal und Valence Daten zu 1000 ausgewählten Songs gesammelt.

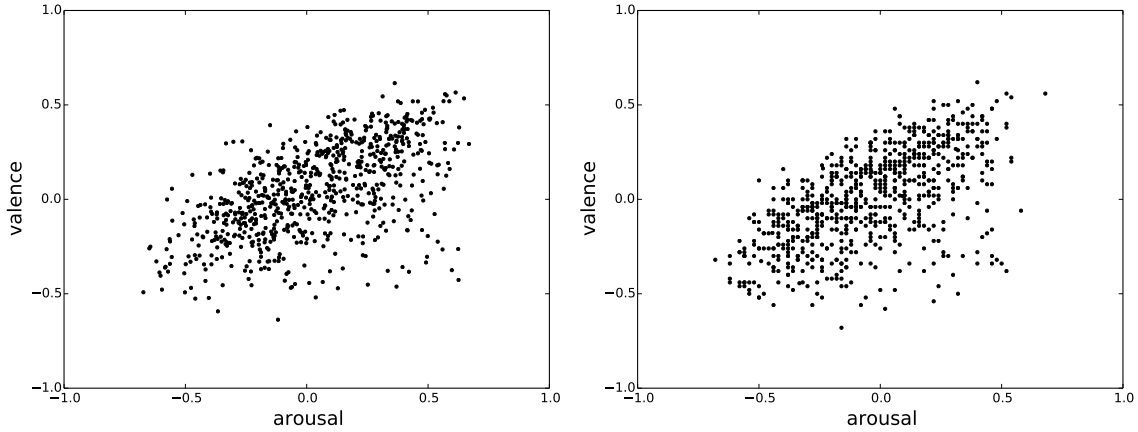


Abbildung 2.3: Verteilung von Arousal und Valence
Links: dynamisch, Rechts: statisch

Die verwendeten Musiktitel stammen von Free Music Archive (FMA)[5] und sind unter *Creative Commons*[6] lizensiert, wodurch der 1000 Songs Datensatz nicht nur die gewonnenen Annotationen, sondern auch alle Musikdaten enthält und frei heruntergeladen werden kann[7]. Aus den Genres Blues, Electronic, Rock, Classical, Folk, Jazz, Country und Pop wurden jeweils die 300 meistgehörten Songs nach FMA Statistik gewählt. Musikstücke mit weniger als einer Minute und mehr als 10 Minuten Spieldauer wurden daraus aussortiert. Anschließend sind die besten 125 Titel aus jedem Genre in die finale Auswahl übernommen worden. Es ergaben sich dadurch 53 - 100 verschiedene Künstler pro Genre, weshalb keine weiteren Limitierungen getroffen werden mussten. Um eine gute Qualität der Annotationen zu erhalten, war eine Qualifikation der Teilnehmer erforderlich, indem sie ihr Verständnis des Arousal-Valence Modells zeigten. Dazu wurden Songs mit stark dynamischen Arousal und Valence Verhalten vorgespielt. Die korrekte Angabe des Verlaufs von Arousal und Valence (steigend oder sinkend) und die Frage nach Genre und einer kurzen Beschreibung des Gehörten qualifizierte 287 Teilnehmer.

[4] Amazon Mechanical Turk: `mturk.com`, aufgerufen am 10.2.2016
[5] Free Music Archive: `freemusicarchive.org`, aufgerufen am 10.2.2016
[6] Creative Commons: `creativecommons.org`, aufgerufen am 10.2.2016
[7] 1000 Songs Database: `http://cvml.unige.ch/databases/emoMusic/`, aufgerufen am 9.11.2015

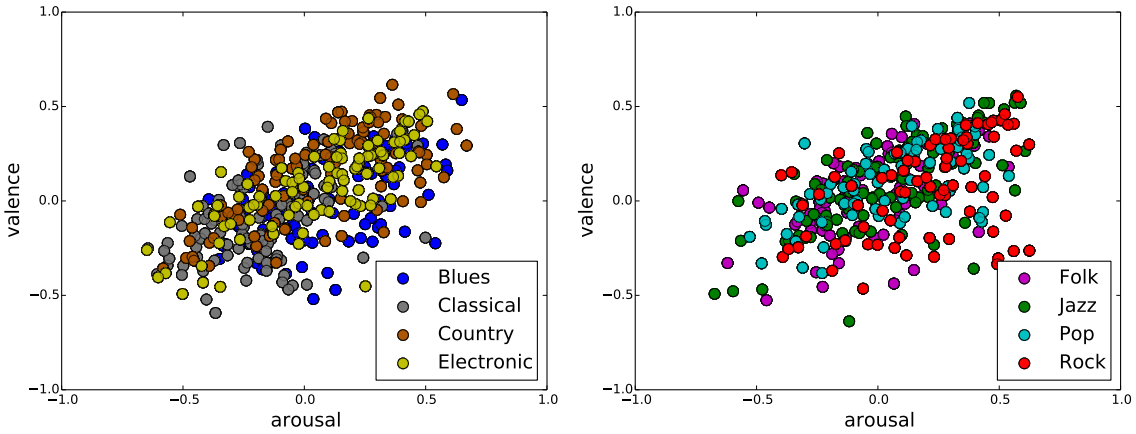


Abbildung 2.4: Verteilung von Arousal und Valence nach Genres

100 Personen davon nahmen an den Haupttests teil, welche in 334 *micro-tasks* mit je drei 45 Sekunden Ausschnitten gegliedert wurden. Zu Ende eines Songs wurden die Teilnehmer aufgefordert, dem gesamten Ausschnitt jeweils Arousal und Valence Werte auf einer Skala von 0 bis 10 zuzuweisen. Zusätzlich wurden mit einer Frequenz von 2 Hz dynamische Annotationen, durch Auslesen des Mauszeigers im 2D Arousal-Valence Raum, erfasst. Die Verteilungen der über ganze Songs gemittelten dynamischen, sowie statischen Daten sind in Abbildung 2.3 im 2D Arousal-Valence Raum dargestellt. Abbildung 2.4 zeigt in beiden Grafiken die gemittelten dynamischen Werte aller verwendeten Musiktitel, wobei die acht Genres jeweils farblich kodiert sind. Hierbei fällt auf, dass Titel des Genres Klassik eher im 3. Quadrant zu finden sind. Dies lässt auf einen Großteil eher ruhiger oder melancholischer Musikstücke im Genre Klassik schließen. Für den Datensatz fand eine Verkürzung der Annotationen auf die letzten 30 Sekunden jedes Musiktitels statt, da zu Anfang Arousal und Valence Angaben, aufgrund der den Teilnehmern überwiegend unbekannten Liedern, unzuverlässig waren. Die Verwendung von *Creative Commons* lizensierter Musik hat einen zusätzlichen Vorteil für die Emotionsanalyse. Sie wird selten im Radio gespielt und ist den Teilnehmern daher oft nicht bekannt, wodurch keine eigenen emotionalen Verbindungen mit den gewählten Musiktiteln bestehen und so das Ergebnis nicht verfälscht wird. M. Soleymani et al. zahlten pro abgeschlossene Qualifikationsaufgabe 0.25 USD und pro Hauptaufgabe 0.40 USD. Für 1784,50 USD wurden somit insgesamt 20000 Annotationen gesammelt. Die Interpretation, gerade von Emotionen, kann unter Teilnehmern variieren. Jeder Song, der von mindestens 10 Personen annotiert war, konnte zugelassen werden, um eine ausreichend allgemeingültige Aussage der Emotionen zu erhalten. Somit enthält die 1000 Songs Database zu 744 Musikstücken Emotionsdaten.

Kapitel 3

Merkmale

Bereits früh wurde festgestellt, dass die wahrgenommene Stimmung in Musik unabhängig von der musikalischen Erfahrungen eines Menschen ist [19]. Bestimmte Strukturen sind besonders entscheidend für das Empfinden von Emotionen. Als Beispiele lassen sich Tempo, Tonlage, Lautstärke und Klangfarbe für Arousal als relevante musikalische Merkmale aufzählen. Tonart und Harmonie können als für Valence wichtig angesehen werden [65, 15]. Für die MER ist es daher von grundlegender Bedeutung, diese Strukturen in Form von Merkmalen (Features) zu extrahieren und zu analysieren.

Einige Merkmale werden typischerweise für kleine Zeitfenster von 20 ms [30] bis 4 s [50] berechnet und bestehen aus einer Zahl oder bei Merkmalen mit mehreren Dimensionen einem Vektor, der das entsprechende Merkmal repräsentiert. Das Vorgehen der Merkmalsextraktion ist in Abbildung 3.1 schematisch dargestellt. Eine Überlappung um 50% der Extraktionsfenster wird teilweise empfohlen, um Abschnitte zwischen zwei Fenstern nicht zu vernachlässigen [5]. Für ein 30 s langes Musiksignal würden somit bei 20 ms Zeitfenstern etwa 3000 Vektoren pro Merkmal generiert. Die Art der Features reicht dabei von einfachen statistischen Kennzahlen des Musiksignals (z.B. Zero-Crossing Rate

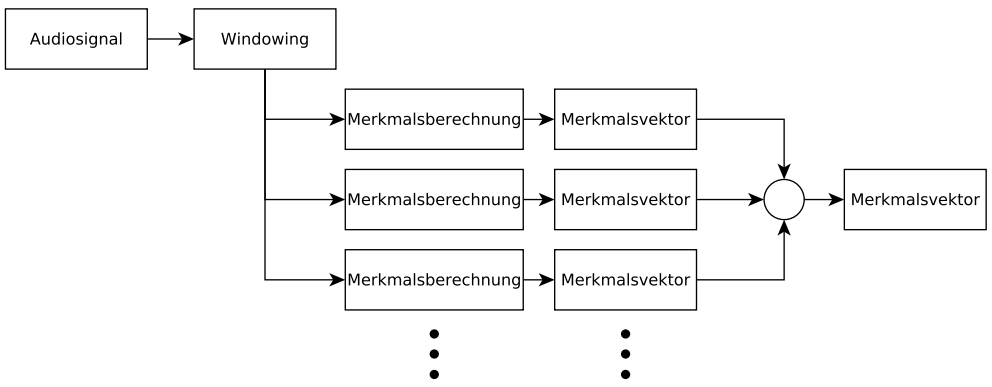


Abbildung 3.1: Merkmalsextraktion als Blockdiagramm

oder Root Mean Square) bis zu Analysen im Frequenzspektrum oder Cepstrum (Kapitel 3.2). Um einen Song oder einen Ausschnitt zu Klassifizieren bzw. im Arousal-Valence Modell vorherzusagen, werden die Merkmale über die Zeit zu einem Vektor zusammengefasst, welcher das gesamte Musiksegment im Sinne der Merkmale wiedergibt. Durch die Art der Vorverarbeitung kann das Ergebnis mitunter beeinflusst werden. Weiterhin ist es möglich, durch das Hinzufügen verschiedener Vorverarbeitungen den resultierenden Merkmalsvektor zu erweitern, um z.B. sein zeitliches Verhalten mit einzuschließen [36, 21]. Da diese Merkmale nur die Beschaffenheit eines Musiksignals erfassen, lassen sich keine direkten Aussagen zu Stimmungen oder Emotionen treffen, zumindest hat sich noch kein einzelnes dominantes Feature herausgestellt [24]. Aufgrund dessen wird meist eine Menge an Features kombiniert, um so mittels Regression oder Klassifikation auf das angestrebte höhere Merkmal wie Genre oder Emotion zu schließen. Nicht alle Features haben die gleiche Bedeutung für eine bestimmte Aufgabe, sodass eine Vorselektion der Merkmale sinnvoll ist (siehe Kapitel 4.3). In den folgenden Abschnitten 3.1 und 3.2 sollen ein paar der als später wichtig herausgestellten Merkmale näher beschrieben werden. Zur Veranschaulichung sind jeweils die zeitlichen Verläufe von vier der in Tabelle 3.1 aufgelisteten Songausschnitte dargestellt.

Titel	**Artist**	**ID**	**Genre**	**Beispiel für**
Bip Bop Bip	Barrence Whitfield and The Savages	115	Blues	+ Arousal
Cold Summer Landscape	Blear Moon	488	Electronic	- Arousal
Clear Blue Sky	Chatham County Line	343	Country	+ Valence
Maia	Kreng	745	Jazz	- Valence

Tabelle 3.1: Beispielsongs

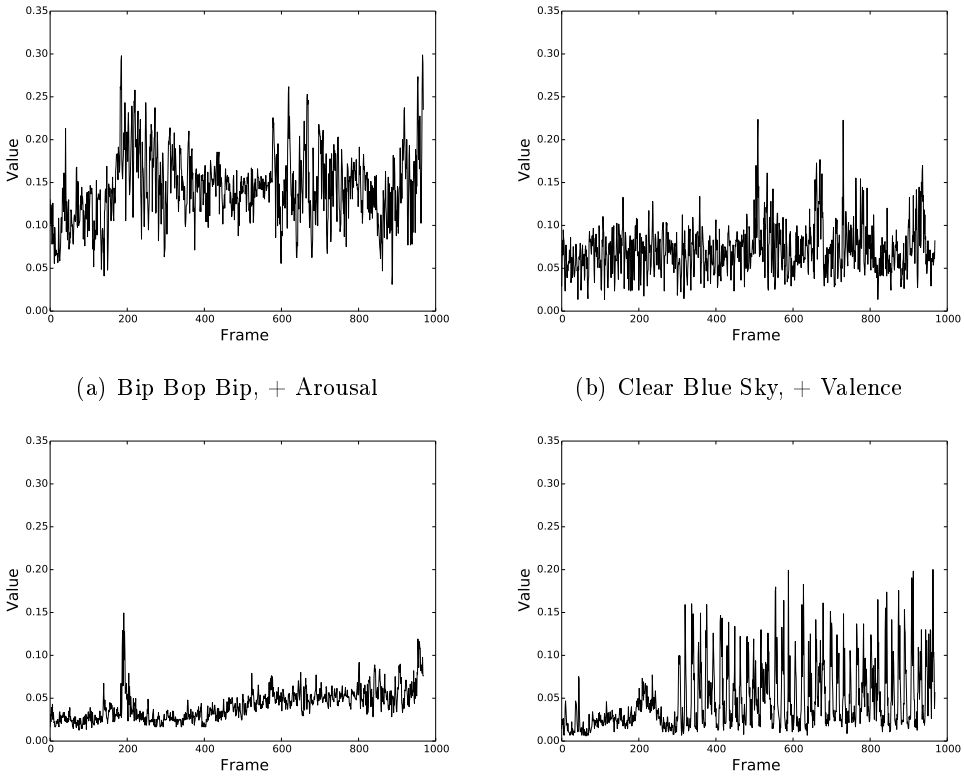
3.1 Nicht-cepstrale Merkmale

Merkmale lassen sich durch ihre Art in verschiedene Gruppen einteilen, über die sie eine Aussage treffen. In dieser Arbeit soll die Bedeutung von cepstralen Merkmalen (Kapitel 3.2) auf verschiedene Konstellationen von nicht-Cepstralen Merkmalen analysiert werden. Die nicht-Cepstralen Merkmale der hier verwendeten „1000 Songs Database“ [53] (Kapitel 2.2) wurden mithilfe von AMUSE (Advanced Music Explorer) [60] extrahiert. AMUSE ist ein Framework, welches eine Vielzahl von bekannten MIR Tools enthält und somit die Anwendung und das Zusammenspiel dieser vereinfacht. Darin enthalten sind Marsyas [57], jMIR [35], MusicMiner [39], MIR Toolbox [27], Chroma Toolbox [40] und RapidMiner [38]. Neben den cepstralen Merkmalen sind Zuordnungen in die Gruppen Energie, Klangfarbe

(*Timbre*), Harmonie und Melodie, sowie Tempo und Rhythmus vorgenommen worden, wie sie bereits in ähnlicher Weise von Tzanetakis und Cook [58] vorgeschlagen wurden.

3.1.1 Energie

Zero-crossing rate



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

(d) Maia, - Valence

Abbildung 3.2: Zero-crossing rate, 23.2 ms Fenstergröße

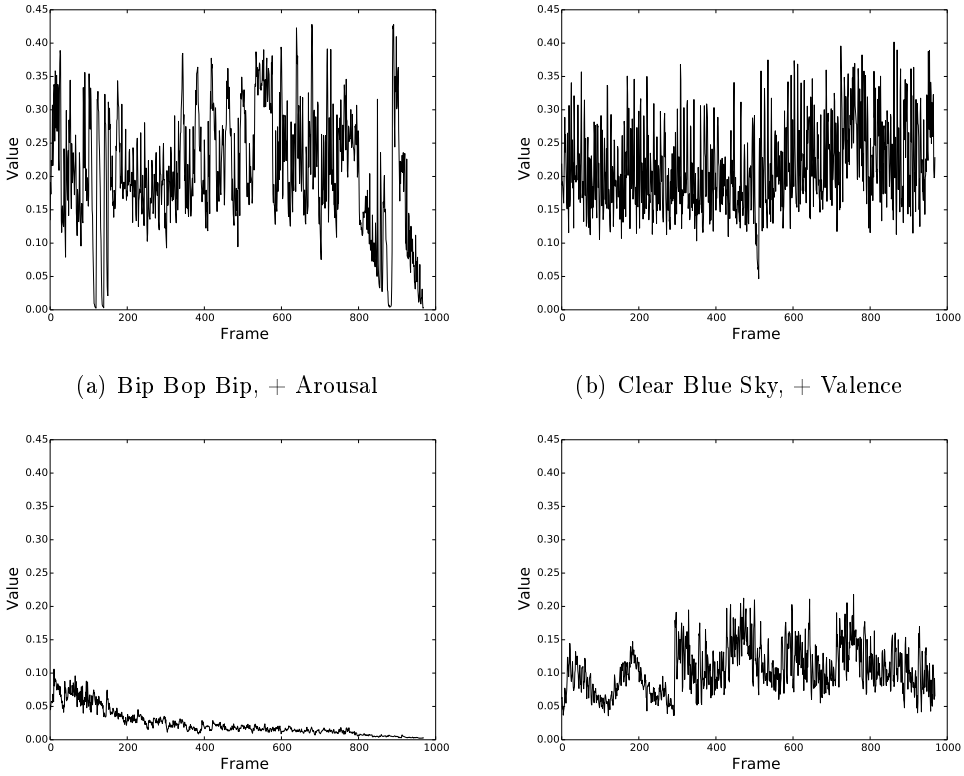
Das Merkmal *Zero-crossing rate* [55] beschreibt, wie oft ein Zeitsignal in einem gegebenen Fenster der Länge N die Nulllinie kreuzt. Dies geschieht in Gleichung 3.1 durch Zählen der Vorzeichenwechsel und anschließendes Normieren auf die Anzahl der Samples.

$$ZCR = \frac{1}{2(N-1)} \sum_{i=0}^{N-2} |sign(x(i+1)) - sign(x(i))| \tag{3.1}$$

Der daraus resultierende Zahlenwert gibt Aufschluss über das Vorkommen von hohen Frequenzen und kann damit als Maß für den Rauschanteil gewertet werden. „Cold Summer Landscape“ (Abbildung 3.2 c) ist ein sehr stilles und ruhiges Musikstück, die Zero-crossing rate fällt hier im Gegensatz zu „Bip Bop Bip“ (a) sehr gering aus. Die ab Frame 300 in

„Maia“ (d) auftretenden Schwankungen lassen sich auf das hochfrequente Rasseln zurückführen.

Root mean square



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

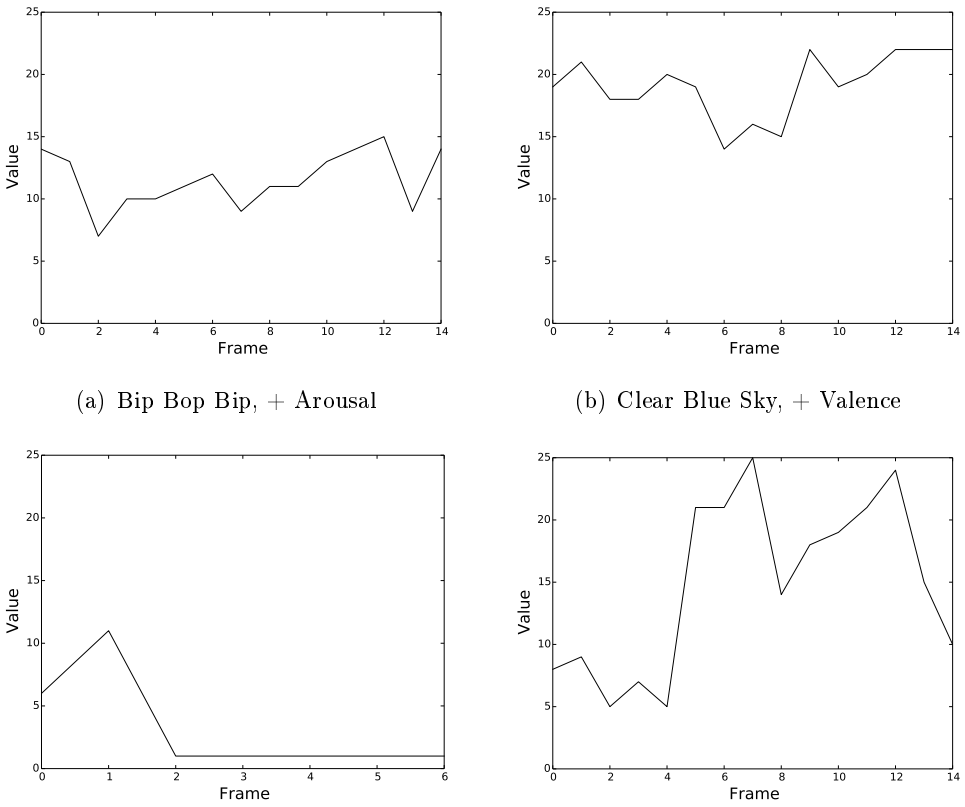
(d) Maia, - Valence

Abbildung 3.3: Root mean square, 23.2 ms Fenstergröße

Mit der Gleichung 3.2 wird über ein Zeitfenster N der quadratische Mittelwert (Root mean square) berechnet [55]. In der Elektrotechnik findet dieser RMS-Wert Anwendung, um den Effektivwert einer Wechselspannung zu bestimmen. Ebenso kann mit ihm der Energiegehalt eines Zeitsignals bestimmt werden. Abbildung 3.3 zeigt, dass die hier gezeigten Beispiele für starke (Grafik a) und geringe (Grafik c) Erregung eine hohe Korrelation zum RMS aufweisen, was vor allem an der Bedeutung der Lautstärke für die Erregung liegen kann.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x(i)^2} \tag{3.2}$$

RMS peak number



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

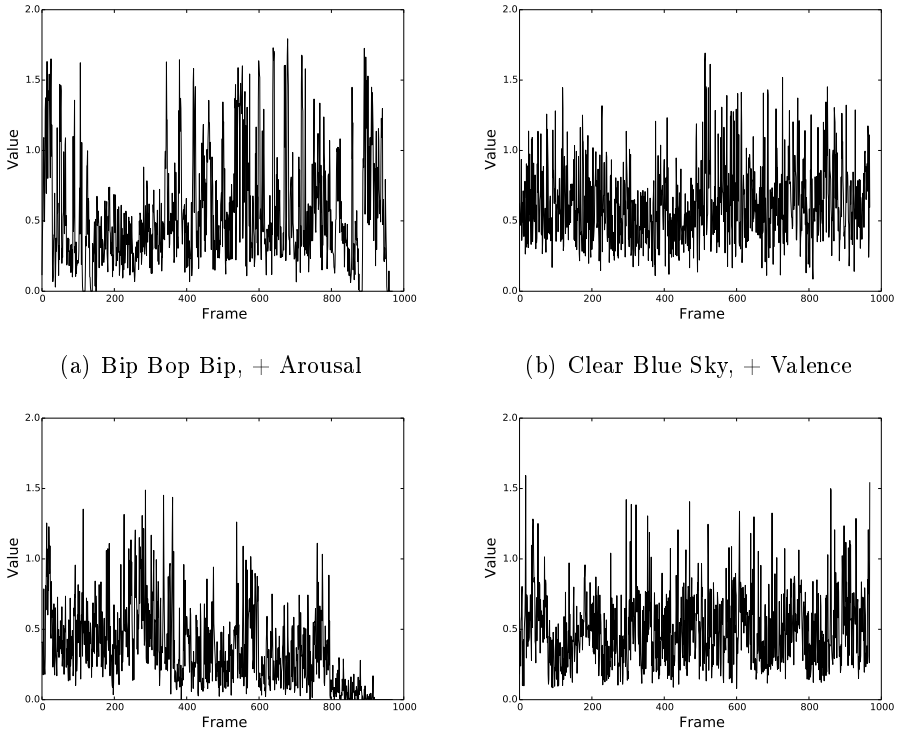
(d) Maia, - Valence

Abbildung 3.4: RMS peak number, 3000 ms Fenstergröße

Aus dem Verlauf von RMS über die Zeit ist das Merkmal der *RMS peak number* abgeleitet. Innerhalb eines Zeitfensters von z.B. 3 Sekunden (Beispiel von Abbildung 3.4) wird die Anzahl der lokalen Maxima gezählt. Das darauf aufbauende Merkmal *RMS peak number above mean amplitude* zählt nur die Momente, bei denen die Amplitude einen Mindestwert überschreitet. Dieser Schwellwert ist hierbei auf die Hälfte des in dem gesamten Signal vorkommenden Maximalwertes festgelegt. Die hohen Werte für Beispielsong b in Abbildung 3.4 können durch das Banjo als Zupfinstrument mit kurzzeitig klingenden Tönen erklärt werden. Da jede Note als Peak gezählt wird fällt die RMS Peak Number für Song a geringer aus, obwohl es bezüglich RMS eine ähnlichen Energieanteil besitzt.

3.1.2 Klangfarbe

Spectral irregularity



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

(d) Maia, - Valence

Abbildung 3.5: Spectral irregularity, 23.2 ms Fenstergröße

Gleichung 3.3 wurde 1994 von Krimphoff et al. [25] vorgestellt und beschreibt die Irregularität als Summe der Amplituden abzüglich dem Durchschnitt der 3 umliegenden Amplituden von Partialtönen (inklusive sich selbst).

$$Irregularity_{krimphoff} = \sum_{i=2}^{N-1} |a_i - \frac{a_{i-1} + a_i + a_{i+1}}{3}| \tag{3.3}$$

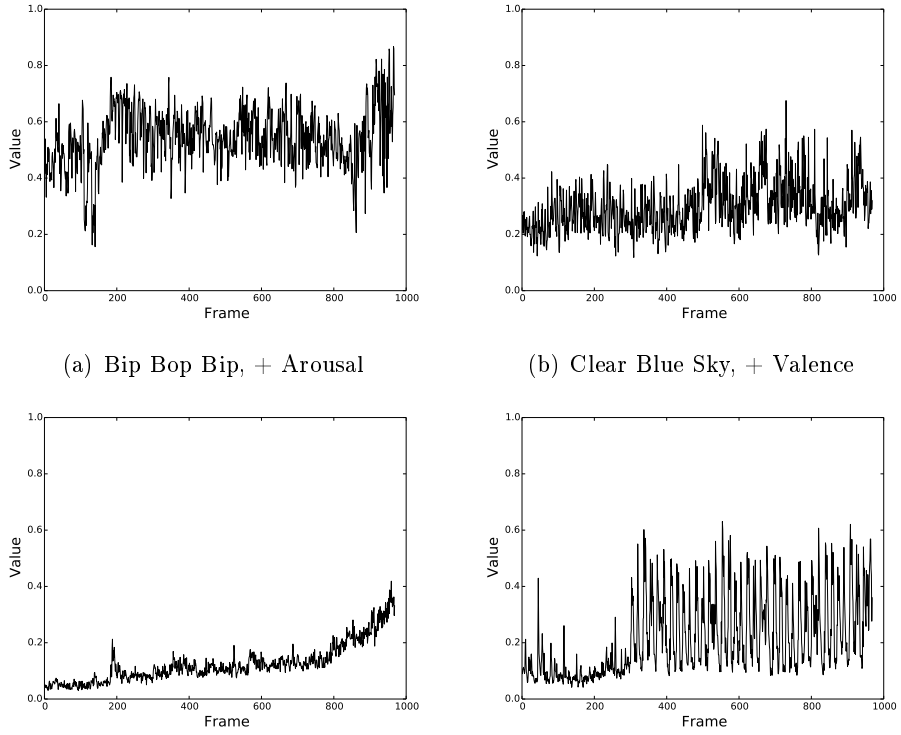
In der MIR Toolbox ist eine alternative Berechnung nach Jensen et al. [20] die Standartimplementierung, bei der die quadrierte Differenz verwendet wird. Anzumerken ist, dass für die Berechnung mit Gleichung 3.4 $a_{N+1} = 0$ ist.

$$Irregularity_{jensen} = \frac{\sum_{i=1}^{N}(a_i - a_{i+1})^2}{\sum_{i=1}^{N} a_i^2} \tag{3.4}$$

Anhand der Beispiele in Abbildung 3.5 lässt sich kein direkter Zusammenhang der Werte zu den vier Emotionsextrema erkennen. Eine Vorverarbeitung, um z.B. die Dynamik in Form

von Ableitung oder Standartabweichung mit einzuschließen, kann für die Aussagekraft dieses Merkmals erforderlich sein.

Spectral brightness



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

(d) Maia, - Valence

Abbildung 3.6: Spectral brightness, 23.2 ms Fenstergröße

Spectral Brightness [26] oder *High-frequency energy* [22] ist ein Merkmal, welches eine hohe Abhängigkeit zur Klangfarbe (engl. timbre) aufweist. Es beschreibt den Energieanteil oberhalb einer gewählten Cutoff-Frequenz von 1500 Hz [26] oder 3000 Hz [22]. Abbildung 3.7 zeigt das Verhältnis im Frequenzspektrum eines möglichen Zeitfensters. Ein Musiksignal mit hoher *Spectral brightness* erzeugt eine Wahrnehmung von scharfer Klangfarbe, ein geringer *Spectral brightness*-Wert führt hingegen zu einer weichen Wahrnehmung [22].

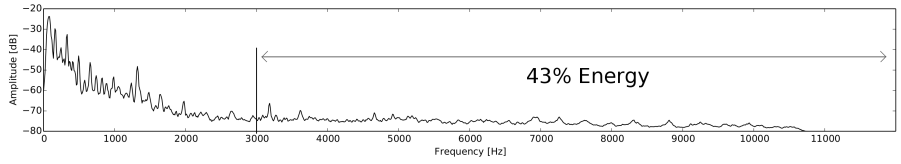


Abbildung 3.7: Spectral brightness im Frequenzspektrum

3.1.3 Harmonie und Melodie

Harmonic change detection function

Für die Erkennung von Akkordwechseln wurde die Methode der *Harmonic change detection function* (HCDF) von Harte und Sandler [18] vorgeschlagen. Ein Audiosegment wird zunächst mittels Konstanter Q-Transformation [8] in den Frequenzbereich überführt (Abbildung 3.8). Die einzelnen Filter haben hier im Gegensatz zur Fourier-Transformation logarithmische Abstände. Die Zentren der Filter können dadurch mit den Frequenzen des Zwölftonsystems zur Übereinstimmung gebracht werden, um so eine gleichbleibende Auflösung über alle Töne zu erhalten. Durch Binning in zwölf Gruppen wird ein Chromagram erstellt, welches Aufschluss über das Vorkommen der Halbtöne gibt. Pro Audiosegment entsteht bis zu diesem Schritt ein 12-dimensionaler Chroma-Vektor. Dieser beschreibt jeweils einen Punkt im Zirkel der Dur-Dreiklänge („Major Thirds“), Moll-Dreiklänge („Minor Thirds“) und im Quintenzirkel („Fifths“), wie in Abbildung 3.9 am Beispiel eines Vektors A dargestellt.

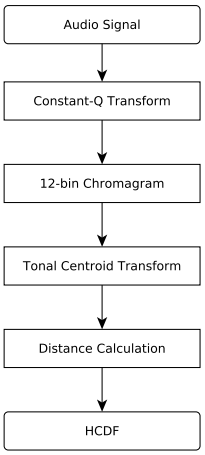


Abbildung 3.8: HCDF Blockdiagramm

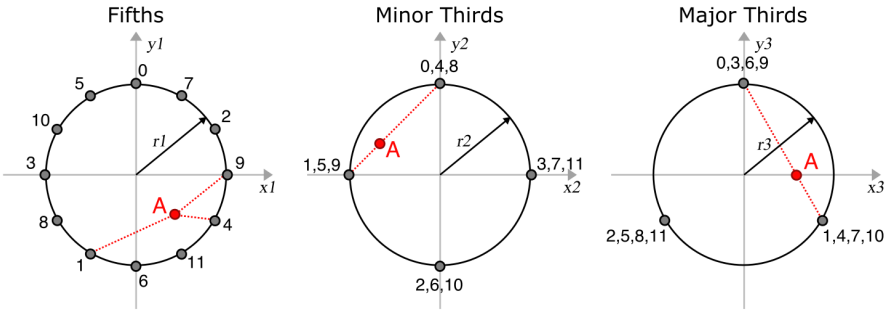
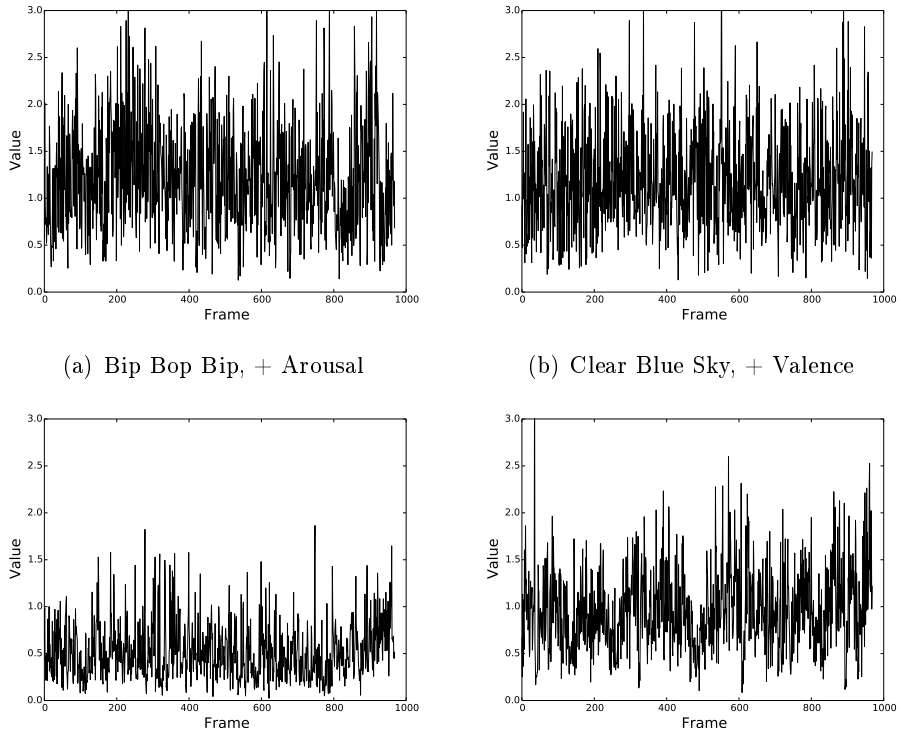


Abbildung 3.9: 6-D Tonaler Raum als drei Kreise [18]

Jeder Punkt innerhalb dieser drei Kreise kann wiederum durch ein 2-D Koordinatensystem beschrieben werden. Durch Aneinanderreihen der sechs Komponenten ergibt sich ein 6-D Tonal Centroid Vector $C = (x_1, y_1, x_2, y_2, x_3, y_3)^T$. Dieser findet sich alleine ebenfalls als Merkmal der Harmonie wieder. Um jedoch Änderungen in der Harmonie zu erkennen wird



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

(d) Maia, - Valence

Abbildung 3.10: Harmonic change detection function, 23.2 ms Fenstergröße

nun der euklidische Abstand zwischen C_{i+1} und C_{i-1} berechnet. i ist dabei der Vektor des i-ten Audiosegments.

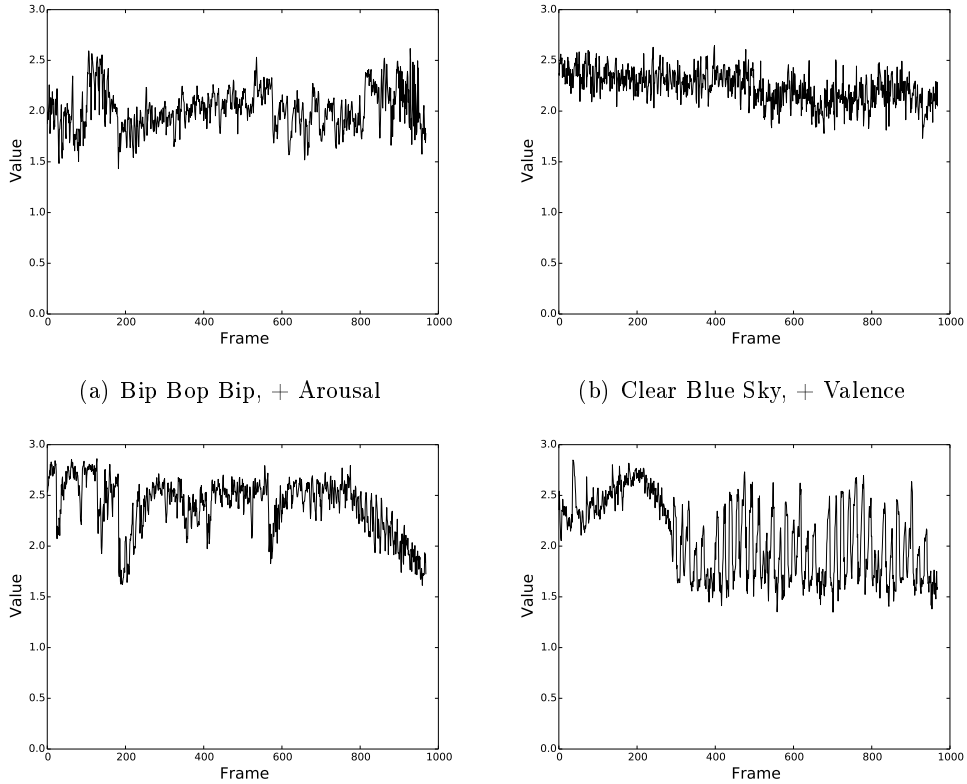
$$HCDF_i = \sqrt{\sum_{d=0}^{5}[C_{i+1}(d) - C_{i-1}(d)]^2} \tag{3.5}$$

Abbildung 3.10 zeigt die durch Gleichung 3.5 berechneten Änderungen in der Harmonie für jedes 23.2 ms lange Zeitfenster.

Chroma DCT-reduced log pitch

Der in Kapitel 3.1.3 angesprochene Chroma-Vektor beschreibt das Vorkommen der zwölf Halbtöne, jedoch wird er von möglichen Obertönen beeinflusst, welche von Instrument zu Instrument unterschiedlich stark ausgeprägt sind [41, 42]. Um die Abhängigkeit zur Klangfarbe eines Instruments zu reduzieren wird das Spektrum zuerst linear transformiert, sowie die ersten Koeffizienten auf Null gesetzt. Die Rücktransformation liefert anschließend ein zur Klangfarbe unabhängigeres Spektrum auf welchem der Chroma-Vektor wie zuvor berechnet wird.

Angles / Distances in phase domain



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

(d) Maia, - Valence

Abbildung 3.11: Angles in phase domain, 23.2 ms Fenstergröße

Für die Betrachtung von (Audio-) Signalen ist es oft hilfreich, sie vom zweidimensionalen Vektorraum in einen anderen zu überführen. Neben dem bereits bekanntem Frequenzwird der Phasenraum von Mierswa und Morik [37] zur Analyse der Signaldynamik vorgestellt.

$$p_i = (x(i), x(i+d), x(i+2d), ...x(i+(m-1)d))^T \tag{3.6}$$

Ein Vektor p_i zum Zeitpunkt i im Phasenraum wird durch Zusammenfügen der Amplituden des Zeitsignals x gebildet. Hierbei bestimmt d die Verzögerung der einzelnen Elemente zueinander und m die Dimension des Phasenvektors. Die Überführung in den Phasenraum erlaubt es, neue Merkmale darauf zu generieren. Die hier gezeigten Features berechnen Winkel (Gl. 3.10) und Distanzen (Gl. 3.11) aufeinanderfolgender Phasenvektoren.

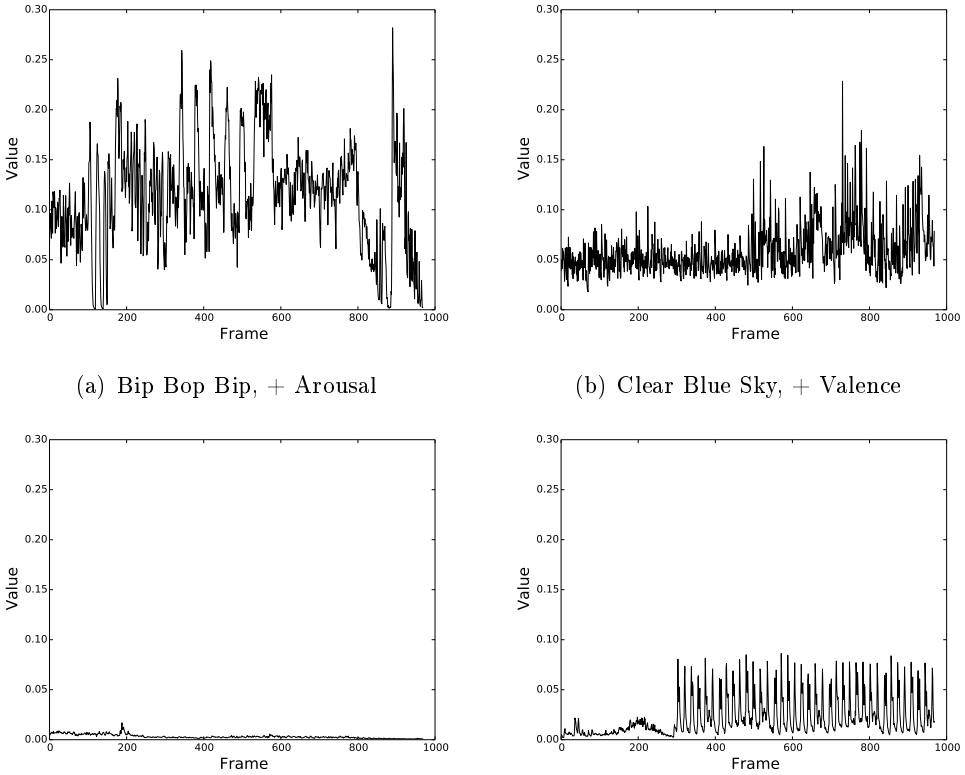
$$p'_i = p_{i-1} - p_i \tag{3.7}$$

$$p''_i = p_{i+1} - p_i \tag{3.8}$$

Die Winkel zwischen zwei Phasenänderungen p' und p'' werden in Gleichung 3.9 durch das Skalarprodukt berechnet. Das endgültige Merkmal *Average Angle* entsteht durch Mitteln mehrerer Winkel α_i innerhalb eines gegebenen Zeitfensters N (Gl. 3.10).

$$\alpha_i = cos^{-1} \frac{{p'_i}^T p''_i}{||p'_i||\,||p''_i||} \tag{3.9}$$

$$Average\ Angle = \frac{1}{N-2-(m-1)d} \sum_{i=1}^{N-2-(m-1)d} |\alpha_i| \tag{3.10}$$



(a) Bip Bop Bip, + Arousal

(b) Clear Blue Sky, + Valence

(c) Cold Summer Landscape, - Arousal

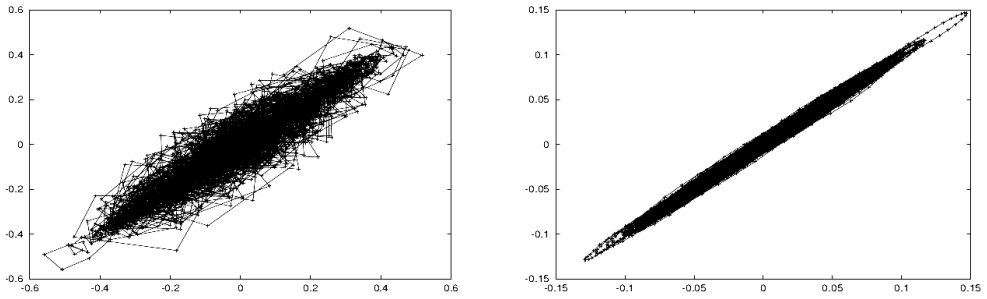
(d) Maia, - Valence

Abbildung 3.12: Distances in phase domain, 23.2 ms Fenstergröße

Die Länge des Vektors p'' beschreibt die Distanz zweier aufeinander folgender Phasenvektoren und wird in Gleichung 3.11 ebenfalls gemittelt.

$$Average\ Distance = \frac{1}{N-2-(m-1)d} \sum_{i=1}^{N-2-(m-1)d} ||p'_i|| \tag{3.11}$$

Die Analyse eines Audiosignals im Phasenraum hat sich für die Unterscheidung von Klassik zu Pop oder Rock als hilfreich erwiesen [37, 42]. Abbildung 3.13 zeigt aneinander gereihte zweidimensionale Phasenvektoren von Beispielsongs aus Pop und Klassik. Das



(a) pop music (b) classic

Abbildung 3.13: Phasenraumdarstellung eines Musikstücks aus Pop (a) und Klassik (b) [37]

Beispiel links wirkt eher ungeordnet, wohin gegen die Phasenvektoren des klassischen Musikstücks eine Ellipse bilden. Abbildungen 3.11 und 3.12 entsprechen jeweils den gemittelten Richtungen bzw. Längen der Phasenvektoren eines Zeitfensters.

3.1.4 Tempo und Rhythmus

Estimated onset number per minute

Zur Analyse von Rhythmik ist das Einsetzten von Tönen oder Schlägen ein wichtiger Indikator. Diese Onsets können durch das Auftreten von Energie-Peaks oder Änderungen der Klangfarbe erkannt werden [12]. Aus den daraus resultierenden Onset-Zeitpunkten wird über ein relativ großes Zeitfenster (hier 10 Sekunden) dessen Anzahl ermittelt und durch den Faktor 6 auf eine Minute hoch gerechnet. Die berechneten Onsets pro Minute sind in Tabelle 3.2 zu finden. Onsets finden ebenfalls in der Beat Detection (Kapitel 3.1.4) Anwendung und sind Grundbestandteil der Zwischen-Onset Methode (Kapitel 3.3.2) zur Vorverarbeitung einer Vielzahl von Merkmalen.

Estimated beat number per minute

Als Beat kann die Menge periodisch vorkommender Schläge bezeichnet werden, zu der sich einfach gesagt ein Klatschrhythmus finden lässt. Eine verbreitete Methode ist es, die Periodizität der Onsets beispielsweise mit Hilfe von Autokorrelation zu ermitteln [12, 16].

$$\rho(l) = \sum_{i=0}^{N-1} x(i)x(i-l), \qquad 0 \leq l \leq N-1 \tag{3.12}$$

$\rho(l)$ zeigt in Gleichung 3.12 innerhalb eines Zeitfensters N zu den Verschiebungen l ein Maximum, an denen sich Onsets wiederholen. Statt fester Onsetzeiten kann die Autokorrelation auch auf einem kontinuierlichen Energieverlauf E_j angewandt werden, welcher durch

Kurzzeit-Fourier-Transformation ($STFT$) auf typischerweise 5-10 logarithmisch verteilten Frequenzbändern j berechnet wird (Gl. 3.13) [16].

$$E_j(i) = \sum_{k \in \kappa_j} |STFT_x^w(i,k)|^2 \tag{3.13}$$

$$D_j(i) = \frac{E_j(i+1) - E_j(i-1)}{3} \tag{3.14}$$

Durch den Einsatz von Linearer Regression kann E_j in eine für Event und somit Beat Detection bessere Darstellung D_j gebracht werden (Gl. 3.14). Maxima auf D_j zeigen den exakten Beginn und nicht nur den lautesten Punkt eines Ereignisses. Anschließend zählt *Estimated beat number per minute* (BPM) ebenfalls die erkannten Beats über ein größeres Fenster und normalisiert sie auf eine Minute. Tabelle 3.2 zeigt Onset und Beat number per minute der vier Beispielsongs im direkten Vergleich. Im Song „Cold Summer Landscape“ wurden ab etwa der Hälfte des analysierten Ausschnitts aufgrund nicht ausreichend starker Noteneinsätze keine Onsets erkannt. „Bip Bop Bip“ und „Clear Blue Sky“ sind Songs mit einem ausgeprägten Rhythmus mit konstantem Tempo, was an der analysierten Beat Number per Minute ersichtlich wird. Da für die Erkennung der BPM eine Periodizität gesucht wird, kann die Anzahl der Onsets trotz stetigem Tempo stark variieren.

Titel	**Beispiel für**	**Onset Number**	**Beat Number**
Bip Bop Bip	+ Arousal	340	127
		322	133
		421	133
		334	133
Cold Summer Landscape	- Arousal	415	92
		92	92
		/	87
		/	69
Clear Blue Sky	+ Valence	444	150
		444	150
		427	150
		490	150
Maia	- Valence	386	138
		438	150
		507	144
		524	144

Tabelle 3.2: *Estimated onset-* und *beat number per minute* der Beispielsongs

3.2 Cepstrale Merkmale

Das erstmals von Bogert, Healy und Tukey [7] benannte *Cepstrum* wird durch die inverse Fourier-Transformation des logarithmierten und quadrierten Spektrums eines Signals gebildet [43, 32]. Der Begriff Cepstrum ist neben anderen in dieser Arbeit eingeführten Begriffen wie „Quefrency“ oder „Rhamonics“ ein Wortspiel zu „Spectrum“, bei dem die vier ersten Buchstaben des Wortes vertauscht wurden. Auf diese Art wollten die Autoren ausdrücken, dass es sich dabei weder um ein Zeitsignal handelt, noch das Frequenzspektrum in seinem üblichen Verständnis gemeint ist. Zuerst zur Erkennung von Echos in einem Signal, die in dessen Cepstrum als Maximum auftreten [43], wird diese Art der Betrachtung neben Zeit- und Frequenzdarstellung gerne für die Sprach- und Musikanalyse angewandt [30] (siehe „Mel-Frequency Cepstral Coefficients“ in Kapitel 3.2.1). Der Grund, die Amplituden des Frequenzspektrums zu logarithmieren ist mit dem menschlichen Gehör und der Wahrnehmung von Tönen begründet. Je lauter ein Signal, desto geringer werden Änderungen in der Lautstärke empfunden. In Hinblick auf die Mathematik überführt eine Logarithmierung die Multiplikation in eine Addition [43, 10]. Das Anwenden von Filtern wird somit vereinfacht. Die Mathematische Definition des Cepstrums (Gl. 3.15) erlaubt eine imaginärwertige Transformation [33].

$$x_c(q) = \frac{1}{N} \sum_{i=0}^{N-1} ln(|X(i)|^2) e^{j \frac{2\Pi q i}{N}} \tag{3.15}$$

Ein segmentiertes Zeitsignal $x(i)$ wird zuvor mittels Diskreter Fourier-Transformation (DFT) in das Spektrum $X(i)$ überführt. $x_c(q)$ beschreibt nach Einsetzen in die Gleichung das Signal im Cepstrum über die Quefrenzen q. Für die Charakterisierung des Spektrums genügt es jedoch den Realteil zu betrachten.

3.2.1 MFCC

Die *Mel-Frequency Cepstral Coefficients* (MFCC) sind ein beliebtes Merkmal für die Erkennung von Sprache [30]. Durch Skalierung der Quefrenzen [7, 43] („Frequenzen“ im Cepstrum) durch die Mel-Skala (Gl. 3.16, Abbildung 3.14) wird die Signaldarstellung weiter an die der menschlichen Wahrnehmung angepasst.

$$Mel(f) = 2595 log_{10}(1 + f/700) \tag{3.16}$$

Durch Befragungen und Experimente wurde festgestellt, dass die Tonhöhe, wie sie empfunden wird, über 1 kHz logarithmisch zur tatsächlichen Frequenz verläuft, darunter wird der Zusammenhang als eher linear beschreiben [14] [30]. Gleichung 3.16 ist eine Approximation von Frequenz (in Hz) zur Tonheit (Einheit Mel). Der Begriff „Mel“ kommt von „Melody“ und soll einen Bezug zur wahrgenommenen Tonhöhe suggerieren.

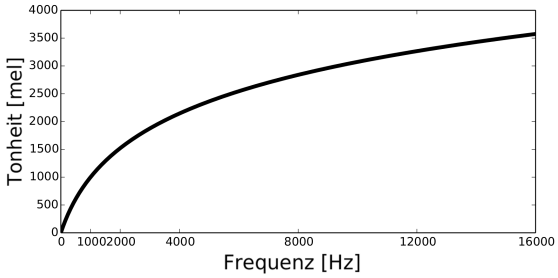


Abbildung 3.14: Gleichung 3.16: Zusammenhang zwischen Frequenz und Tonheit [45] [14]

Um nun Merkmale für die Sprach- oder Musikanalyse zu gewinnen, wird durch diskrete Kosinustransformation eine Menge von unkorrelierten Koeffizienten erzeugt, von denen z.B. die ersten 13 [30] als Merkmalsvektor verwendet werden. Der genaue Ablauf ist in Abbildung 3.15 dargestellt.

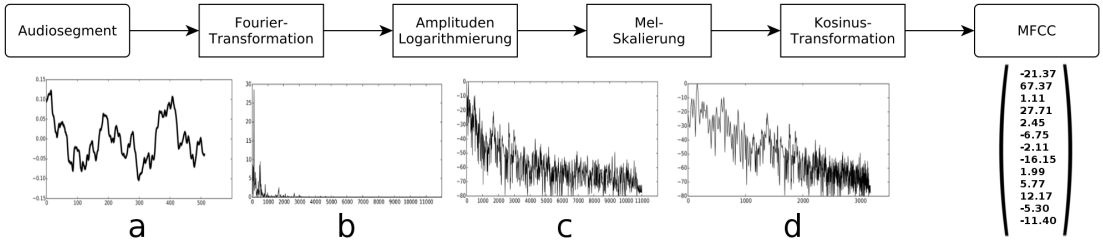


Abbildung 3.15: MFCC Extraktion

Ein MFCC Vektor wird jeweils für ein Zeitfenster erzeugt, typischerweise in einer Größenordnung von 20ms [30]. Das jeweilige Audiosegment muss zunächst mittels diskreter Fourier-Transformation (DFT) in den Frequenzbereich überführt werden. Bei der DFT ist die Anzahl der Komponenten auf z.B. 256 [30] beschränkt. Ein weiterer Schritt hin zum Cepstrum ist die Logarithmierung der Amplituden, wodurch leise Bereiche stärker angehoben werden (vgl. Übergang von Grafik b zu c in Abbildung 3.15). An dieser Stelle kommt die Mel-Skala zum Einsatz, wodurch eine Stauchung der Frequenzen über 1 kHz stattfindet. Die zuvor 256 Komponenten werden dafür per Binning in z.B. 40 Frequenzgruppen [30] eingeteilt und gemittelt. Dies führt zu einer zusätzlichen gewünschten Glättung. Die entstandenen 40 Komponenten sind allerdings stark untereinander abhängig. Um sie zu de-korrelieren wird eine diskrete Kosinustransformation (DCT) ausgeführt. Sie ist eine gute Approximation der Karhunen-Loève Transformation bzw. Hauptkomponentenanalyse und liefert eine Anzahl von z.B. 13 Koeffizienten [30], welche als MFCC-Merkmalsvektor bezeichnet werden. Obwohl die Mel-Frequenz Skala zuerst nur für die Sprachanalyse ent-

wickelt wurde, konnten deutliche Vorteile gegenüber einer linearen Skalierung auch auf Musiksignalen nachgewiesen werden [30].

3.2.2 Spectral Contrast

Ursprünglich waren *Mel-Frequency Cepstral Coefficients* für die Sprachanalyse gedacht, brachten in der Musikanalyse dennoch gute Resultate [30]. Als Kritik sehen Jiang et al. [21] die Verwendung der Mel-Skala, sowie das Mitteln der spektralen Amplituden innerhalb der Frequenzbänder. Das in dieser Arbeit vorgestellte *Octave-Based Spectral Contrast Feature* (OBSC) teilt die Frequenzen, wie der Name suggeriert, in ein Frequenzband pro Oktave ein. Harmonische Anteile werden im Spektrum als Maximum sichtbar, Rauschen als Minimum. Um die Information der relativen Verteilung von harmonischen und nicht-harmonischen Anteil zu erhalten, werden Peak und Valley eines jeden Frequenzbandes berechnet. Der namensgebende *Spectral Contrast* errechnet sich aus dessen Differenzen. Der endgültige Merkmalsvektor wird durch den Einsatz der Karhunen-Loève Transformation aus Spectral Contrast und Valleys gebildet.

Abbildung 3.16: Octave-Based Spectral Contrast Extraktion

Abbildung 3.16 zeigt den Ablauf der OBSC-Berechnung für ein Zeitfenster. In [21] wird eine Länge von 200ms und eine 50 prozentige Überlappung angegeben. Eine Evaluierung des OBSC Merkmals ist in Kapitel 5.3 zu finden. Nach der Segmentierung wird das jeweilige Signal per Fourier-Transformation in den Frequenzbereich überführt. Anschließend werden die Frequenzen in $k = 6$ nicht überlappende Bänder in Abständen einer Oktave mit 0Hz bis 200Hz, 200Hz bis 400Hz, 400Hz bis 800Hz, 800Hz bis 1.6kHz,1.6kHz bis 3.2kHz und 3.2kHz bis 8kHz eingeteilt. Es ergeben sich so Zuteilungen der Frequenzamplituden X in sechs Vektoren $X_k = (X_{k,1}, X_{k,2}, ...X_{k,N_k})$ mit $k = 1, 2, ...6$, wobei N_k die jeweilige Anzahl der im Band k enthaltenden Werte angibt. Zur Berechnung von Peak und Valley müssen zunächst die Frequenzbänder X_k absteigend nach Amplituden sortiert werden. $X'_k = (X'_{k,1}, X'_{k,2}, ...X'_{k,N_k})$ ist der resultierende sortierte Vektor. Es gilt $X'_{k,1} > X'_{k,2} > ...X'_{k,N_k}$. Gleichungen 3.17 und 3.18 dienen zur Berechnung von Peak (P_k) und Valley (V_k) auf X'_k.

$$P_k = log(\frac{1}{\alpha N_k} \sum_{i=1}^{\alpha Nk} X'_{k,i}) \tag{3.17}$$

$$V_k = log(\frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N k} X'_{k,N_k-i+1}) \tag{3.18}$$

Hier wird auf den ersten Blick wie bei der MFCC Berechnung der Mittelwert über die Amplituden des entsprechenden Frequenzbandes berechnet. Der Wert α mit $0 < \alpha \leq 1$ beschränkt allerdings mit sinkendem α die Menge der in die Mittelwertbildung einfließenden Werte. Da X'_k Amplituden in absteigender Sortierung enthält, stellen anders gesagt Peak und Valley das Maximum und Minimum mit Einfluss der umliegenden Amplituden dar. Werte für α sind eher gering zu wählen. In [21] wurden Tests mit $\alpha = 0.02$ bis $\alpha = 0.2$ gemacht, es konnte allerdings kein signifikanter Einfluss auf die Erkennungsraten der dort verwendeten Genre Klassifizierung festgestellt werden.

$$SC_k = P_k - V_k \tag{3.19}$$

Der Spectral Contrast berechnet sich als Differenz zwischen Peak und Valley (Gl. 3.19). Ein Vorläufiger Merkmalsvektor wird wie folgt aus Contrast und Valley zusammengesetzt.

$$OBSC = (SC_1, SC_2, ...SC_6, V_1, V_2, ...V_6) \tag{3.20}$$

Wie schon bei der MFCC Berechnung, sind die einzelnen Komponenten dieses Vektors untereinander korreliert. Eine Karhunen-Loève Transformation wird auf dem $OBSC$ Vektor angewandt, um ihn zu de-korrelieren.

3.2.3 CMRARE

Cepstral Modulation RAtio REgression (CMRARE) [32] Merkmale stellen eine weitere Art der Darstellung des Cepstrums dar. Ihr Ziel ist es im Gegensatz zu MFCC und OBSC die feine harmonische Struktur des Cepstrums zu erhalten, indem auf ihm ein Modulationsspektrum berechnet wird. Dazu verwenden Martin und Nagathil die DFT mit Sliding Window, um so den Verlauf über die Quefrenzen einfließen zu lassen. Die Modulationsspektra werden daraufhin durch Division auf das Nullte Modulationsfrequenzband normalisiert, welches in die darauf folgende Regression nicht mit einfließt. Dadurch wird die Unabhängigkeit von der Lautstärke des Musiksignals erreicht. Die daraus resultierenden Cepstral Modulation Ratios können anschließend durch ein Polynom mit Grad p durch die Methode der kleinsten quadratischen Differenzen approximiert werden. Der CMRARE Merkmalsvektor setzt sich aus den errechneten Polynomen zusammen. Auswirkungen von Polynomgrad auf die Emotionsvorhersage werden in Kapitel 5.4 untersucht.

3.3 Extraktion und Vorverarbeitung

Dem „1000 Songs“-Datensatz liegt eine Menge von bereits extrahierten Features bei. Für die im Zusammenhang mit dieser Arbeit in Kapitel 5 gemachten Studien sollen jedoch

Auswirkungen der Extraktionsparameter zusätzlich betrachtet werden. Ein Großteil davon bezieht sich auf ein voranschreitendes Zeitfenster, dessen Länge Einfluss auf die Vorhersage haben kann. Die Extraktion aller nicht-cepstralen Merkmale fand mit AMUSE statt (siehe Abschnitt 3.1). Eine Übersicht der verwendeten Merkmale zeigen Tabellen 3.3, 3.4, 3.5 und 3.6. Dort aufgelistet sind die insgesamt 43 verschiedenen Merkmale der vier Gruppen *Energy*, *Timbre*, *Harmony and Melody* und *Tempo and Rhythm* von denen 24 zusätzlich zu unterschiedlichen Extraktionsfenstern vorhanden sind. Das cepstrale Merkmal CMRARE (Abschnitt 3.2.3) besitzt neben Fenstergröße den Polynomgrad als wichtigen Parameter, der ebenso in die Betrachtungen einbezogen werden sollte. Zur Anwendung von AMUSE wurden ganze Musikstücke aus dem Datensatz zuerst auf 22050 Hz herunter gerechnet und die Merkmale ohne Überlappung auf den in Tabellen 3.3, 3.4, 3.5 und 3.6 angegebenen Zeitfenstern extrahiert. Anschließend brauchten nur die Merkmale der 45 s Segmente, die bei den Umfragen zur Gewinnung der Annotationen verwendet wurden, behalten werden. Für Musiksegmente die nicht mit dem original Musikstück beginnen bzw. enden konnten somit Merkmale mit Zeitfenstern die über die des 45 s Segments hinausgehen berechnet werden. Dazu zählen u.A. *Estimated onset-* und *Estimated beat number per minute* mit etwa 10s Fensterlänge. Welches Tool von AMUSE zur Extraktion verwendet wurde kann in [59] ab Seite 135 nachgeschlagen werden. Die *Mel-Frequency Cepstral Coefficients* (Abschnitt 3.2.1) sind ein weit verbreitetes Merkmal in der Audioanalyse und besitzen mehrere Parameter, deren Optimierung die Vorhersage positiv beeinflussen kann. Dazu zählen Längen der Extraktionsfenster und deren Überlappung zueinander, Anzahl der Koeffizieten, FFT- und Mel-Bins, sowie der betrachtete Frequenzbereich. Ebenso ist das Merkmal *Octave-Based Spectral Contrast* (Abschnitt 3.2.2) für die Analyse von Musiksignalen vielversprechend. Dort lassen sich die Frequenzen der einzelnen Bänder und der α-Wert, der das Quantil beschreibt, angeben. MFCC, sowie OBSC wurden mit *librosa*[1] [34] in Version 0.4.1 extrahiert, einer Python Bibliothek mit vielen Funktionen für die Audioanalyse. Um Rechenzeit zu sparen wurden diese beiden Merkmale auf den der „1000 Songs Database“ ebenso beiliegenden 45 s Audioclips mit 44100 Hz Samplerate extrahiert. Angaben von Zeitpunkten in Samples beziehen sich somit für MFCC und OBSC auf 44100 Hz, für CMRARE und alle nicht-cepstralen Merkmale auf 22050 Hz. Entsprechende Zeiten werden deshalb im Folgenden zur besseren Vergleichbarkeit mit angegeben.

[1] librosa: `https://github.com/bmcfee/librosa`, aufgerufen am 12.11.2015

Merkmal	Dim.	Fenster (Samples)	Fenster (ms)
Zero-crossing rate	1	512, 1024, 2048	23.2, 46.4, 92.9
Root mean square	1	512, 1024, 2048	23.2, 46.4, 92.9
Low energy	1	512, 1024, 2048	23.2, 46.4, 92.9
RMS peak number in 3 seconds	1	66150	3000
RMS peak number above half of maximum peak in 3 seconds	1	66150	3000
Sub-band energy ratio	4	512, 1024, 2048	23.2, 46.4, 92.9

Tabelle 3.3: Verwendete Merkmale der Gruppe *Energy*

Merkmal	Dim.	Fenster (Samples)	Fenster (ms)
Spectral centroid	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral irregularity	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral bandwidth	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral skewness	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral kurtosis	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral crest factor	4	512, 1024, 2048	23.2, 46.4, 92.9
Spectral flatness measure	4	512, 1024, 2048	23.2, 46.4, 92.9
Spectral extent	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral flux	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral brightness	1	512, 1024, 2048	23.2, 46.4, 92.9
Sensory roughness	1	512, 1024, 2048	23.2, 46.4, 92.9
Spectral slope	1	512, 1024, 2048	23.2, 46.4, 92.9
Angles in phase domain	1	512, 1024, 2048	23.2, 46.4, 92.9
Distances in phase domain	1	512, 1024, 2048	23.2, 46.4, 92.9

Tabelle 3.4: Verwendete Merkmale der Gruppe *Timbre*

3.3.1 Vorverarbeitung

Der zu Anfang diesen Kapitels in Abbildung 3.1 als Blockdiagramm dargestellte Ablauf deutet an, dass für den endgültig verwendeten Merkmalsvektor mehrere Vektoren, der aus den kleineren Zeitfenstern entstandenen Merkmale, zusammengefasst werden müssen. Dies führt dazu, dass die zeitliche Abhängigkeit, die durch einfaches aneinanderreihen entstehen würde, entfällt. Es soll, anders gesagt, jedes angewandte Merkmal für ein gesamtes Musikstück oder vorherzusagendes Audiosegment bestimmt werden. Um dies zu erreichen, wird der Durchschnitt über alle Dimensionen der Merkmale berechnet. In [3] ermitteln die

Merkmal	Dim.	Fenster (Samples)	Fenster (ms)
Tristimulus	2	512	23.2
Inharmonicity	1	512, 1024, 2048	23.2, 46.4, 92.9
Major/minor alignment	1	512, 1024, 2048, 4096	23.2, 46.4, 92.9, 185.8
Strengths of major keys	12	512, 1024, 2048, 4096	23.2, 46.4, 92.9, 185.8
Strengths of minor keys	12	512, 1024, 2048, 4096	23.2, 46.4, 92.9, 185.8
Tonal centroid vector	6	512, 1024, 2048, 4096	23.2, 46.4, 92.9, 185.8
Harmonic change detection function	1	512, 1024, 2048, 4096	23.2, 46.4, 92.9, 185.8
Chroma DCT-Reduced log Pitch	12	4410	200
Number of different chords in 10s	1	220500	10000
Number of chord changes in 10s	1	220500	10000
Shares of the most frequent 20, 40 and 60 percents of chords with regard to their duration	3	220500	10000

Tabelle 3.5: Verwendete Merkmale der Gruppe *Harmony and Melody*

Merkmal	Dim.	Fenster (Samples)	Fenster (ms)
Characteristics of fluctuation patterns	7	32768	1486.1
Rhythmic clarity	1	66150	3000
Estimated onset number per minute	1	229376	10402.5
Estimated beat number per minute	1	229376	10402.5
Estimated tatum number per minute	1	229376	10402.5
Tempo based on onset times	1	32768	3000
Five peaks of fluctuation curves summed across all bands	5	229376	10402.5

Tabelle 3.6: Verwendete Merkmale der Gruppe *Tempo and Rhythm*

Autoren neben Mittelwert auch die Varianz, welche Auskunft über die Streuung der Werte gibt und fügen sie dem Merkmalsvektor hinzu. Andere statistische Kennzahlen wie z.B. Median, Differenz zwischen Minimum und Maximum oder beliebige Quantile sind ebenso möglich.

3.3.2 Zwischen-Onset Methode

Mit nur wenigen Millisekunden sind die Extraktionsfenster vieler cepstraler und nicht-cepstraler Merkmale sehr gering, sodass sie jeweils nur einen geringen Teil eines vorherzusagenden Audioclips wiedergeben. Dadurch fallen extreme Werte, wie zu Zeitpunkten

eines Noteneinsatzes (*Onsets*) oder Schlags durch perkussive Instrumente, für Merkmale, die z.B. nur Harmonie betrachten, bei der Vorverarbeitung (Abschnitt 3.3.1) negativ ins Gewicht. Das *Attack-Decay-Sustain-Release* Modell [44] hilft, das Verhalten bezüglich der Amplitude von Tönen über den zeitlichen Verlauf zu verstehen. *Attack* bestimmt die Zeit die ein Ton zum *Anschwellen* benötigt, gefolgt vom Abfall (*Decay*) (Abbildung 3.17). Am Beispiel eines Klaviers lässt sich die Ausklingzeit (*Sustain*) als die Dauer beschreiben, die der Ton nach dem Drücken und gedrückt halten einer Taste erhalten bleibt und langsam leiser wird. Die Zeit, bis ein Ton nach dem Loslassen ausgeklungen ist, beschreibt der Begriff *Release*.

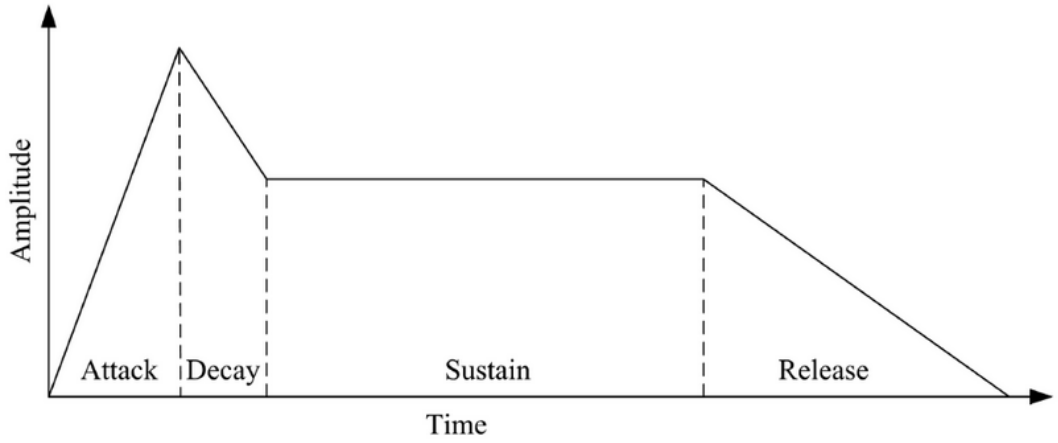


Abbildung 3.17: Darstellung von *Attack*, *Decay*, *Sustain* und *Release* [44]

Für die Anwendung in der Merkmalsvorverarbeitung ist diese Aufteilung sehr komplex. Eine Vereinfachung stellt das *Attack-Onset-Release* Modell dar [59, p. 40], in dem nur der Beginn, der Zeitpunkt mit höchster Amplitude, sowie das Ende eines Tons bestimmt werden. Diese Zeitpunkte lassen sich mit den in Abschnitt 3.1.4 vorgestellten Methoden ermitteln. Um Merkmale, dessen Zeitfenster einen solchen Noteneinsatz beinhalten auszuschließen, kann die so genannte Zwischen-Onset Methode angewendet werden. Hier wird nur der Mittelwert über solche Merkmale berechnet, die in einem bestimmten Bereich zwischen zwei aufeinander folgenden Onsets liegen.

Kapitel 4

Grundlagen

Im Folgenden sollen Methoden erläutert werden, welche für die Studien dieser Arbeit Anwendung fanden. Darunter die Lineare Regression (Abschnitt 4.1), die es erlaubt, einen linearen Zusammenhang mehrerer numerischer Merkmale auf den Arousal- oder Valence-Wert herzustellen. Das Regressionsmodell wird auf einer Menge Trainingsdaten angelernt und soll darauf hin die in der Testmenge befindlichen Musikstücke bezüglich ihrer Emotionen vorhersagen. Abschnitt 4.2 erklärt dazu die Kreuzvalidierung, welche die Aussagekraft erhöht, indem mehrere Test- und Trainingsmengen bestimmt und für die Regression verwendet werden. Eine Methode zur Auswahl von Merkmalen stellt MRMR dar (Abschnitt 4.3). Mit ihr wird Relevanz und Redundanz der zur Auswahl stehenden Features in ein Verhältnis gesetzt, was es erlaubt, eine approximativ gute Zusammenstellung zu finden.

4.1 Multiple Lineare Regression

Die Lineare Regression [4, 66] ist ein statistisches Werkzeug, um den linearen Zusammenhang einer Variable zu einer oder mehreren unabhängigen Variablen zu modellieren. In den folgenden Gleichungen ist y die zu bestimmende abhängige Variable, im Anwendungsfall der Emotionsvorhersage entspricht diese dem Arousal- oder Valence-Wert. Unabhängige Variablen (Regressanden) $x_1 ... x_k$ entsprechen den k verwendeten Merkmalen. Der lineare Zusammenhang wird über die Regressionskoeffizienten $\beta_0 ... \beta_k$ wie in Gleichung 4.1 ausgedrückt. Im 2D-Fall ($k = 1$) der *einfachen linearen Regression* wird hiermit eine Gerade approximiert. Bei mehr als einer unabhängigen Variable ist von der *multiplen linearen Regression* die Rede. Der jeweilige Approximationsfehler ϵ soll dabei möglichst gering sein. Um hohe Abweichungen stärker ins Gewicht fallen zu lassen wird die Summe der quadrierten Fehler $||y - X\beta||^2$ minimiert.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \epsilon_i \tag{4.1}$$

$$y = X\beta + \epsilon \quad , X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & ... & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & ... & x_{k,2} \\ 1 & ... & ... & ... & ... \\ 1 & x_{1,n} & x_{2,n} & ... & x_{k,n} \end{bmatrix} \tag{4.2}$$

Die Regressionskoeffizienten β_i können auf n gegebenen Trainingsdaten y und X berechnet werden. Gleichung 4.2 zeigt die in Matrixschreibweise überführte Gleichung 4.1. Indem zuerst mit der transponierten von X X^T (Gl. 4.3) und anschließend mit der inversen von $X^T X$ erweitert wird kann β durch Lösen des Linearen Gleichungssystems berechnet werden.

$$X^T X\beta = X^T y \tag{4.3}$$

$$\beta = (X^T X)^{-1} X^T y \tag{4.4}$$

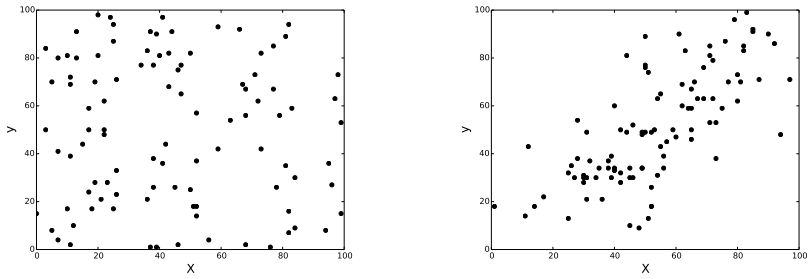
$$\hat{y} = X\beta \tag{4.5}$$

Nachdem die Modelle für Arousal und Valence trainiert wurden, können durch Einsetzten der extrahierten Merkmale x_i in Gleichung 4.5 unter Vernachlässigung des Fehlers ϵ aus Gleichung 4.2 Emotionsvorhersagen gemacht werden.

Eine Aussage darüber, wie gut dieses lineare Modell den Zusammenhang von X und y darstellt gibt das Bestimmtheitsmaß R^2 [66]. Im Fall der einfachen linearen Regression entspricht es dem quadrierten Korrelationskoeffizient nach Bravais und Pearson. Allgemein wird R^2 durch das Verhältnis der quadrierten Abweichungen von Regressions- und y-Werten beschrieben (Gl. 4.6). Der Vektor $\hat{y} = (\hat{y}_0, \hat{y}_1, ...\hat{y}_N)$ aus Gleichung 4.5 enthält die vorhergesagten Werte. $\bar{y} = \frac{1}{N}\sum_{i=0}^{N} y_i$ ist das arithmetische Mittel über alle y_i.

$$R^2 = \frac{\sum_{i=0}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=0}^{N}(y_i - \bar{y})^2} \tag{4.6}$$

Abbildung 4.1 zeigt mögliche Verteilungen von X und y der vier angegebenen Werte für R^2. Ein Bestimmtheitsmaß von 0 zeigt eine Unabhängigkeit von X zu y, wohingegen $R^2 = 1$ einen maximal linearen Zusammenhang der beiden Variablen wiederspiegelt. Diese Korrelation muss nicht wie im Beispiel gezeigt positiv sein. Bei der Anwendung von linearer Regression für die Vorhersage von Emotionen in Musik sind Bestimmtheitsmaße im Bereich 0.5 für Arousal und 0.1 für Valence zu erwarten [53]. Im Rahmen dieser Arbeit soll R^2 als Hauptindikator für die Güte der Regression herangezogen werden.



(a) R^2=0 (b) R^2=0.5

(c) R^2=0.8 (d) R^2=1

Abbildung 4.1: Beispiele für verschiedene Bestimmtheitsmaße

4.2 Kreuzvalidierung

Die „1000 Songs Database“ (Kapitel 2.2) enthält 744 Songs mit Arousal und Valence Werten. Um die Vorhersage dieser zu Testen muss eine Menge von Songs für das Training des Modells bestimmt werden. Die Musiktitel, die nicht für das Training verwendet wurden, bilden die Testmenge. Dazu werden zuerst alle Datenpaare in N_{cv} Partitionen P_i eingeteilt, auch *Folds* genannt. $N_{cv} - 1$ Partitionen bilden jeweils die Trainingsmenge. Zu jeder der N_{cv} verschiedene Aufteilungen werden im Rahmen der Tests statistische Werte berechnet, darunter auch R^2. Das Mitteln der Ergebnisse führt anschließend zu den Endresultaten der Kreuzvalidierung.

4.3 MRMR

Durch die Vielzahl von cepstralen und nicht-cepstralen Merkmalen (Kapitel 3), die auf Musiksignalen generiert werden können, ließe sich ein Featurevektor mit mehreren Dutzend Dimensionen zusammenstellen. Wie jedoch die Evaluierung der MFCC's in Kapitel 5.2 zeigt, steigt zwar das Bestimmtheitsmaß R^2 der linearen Regression auf den Trainingsdaten der Kreuzvalidierung, sinkt allerdings auf Testdaten ab einer gewissen Dimensionsanzahl. Dieses Verhalten ist darauf zurückzuführen, dass der für die entsprechende Regression benötigte Informationsgehalt der Merkmale ab diesem Punkt nicht mehr steigt,

sondern zunehmend redundante Daten enthält, wodurch das Modell überbestimmt wird oder anders gesagt, das Rauschen zunimmt. Eine empirische Bestimmung der optimalen Anzahl Koeffizienten für Merkmale wie MFCC oder CMRARE (wie in Kapitel 5.2 und 5.4 angewandt) ist hierfür, bezogen auf ihre Laufzeit, realistisch. Anders ist es bei der Menge nicht-cepstraler Merkmale. Es soll für jede der vier Untergruppen eine Menge von K Merkmalen gefunden werden, die eine hohe Relevanz und gleichzeitig eine verhältnismäßig geringe Redundanz aufzeigen. Die Gesamtheit möglicher Kombinationen aus $N_{features}$-Merkmalen beziffert sich auf $(N_{features}+1)^2-1$. Mit einer Begrenzung auf K ausgewählte Merkmale liegt sie weiterhin mit $\sum_{i=0}^{K}(N_{features}-i)$ zu hoch für den empirischen Ansatz. Aufgrund der Komplexität des Zusammenhangs der unterschiedlichen Merkmale auf das Regressionsproblem ist ein effizientes Vorgehen zur Bestimmung der besten Auswahl hier nicht möglich.

Minimum Redundancy — Maximum Relevance (MRMR) [46, 11] ist ein Verfahren, welches Relevanz und Redundanz in Verhältnis stellt, um so die Wichtigkeit eines Merkmals zu beurteilen. Die Transinformation (Mutual Information) $I(x,y)$ wird durch Gleichung 4.7 ausgedrückt und ist ein Maß für die Ähnlichkeit der Variablen x und y.

$$I(x,y) = \sum_i \sum_j p(x_i,y_j) log \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \quad [11] \tag{4.7}$$

Die Redundanz einer Menge Merkmale S kann als normalisierte Summe der Transinformationen aller Kombinationen mit zwei Merkmalen i und j beschrieben werden (Gl. 4.8). Ziel ist es, eine Teilmenge S aus der Menge aller verfügbaren Merkmale Ω zu finden, für die W_I minimal wird.

$$W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j) \tag{4.8}$$

Die gewählten Merkmale dürfen nicht nur möglichst verschieden sein, sie sollen gleichzeitig eine hohe Aussagekraft für die Klassifikationsvariable h aufweisen. Im Fall der Emotionsvorhersage entspricht h entweder Arousal oder Valence. V_I in Gleichung 4.9 ist dazu die normalisierte Summe aller Ähnlichkeiten der Merkmale in S zu h und soll daher maximal werden.

$$V_I = \frac{1}{|S|} \sum_{i \in S} I(h,i) \tag{4.9}$$

Um die Bedingungen $min\ W_I$ und $max\ V_I$ Zeitgleich zu optimieren, werden in [11] zwei Kriterien angegeben. Das *Mutual Information Difference criterion* (MID) (Gl. 4.10) maximiert die Differenz zwischen V_I und W_I, das *Mutual Information Quotient criterion* (MIQ) (Gl. 4.11) maximiert den Quotienten beider Bedingungen.

$$max(V_I - W_I) \tag{4.10}$$

$$max(\frac{V_I}{W_I}) \tag{4.11}$$

Der Suchaufwand kann deutlich verringert werden, indem eine näherungsweise optimale Lösung durch schrittweises hinzunehmen von Merkmalen gesucht wird. Sei nun S die Menge aller bisher gewählten Merkmale aus Ω, so enthält sie im ersten Schritt das Merkmal $i \in \Omega$, für welches $I(h,i)$ maximal ist. Sukzessiv werden nach dem MID $max(V_{SI}-W_{SI})$ oder MIQ Kriterium $max(V_{SI}/W_{SI})$ weitere nicht bereits in S enthaltene Merkmale aus $\Omega_S = \Omega - S$ hinzugenommen. Gleichungen 4.12 und 4.13 entsprechen hierbei Gl. 4.8 und 4.9 für jeweils ein auszuwählendes Merkmal i.

$$W_{SI} = \frac{1}{|S|} \sum_{j \in S} I(i,j), i \in \Omega_S \tag{4.12}$$

$$V_{SI} = I(h,i), i \in \Omega_S \tag{4.13}$$

Das schrittweise hinzunehmen einzelner Merkmale führt zu einer effizienten Vorgehensweise und liefert zudem eine Rangfolge der Merkmale, in der ein Hinzunehmen zur Regression sinnvoll ist.

Kapitel 5

Studien

Die Implementierung der Tests fand in Python statt. Verwendete Merkmale mussten vorerst extrahiert werden, um sie erst während der Tests, durch z.B. Mittelwertberechnung über das für die Vorhersage verwendete Zeitfenster, vorzuverarbeiten. Vorhersagen fanden bei einem Großteil der Tests für jeweils ganze 45 Sekunden lange Musikclips der „1000 Songs Database“ statt. Dazu wurden die der Datenbank beiliegenden statischen Annotationen verwendet, welche jeweils für Arousal und Valence einen Wert zwischen -1 und 1 annehmen können. Vereinzelt konnten nicht-cepstrale Merkmale auf bestimmten Songs nicht fehlerfrei berechnet werden. 741 Musiktitel wurden daher teilweise nur verwendet. Während dem Laden der Merkmale vor jedem Test fand der Einfachheit halber eine Ersetzung von „NaN“-Werten (Not a Number) statt. Diese zeigen hierbei an, dass der Wert, aufgrund fehlender Informationen im Musiksignal, wie z.B. zu Beginn oder am Ende eines Segments, nicht bestimmt werden konnte. Das Vorhandensein reeller Zahlenwerte ist für die Anwendung der Regression notwendig. Eine korrekte Behandlung der NaN-Werte würde ein Verständnis jedes einzelnen Merkmals und der darauf folgenden Vorverarbeitung voraussetzen. Da diese Werte jedoch als Ausnahme betrachtet werden, soll diesem Sachverhalt allerdings keine zu große Bedeutung beigemessen werden. Muss ein Wert ersetzt werden, geschieht dies durch den jeweils nächst oder zuletzt gültigen Wert innerhalb eines Musikstücks. Die Signaldynamik, sowie der Mittelwert ändern sich dadurch nur geringfügig.

Eine achtfache Kreuzvalidierung soll zur Gewinnung repräsentativer Ergebnisse herangezogen werden. Aufgrund der in der „1000 Songs Database“ vertretenden Anzahl von acht Genres, wurde die Kreuzvalidierung ebenfalls so angewandt, dass ein fairer Vergleich von genrespezifischen vs. zufälligem Training gemacht werden kann. Musiktitel wurden hierzu zuvor zufällig gemischt. Die Einteilung der Songs in acht Test- und Trainingsfolds bleibt jedoch für alle Tests zugunsten der Vergleichbarkeit erhalten. Die Evaluierungen der cepstralen Merkmale MFCC und OBSC in Abschnitt 5.2 und 5.3 fand mit nicht zufälligen Einteilungen der Songs statt. Aufgrund der Vorsortierung der „1000 Songs Database“ nach Genres enthalten Testfolds dabei maximal Songs aus zwei Genres. Hierbei gewonne-

ne Extraktionsparameter können dennoch für den weiteren Verlauf verwendet werden, da sich durch die Korrektur auf zufällige Einteilungen der Folds eine Verbesserung über alle Vorhersagen zeigte. Aufgrund der zeitintensiven Extraktion musste auf eine Wiederholung der Tests verzichtet werden. Die Auswahl nicht-cepstraler Merkmale, die Evaluierung von CMRARE, sowie die abschließende Betrachtung der Relevanz beziehen sich jedoch auf Ergebnisse, die durch zufällige Einteilungen in Test- und Trainingsmenge entstanden sind. Als Gütemaß dient in erster Linie das Bestimmtheitsmaß R^2, welches wie in Abschnitt 4.1 beschrieben auf den jeweils verwendeten Testdaten berechnet und über die acht Ergebnisse gemittelt wird. Die Bestimmung bezüglich Arousal und Valence sind als zwei unabhängige Vorhersagen zu sehen. Evaluierungen werden daher separat für beide Variablen durchgeführt.

5.1 Testablauf

Zuerst sollen die jeweiligen Parameter der drei cepstralen Merkmale MFCC, OBSC und CMRARE unabhängig voneinander optimiert werden. Des Weiteren wird eine Auswahl der nach MRMR (Kapitel 4.3) besten drei nicht-cepstralen Merkmale jeder Gruppe stattfinden. Die Aussagekraft einzelner und mehrerer Gruppen für die Vorhersage von Emotionen soll mit den 16 möglichen Kombinationen festgestellt werden. Anschließend werden die verschiedenen cepstralen Merkmale hinzugenommen und somit geprüft, ob diese eine relevante Verbesserung auf die unterschiedlichen Zusammensetzungen erzielen.

5.2 MFCC-Evaluierung

Zur Extraktion von *Mel-Frequency Cepstral Coefficients* (Abschnitt 3.2.1) stehen in librosa, einer Python-Bibliothek zur Musikanalyse, mehrere Parameter zur Verfügung. Viele davon sind untereinander abhängig, sodass eine naive Optimierung einzelner Parameter nur lokale Optima liefert. Aufgrund der Komplexität und Anzahl der möglichen Einstellungen ist das Testen aller Parameter zeitintensiv. Daher sollen für die MFCC Evaluierung bestimmte Werte festgelegt werden. Darunter fällt zum einen der betrachtete Frequenzbereich. Das Gehör eines erwachsenen Menschen ist fähig, Töne mit Frequenzen von wenigen Herz bis zu 10 oder 12 kHz wahrzunehmen [48, p. 80]. Es ist somit ausreichend, Frequenzen von 0 bis 16000 Hz für die Audioanalyse zu betrachten. Berechnet werden die MFCC's auf fortschreitenden Zeitfenstern. Die Länge der Zeitfenster ist ein wichtiger Parameter, der in fast allen Merkmalen Verwendung findet. Muss ein Audiosignal im Frequenzbereich analysiert werden, wird es durch Fourier-Transformation in diesen überführt. Die Anzahl der Datenpunkte (Samples) entspricht hierbei typischerweise einer Zweierpotenz, da die Berechnung effizienter erfolgen kann. Während die Fensterlänge evaluiert werden soll, kann der prozentuale Anteil, um den sich aufeinander folgende Fenster überlappen, auf 50% festgelegt

werden. Dies stellt sicher, dass Ereignisse im Signal nicht durch Fenstergrenzen getrennt werden. Zu den zu optimierenden Parametern gehört neben der Fenstergröße (*FFT window*) die Anzahl der *Mel bins*, sowie die Menge der durch die Kosinus-Transformation berechneten Koeffizienten. In [30] wird eine Fensterlänge von 25.6 ms vorgeschlagen. Das entspricht in etwa 1024 Samples bei 44100 Hz Abtastrate, wie sie hier für die MFCC Extraktion verendet wird. 23.2 ms (1024 Samples) sollen für die Evaluierung daher als Startwert dienen. Während der Berechnung des Merkmals werden die nach Mel skalierten Frequenzen in Frequenzbänder eingeteilt, dessen Anzahl der Wert *Mel bins* beschreibt. In [30] findet ein Wert von 40 Anwendung. Auch wird dort eine Menge von 13 MFCC Koeffizienten angegeben. Mit zunehmender Anzahl enthalten diese aufgrund der de-Korrelation immer weniger relevante Informationen. Ab einer bestimmten Anzahl ist ein Sinken des Bestimmtheitsmaßes auf den Testdaten daher anzunehmen. Um einen Überblick zu bekommen werden alle Kombinationen aus den in Tabelle 5.1 angegebenen Parametern getestet.

Parameter	**Werte**
Fenster	256, 512, 1024, 2048, 4096, 8192
Mel bins	32, 64, 128, 256, 512, 1024
Koeffizienten	5, 10, 20

Tabelle 5.1: Startwerte für die MFCC Evaluierung

Die durchschnittlichen Bestimmtheitsmaße der Testdaten über die acht Folds sind in Tabellen 7.1 bis 7.6 im Anhang aufgelistet. Das beste Ergebnis von $R^2 = 0.4964$ für die Vorhersage von Arousal wurde mit einem Fenster von 1024 Samples (23.2 ms), 1024 Mel bins und fünf Koeffizienten erreicht. Ebenfalls war eine sinkende Tendenz mit steigenden Koeffizienten ab fünf für Valence zu erkennen. Dort lag R^2 mit 0.0944 für 8192 Samples (185.8 ms) und 32 Mel bins am höchsten. Weitere Tests für MFCC Koeffizienten unter zehn waren daher notwendig. Mit diesem Überblick über punktuelle Testergebnisse ist eine Optimierung einzelner Variablen von diesem Punkt an als gerechtfertigt anzusehen. Tabellen 7.7 (für Arousal) und 7.8 (für Valence) zeigen die Ergebnisse für Koeffizienten von zwei bis neun, wobei die zuvor ermittelte Fensterlänge jeweils beibehalten wurde. Auch die Anzahl der Mel bins wurde übernommen und um 512 und 2048 bzw 16 und 64 erweitert, um Abhängigkeiten und die damit sich ändernden Optima mit zu betrachten. Das Optimum für Valence blieb innerhalb der Ergebnisse dieses Tests bei fünf Koeffizienten und 32 Mel bins. Ein minimal höheres Bestimmtheitsmaß von 0.4967 konnte mit einer Verringerung auf vier MFCC's erreicht werden. Da eine neue beste Anzahl an Koeffizienten für Arousal gefunden wurde, musste der erste Test mit vier MFCC's in einem kleineren Bereich wiederholt werden, um dort das lokale Optimum sicher zu stellen. In Tabellen 7.5 und 7.6 liegt der höchste Wert für R^2 an der oberen Grenze der Fensterlänge. Diese wurde daher für die Vorhersage von Valence mit 2048 (46.4 ms) bis 65535 Samples (1486.1 ms)

und anschließender Variation der Anzahl Mel bins getestet. So ergab sich bei Valence $R^2 = 0.1065$ für 32768 Samples (743 ms) Fensterlänge und 64 Mel bins (Tabelle 7.10). Darauf folgende Tests über die Anzahl der Koeffizienten ergab keine Änderung des Optimums. Das Bestimmtheitsmaß nahm für Tests auf Arousal für 512 Samples (11.6 ms) und 2048 Mel bins zu (Tabelle 7.9). Auch hier bleibt das Optimum für vier MFC-Koeffizienten bestehen. Tabelle 5.2 zeigt die endgültigen Ergebnisse der MFCC-Evaluierung. Im Gegensatz zu den für vorangehende Vergleiche verwendeten Werten, sind Ergebnisse dieser Tabelle durch zufällige Test- und Trainigsmengen entstanden. Die Varianz der Fehler liegt für Valence mit 0.0189 etwas höher als 0.0122 bei Arousal. In den Studien zur Relevanz der cepstralen Merkmale (Abschnitt 5.6) zeigten sich MFCC's als nicht besonders aussagekräftig für die Vorhersage der Valence. Auch die Hinzunahme zu anderen Merkmalen liefert keine zu OBSC oder CMRARE überlegenen Resultate.

Test	**Frequenz-bereich**	**Fenster**	**Fenster-überlapp**	**Mel bins**	**Koef-fizienten**	**Test-R^2**	**Fehler**
Arousal	0-16 kHz	512	50%	2048	4	0.5395	0.1463
Valence	0-16 kHz	32768	50%	64	5	0.1762	0.1793

Tabelle 5.2: Gefundene Einstellungen für MFCC mit höchstem Bestimmtheitsmaß. R^2 und Fehler sind korrigierte Ergebnisse bei zufällig eingeteilten Mengen für Test- und Training

5.3 OBSC-Evaluierung

Zur Evaluierung des *Octave-Based Spectral Contrast* Merkmals sind durch die Implementierung in librosa neben FFT Fensterlänge auch die Grenzen der sechs Frequenzbänder, sowie der α-Wert einstellbar. Da die Wahl des Quartils (α-Wert) hat zur Glättung der Maxima und Minima bei der Berechnung von *Peak* und *Valley* laut Jiang et al. [21] keinen großen Einfluss auf den endgültigen Merkmalsvektor und wird daher hier auf den dort angegebenen Wert von $\alpha = 0.02$ festgelegt. Frequenzbänder können frei angegeben werden, jedoch ist das OBSC-Feature für sechs Frequenzbänder mit einem Abstand von jeweils einer Oktave definiert. Die Angabe der ersten Frequenzgrenze F_{min} ist somit ausreichend, da ein Abstand einer Oktave eine Verdoppelung der Frequenz bedeutet. Hier werden zunächst FFT-Zeitfenster in einem Bereich von 256 bis 65535 Samples, sowie der Parameter F_{min} von 5 bis 400 Hz kombiniert (Tabellen 7.11 und 7.12). Der Überlapp der Zeitfenster ist fest auf 50% eingestellt. In beiden Fällen sind Zeitfenster von 16384 Samples (371.5 ms) und 32768 Samples (743 ms) hinsichtlich R^2 zu bevorzugen. In [3] wird bei der Vorverarbeitung dem Merkmal der Wert der Varianz hinzugefügt. Dadurch sollen Informationen der Werteverteilung, die bei der alleinigen Verwendung des Mittelwertes verloren gehen, behalten

werden. Tabellen 7.13 und 7.14 zeigen eine genauere Evaluierung über die erste Frequenzgrenze F_{min}, wobei nur der Mittelwert der Merkmale benutzt wurde. Die Hinzunahme der Varianz brachte nur für die Vorhersage von Arousal eine Verbesserung. Hier stiegt R^2 von 0.4549 auf 0.4703. F_{min} liefert mit 30 Hz hier das beste Ergebnis. Bei der Evaluierung für Valence konnte mit $F_{min} = 10Hz$ und 32768 Samples für ein Fenster der FFT ein Wert von 0.1264 erreicht werden. Die Einstellungen für die Ermittlung von Arousal und Valence durch OBSC sind in Tabelle 5.3 aufgelistet und werden für die folgenden Tests herangezogen.

Test	**F_{min}**	**Fenster**	**Fenster-überlapp**	α	**Vorverarbeitung**	**Test-R^2**	**Fehler**
Arousal	30 Hz	16384	50%	0.02	Mean & Varianz	0.5334	0.1507
Valence	10 Hz	32768	50%	0.02	Mean	0.181	0.1787

Tabelle 5.3: Gefundene Einstellungen für OBSC mit höchstem Bestimmtheitsmaß R^2 und Fehler sind die Ergebnisse bei zufällig eingeteilten Mengen für Test- und Training

5.4 CMRARE-Evaluierung

Cepstral Modulation Ratio Regression wurde für die Evaluierung bezüglich seines Fensters für die Fouriertransformation und dem Grad des Polynoms der Regression betrachtet. Der Polynomgrad bestimmt zudem auch die Dimension des Merkmalsvektors. Wie in Abschnitt 3.3 beschrieben, wurde CMRARE mit dem AMUSE-Framework extrahiert. Die Abtastfrequenz der Audiosignale war daher auf 22050 Hz festgelegt. Bei diesem Merkmal sind im Vergleich zu MFCC und OBSC die Zeitfenster mit mehreren Sekunden deutlich länger. 110250 und 220500 Samples (5 s und 10 s) ohne gegenseitige Überlappung wurden hier für die Evaluierung gewählt. In [32] wurde ein Polynomgrad von 3 für die Klassifizierung von Sprache, Musik und Geräuschen verwendet. Für die Betrachtung der Abhängigkeit des Grades wurde CMRARE für Grad 5 und 10 mit 5 s und 10 s Fensterlänge extrahiert. Tabelle 7.15 zeigt die Ergebnisse bezüglich R^2 der Vorhersagen von Arousal und Valence auf ganzen Musikstücken. Arousal erreichte mit $R^2 = 0.4712$ bei Polynomgrad 10 und einem 10 s Zeitfenster einen geringeren Wert als MFCC oder OBSC. Mit selben Polynomgrad und Zeitfenster erreichte CMRARE bei Valence eine Bestimmtheit von 0.2897 und ist damit deutlich besser als die anderen cepstralen Merkmale MFCC und OBSC. Tabelle 5.4 zeigt beide im späteren Verlauf verwendeten Einstellungen für CMRARE.

Mit angegeben ist in der Tabelle der minimale R^2-Wert der achtfachen Kreuzvalidierung. Dieser liegt hier bei Valence mit etwa 0.11 im Größenbereich von MFCC und OBSC.

Test	Fenster	Polynom-grad	Test-R^2	Minimaler Test-R^2	Fehler	Fehler Varianz
Arousal	220500	10	0.4712	0.2433	0.1584	0.014
Valence	220500	10	0.2897	0.1117	0.1681	0.0147

Tabelle 5.4: Gefundene Einstellungen für CMRARE mit höchstem Bestimmtheitsmaß

Aufgrund dieser Ergebnisse lässt sich zusammenfassen, dass CMRARE für die Vorhersage der Valence, im Rahmen der getesteten Parameter, als bevorzugendes Merkmal anzusehen ist.

5.5 Auswahl nicht-cepstraler Merkmale

AMUSE bietet mehrere Merkmale, eingeteilt in die fünf Gruppen *Energy*, *Timbre*, *Harmony and Melody*, *Tempo and Rhythm* und *Cepstral* an. Da cepstrale Merkmale in den hier gemachten Studien zunächst separat betrachtet werden, spielen für die Auswahl nicht-cepstraler Merkmale die ersten vier Gruppen mit insgesamt 43 verschiedenen Features eine Rolle. Um hier den Einfluss der Fenstergröße mit zu untersuchen, wurden 24 der Merkmale zusätzlich zu zwei anderen Zeitfenstern hinzugenommen. Tabellen 3.3, 3.4, 3.5 und 3.6 geben eine Übersicht über alle verwendeten Merkmale mit den extrahierten Zeitfenstern. Wie in Abschnitt 4.3 bereits angemerkt, ist die Erstellung eines großen Merkmalsvektors durch Zusammenfügen aller zur Verfügung stehenden Merkmale nicht zielführend. Werden diese unabhängig voneinander für die Regression angewandt, können sie gute Ergebnisse liefern, die Kombination der Besten Features aus dieser getrennten Betrachtung kann jedoch bezüglich des Bestimmheitsmaßes auf den Testdaten bei der Kreuzvalidierung deutlich schlechter ausfallen. Schuld ist die Redundanz zwischen den Merkmalen. *Minimum Redundancy — Maximum Relevance* (Abschnitt 4.3) wählt zunächst das Merkmal mit der höchsten Relevanz bezüglich der gewählten Klassifikationsvariable. Anschließend werden sukzessiv weitere Merkmale, nach einem Kriterium, welches Relevanz zur Redundanz in Beziehung stellt, hinzu genommen. Dieses Vorgehen liefert eine approximativ gute Zusammenstellung von Merkmalen, indem die ersten N Einträge der entstandenen Rangfolge übernommen werden.

Mit der Implementierung von MRMR als Online-Tool[1] fand die Erzeugung der Rangfolge nach dem MID-Kriterium statt, welches die Differenz von Relevanz und Redundanz verwendet. Vorangehende Untersuchungen wurden bezüglich der Wahl von MID oder MIQ als angewandtes Kriterium gemacht, da keine eindeutige Empfehlung diesbezüglich gefunden wurde. Die Ergebnisse bei diesem Vergleich waren sehr ähnlich. In den ersten Plätzen waren nur einzelne Vertauschungen zu erkennen, erst ab der Hälfte der Rangfolge unter-

[1] mRMR Online-Tool: http://penglab.janelia.org/proj/mRMR/, aufgerufen am 15.3.2016

schieden sich die Auflistungen deutlicher. Ausgewählt werden im Folgenden nur die besten drei Merkmale jeder Gruppe, daher ist die Wahl des Kriteriums nicht als kritisch für das Ergebnis anzusehen. Die Generierung der für das Tool benötigten Tabellen fand in Python statt. Hier wurden die Merkmale vorverarbeitet und anschließend jeweils mit der Klassifikationsvariable Arousal und Valence als CSV exportiert. Merkmale wurden bei der Vorverarbeitung durch Mittelwertberechnung zu einem Vektor zusammengefasst. Da das MRMR-Tool auf diskreten Werten arbeitet, mussten diese zuvor auf zwei Nachkommastellen gerundet und mit 100 multipliziert werden, um eine ausreichende Genauigkeit zu erreichen. Diese Tabelle enthält, bis auf die Diskretisierung, die Daten, welche anschließend für die Regression benutzt wurden. Dies bedeutet auch, dass jede Dimension mehrdimensionaler Merkmale als eigenes Feature betrachtet wird. Ist im Folgenden eine bestimmte Dimension eines Merkmals gemeint, wird sie mit „#“, gefolgt von einer Zahl, beginnend mit 1 für die erste Dimension, angegeben. Tabellen 7.16 und 7.17 im Anhang zeigen die Ergebnisse für Arousal, 7.16 und 7.17 für Valence. Dort wurde für die Vorhersage immer das nächste Merkmal der Liste hinzugenommen. Die Fenstergröße ist dort der Merkmalsbezeichnung angefügt, andernfalls ist ein Fenster von 1024 Samples verwendet worden. Die Samplerate betrug bei allen nicht-cepstralen Merkmalen 22050 Hz. Viele der Gruppen erreichen schon mit drei Features 90% ihres Maximums. Tabellen 5.5 und 5.6 zeigen die als Repräsentanten der vier Gruppen gewählten Merkmale.

	Energy Merkmal	**Timbre Merkmal**
1	RMS peak number in 3 seconds	Spectral flatness measure 2048 #1
2	Zero-crossing rate 2048	Distances in phase domain 1024
3	Root mean square 512	Spectral brightness 1024

	Harmony and Melody Merkmal	**Tempo and Rhythm Merkmal**
1	Harmonic change detection function 2048	Estimated onset number per minute
2	Tristimulus 512 #1	Characteristics of fluctuation patterns #3
3	Inharmonicity 512	Rhythmic clarity

Tabelle 5.5: MRMR Rangfolge der besten 3 Merkmale (Arousal)

Die „RMS peak number in 3 seconds“ ist als Merkmal der Energie sowohl für Arousal, als auch für die Vorhersage des Valence-Wertes hoch gewertet. Für die Gruppe „Harmony and Melody“ erwies sich die „Harmonic change detection function“ mit einem 92.9 ms (2048 Samples) Zeitfenster als relevantes Merkmal für beide Dimensionen des Emotionsmodells. Mit einem R^2 von 0.462 erreicht die „Timbre“-Merkmalsauswahl für Arousal den höchsten Wert der vier nicht-cepstralen Gruppen. Obwohl eine Abhängigkeit der Harmonie, Melodie

	Energy Merkmal	**Timbre Merkmal**
1	RMS peak number in 3 seconds	Spectral bandwidth 1024
2	Low energy 512	Spectral irregularity 512
3	Zero-crossing rate 1024	Spectral crest factor 2048 #1

	Harmony and Melody Merkmal	**Tempo and Rhythm Merkmal**
1	Harmonic change detection function 2048	Rhythmic clarity
2	Strengths of minor keys 1024 #8	Estimated onset number per minute
3	Majorminor alignment 4096	Characteristics of fluctuation patterns #3

Tabelle 5.6: MRMR Rangfolge der besten 3 Merkmale (Valence)

oder Klangfarbe (Timbre) zur Wertigkeit der Emotion zu Erwarten wäre, liegen die drei Merkmale der „Tempo and Rhythm“ Zuteilung mit $R^2 = 0.1616$ deutlich vorne. Für folgende Studien wird sich bei den Merkmalen der hier genannten nicht-cepstralen Gruppen auf die in Tabellen 5.5 und 5.6 aufgelisteten Features bezogen.

5.6 Relevanz der cepstralen Merkmale

Die Relevanz der cepstralen Merkmale MFCC, OBSC und CMRARE soll auf den 15 entstandenen Gruppenkombinationen getestet werden. Dazu wird jede der sieben Kombinationen aus MFCC, OBSC und CMRARE den nicht-cepstralen Gruppenkombinationen hinzugefügt. Die Ergebnisse sind in Tabelle 5.7 für Arousal und Tabelle 5.8 für Valence bezüglich durchschnittlichem R^2 der Testfolds zusammengefasst.

Im linken Bereich sind jeweils die für die Zeile verwendeten Merkmale der Gruppen mit einem **X** gekennzeichnet. Der rechte Teil zeigt die Bestimmtheitsmaße der jeweiligen Kombination, wobei jede Spalte einer Kombination der cepstralen Merkmale MFCC, OBSC und CMRARE, die den entsprechenden nicht-cepstralen Gruppenkombinationen hinzugefügt werden, entspricht. Die zur Extraktion verwendeten Parameter wurden in den Abschnitten 5.2, 5.3 und 5.4 ermittelt. Jede nicht-cepstrale Gruppe enthält drei Merkmale, wobei für mehrdimensionale Merkmale einzelne Dimensionen gemeint sind. Der Merkmalsvektor jeder Gruppe besteht somit aus genau drei Einträgen.

Das höchste Bestimmtheitsmaß ist in jeder Spalte hervorgehoben. Das beste Ergebnis von Arousal und Valence ist zudem rot eingefärbt. Mit den Gruppen *Energy*, *Harmony and Melody*, *Tempo and Rhythm*, sowie allen drei cepstralen Merkmalen konnte der höchste Wert bei Arousal von $R^2 = 0.6685$ erreicht werden (Tabelle 5.7). Jedoch liegt das Maß mit 0.5932 nur minimal unter dem genannten Maximum, welches mit weniger als der Hälfte

Energy	Timbre	Harmony and Melody	Tempo and Rhythm	Nur nicht-cepstral	Mit MFCC	Mit OBSC	Mit CMRARE	Mit MFCC & OBSC	Mit MFCC & CMRARE	Mit OBSC & CMRARE	MFCC, OBSC, CMRARE
					0.5395	0.5334	0.4712	0.6150	0.6197	0.5918	0.6500
X				0.4751	0.5780	0.5823	0.5946	0.6230	0.6331	0.6387	0.6565
	X			0.5054	0.5710	0.5871	0.6038	0.6271	0.6391	0.6306	0.6544
		X		0.4872	0.6143	0.5703	0.5674	0.6356	0.6429	0.6125	0.6562
			X	0.2968	0.5800	0.5487	0.5311	0.6235	0.6370	0.6141	0.6591
X	X			0.5525	0.6901	0.598	0.6248	0.6259	0.6403	0.6390	0.6508
X		X		0.5803	0.6248	0.6106	0.6330	0.6417	0.6516	0.6504	0.6623
X			X	0.4918	0.5965	0.5867	0.6000	0.6314	0.6402	0.6432	0.6611
	X	X		0.5674	0.6180	0.6088	0.6195	0.6447	0.6470	0.6395	0.6594
	X		X	0.5500	0.5967	0.5931	0.6242	0.6286	0.6491	0.6407	0.6583
		X	X	0.5369	0.6345	0.5858	0.5988	0.6453	0.6576	0.6310	0.6657
X	X	X		0.5916	0.6212	0.6210	0.6394	0.6417	0.6480	0.6510	0.6567
X	X		X	0.5674	0.6095	0.6014	0.6304	0.6313	0.6462	0.6426	0.6546
X		X	X	0.5951	**0.6407**	0.6185	0.6399	**0.6513**	**0.6598**	0.6562	**0.6685**
	X	X	X	0.5915	0.6350	0.6163	0.6357	0.6494	0.6574	0.6495	0.6650
X	X	X	X	**0.6070**	0.6370	**0.6283**	**0.6468**	0.6404	0.6552	**0.6566**	0.6624

Tabelle 5.7: Bestimmtheitsmaße der Kombinationen aus nicht-cepstralen Gruppen mit den drei cepstralen Merkmalen (Arousal)

Dimensionen des Merkmalsvektors gewonnen werden konnte. Hierfür waren als cepstrales Merkmal nur die MFCC's notwendig, was zusammen einen 13-Dimensionalen Merkmalsvektor ergibt (gegenüber 30 für das beste Ergebnis). In Hinblick auf die Ergebnisse einzelner nicht-cepstraler Gruppen liefert *Tempo and Rhythm* das mit Abstand schlechteste Ergebnis von 0.2968. Merkmale dieser Gruppe sollten demnach nur in Kombination mit anderen Merkmalen zur Vorhersage von Arousal verwendet werden. Für den Wert der Valence ist diese Gruppe jedoch vergleichsweise aussagekräftig (Tabelle 5.8). Auch einen deutlichen Vorteil bietet das Merkmal CMRARE für Valence. Es erreicht bei alleiniger Verwendung bereits ein Bestimmtheitsmaß von 0.2897 und liegt damit über dem Maximum des nur durch nicht-cepstrale Merkmale erreichten Wertes (Tabelle 5.8). Das CMRARE nur einen

Energy	Timbre	Harmony and Melody	Tempo and Rhythm	Nur nicht-cepstral	Mit MFCC	Mit OBSC	Mit CMRARE	Mit MFCC & OBSC	Mit MFCC & CMRARE	Mit OBSC & CMRARE	MFCC, OBSC, CMRARE
					0.1762	0.1810	0.2897	0.2156	0.3243	0.3472	0.3629
X				0.1523	0.2095	0.2084	0.3318	0.2303	0.3441	0.3652	0.3700
	X			0.1269	0.1730	0.2028	0.3168	0.2188	0.3291	0.3529	0.3623
		X		0.1096	0.2051	0.1932	0.2975	0.2249	0.3309	0.3510	0.3658
			X	0.2166	0.2780	0.2614	0.3789	0.2829	0.3899	0.3967	0.3998
X	X			0.1732	0.2076	0.2221	0.3398	0.2333	0.3509	0.3657	0.3737
X		X		0.1794	0.2256	0.2149	0.3335	0.2384	0.3476	0.3660	0.3717
X			X	0.2627	0.3060	0.2853	0.3947	0.3084	0.4004	0.4050	0.4070
	X	X		0.1680	0.2060	0.2188	0.3190	0.2343	0.3303	0.3560	0.3638
	X		X	0.2597	0.2800	0.2811	0.3948	0.2867	0.3997	0.4009	0.4055
		X	X	0.2566	0.2929	0.2737	0.3818	0.2922	0.3943	0.3972	0.4011
X	X	X		0.2018	0.2311	0.2357	0.3401	0.2469	0.3509	0.3672	0.3746
X	X		X	0.2703	0.3053	0.2913	**0.4015**	0.3080	**0.4105**	**0.4067**	**0.4136**
X		X	X	0.2780	0.3128	0.2905	0.3948	0.3117	0.4014	0.4034	0.4060
	X	X	X	0.2850	0.3002	0.2963	0.3963	0.3001	0.3998	0.4020	0.4051
X	X	X	X	**0.2881**	**0.3158**	**0.3007**	0.4006	**0.3143**	0.4087	0.4057	0.4118

Tabelle 5.8: Bestimmtheitsmaße der Kombinationen aus nicht-cepstralen Gruppen mit den drei cepstralen Merkmalen (Valence)

geringen Anteil redundanter Informationen über Valence enthält, zeigt sich an dessen Hinzunahme zu den nicht-cepstralen Gruppen. Die geringste Verbesserung beträgt hier 39%. CMRARE, zusammen mit *Energy*, *Timbre* und *Tempo and Rhythm*, liefert bereits einen, für Valence beachtlichen Wert von $R^2 = 0.4015$. Die Anzahl der Dimensionen des Merkmalsvektors beträgt in diesem Fall 19. Das beste Ergebnis von $R^2 = 0.4136$ wurde mit allen aufgelisteten Merkmalen außer *Harmony and Melody* erreicht. Hier besteht allerdings der für die lineare Regression benutzte Vektor aus 31 Werten. Die Auswahl der Merkmale für einen gegebenen Anwendungsfall sollte demnach mit Hinblick auf deren Menge gemacht werden. Obwohl eine Verbesserung durch Hinzunahme der cepstralen Merkmale für alle 15

Kombinationen, sowohl für Arousal als auch Valence, festzustellen ist, müssen Extraktions- und Berechnungsdauer unter Umständen mit betrachtet werden.

Beide besten Resultate sind in nachfolgender Tabelle 5.9 zusammengefasst. Mit angegeben ist das minimale, über die Tests der Kreuzvalidierung erreichte Bestimmtheitsmaß.

Test	**Gruppen / Merkmale**	**Test-R^2**	**Minimaler Test-R^2**	**Fehler**	**Fehler Varianz**
Arousal	Energy, Harmony and Melody, Tempo and Rhythm, MFCC, OBSC, CMRARE	0.6685	0.5442	0.1251	0.0088
Valence	Energy, Timbre, Tempo and Rhythm, MFCC, OBSC, CMRARE	0.4136	0.3077	0.1523	0.0122

Tabelle 5.9: Höchste, in dieser Arbeit erreichte Bestimmtheitsmaße für Arousal und Valence

Das dieser Wert dennoch relativ hoch ist zeigt, dass die Einteilung der Test- und Trainingsmengen das Endergebnis nicht begünstigen. Die von Soleymani et al. [53] erreichten Ergebnisse konnten hier durch gezielte Auswahl der Merkmale beachtlich übertroffen werden. Insgesamt konnte die Wichtigkeit der cepstralen Merkmale gezeigt werden. Sie enthalten daher für Emotionsvorhersagen relevante Informationen, die sich nicht mit denen der für die vier Kategorien ausgewählten nicht-cepstralen Merkmale ausreichend decken, um das Ergebnis zu verschlechtern.

Kapitel 6

Zusammenfassung

Die Qualität der inhaltsbasierten Musikanalyse ist stark abhängig von verwendeten Merkmalen, deren Aussagekraft für die Emotionsanalyse oft nicht ohne umfassende Tests angegeben werden kann. In dieser Arbeit wurden daher auf dem Raum des sogenannten Cepstrums arbeitende Merkmale mit nicht-cepstralen Merkmalen verglichen. Die Vorhersage von Emotionen fand dazu auf den Musikstücken und Annotationen der „1000 Songs Database“ mittels linearer Regression statt. Hierzu wurde das Arousal-Valence Modell verwendet, welches eine kontinuierliche Emotionsdarstellung erlaubt. Nicht-cepstrale Merkmale lassen sich weiter in vier Gruppen aufteilen. Beispielhaft wurde die Gewinnung der Merkmale aus den Bereichen *Energy*, *Timbre*, *Harmony and Melody* und *Tempo and Rhythm* näher erläutert. *Mel-Frequency Cepstral Coefficients*, das *Octave-Base Spectral Contrast Feature*, sowie das Merkmal *Cepstral Modulation Ratio Regression* basieren auf dem Cepstrum und fanden für die Beantwortung der Fragestellung nach deren Relevanz in den durchgeführten Studien Anwendung. Bevor cepstrale und nicht-cepstrale Merkmale verglichen werden konnten, mussten zunächst repräsentative Extraktionsparameter ermittelt werden. Dazu wurden Evaluationen über ausgewählte Parameter durchgeführt, um die so erhaltenen Einstellungen in darauf folgenden Tests zu verwenden. Durch das *Advanced Music Explorer* Framework (AMUSE) stehen eine Vielzahl nicht-cepstraler Merkmale zur Auswahl. Das Verfahren „*Minimum Redundancy — Maximum Relevance*“ wurde angewandt, um nach den Kriterien von Relevanz und Redundanz eine Rangfolge der Merkmale jeder Gruppe aufzustellen. Als Repräsentanten jeder der vier Gruppen dienten die jeweils besten drei Merkmale. Die Gruppe *Tempo and Rhythm* erwies sich bei den durchgeführten Studien als hilfreich für die Vorhersage der Wertigkeit (Valence, fröhlich oder traurig) der Emotion. Ebenfalls brachte CMRARE als cepstrales Merkmal für die Valence als alleiniges, sowie in Kombination mit anderen Merkmalen eine signifikante Verbesserung der Vorhersagen bezüglich der Valence. Durch Verwendung von Merkmalen der Gruppen *Energy*, *Timbre* und *Tempo and Rhythm*, sowie allen drei cepstralen Merkmalen konnte ein Bestimmtheitsmaß von etwa 0.41 erreicht werden. Dies stellt eine Verbesserung zu bisher publizierten

Ergebnissen in diesem Gebiet dar. Dass die Vorhersage der Erregung (Arousal) genauere Ergebnisse liefert, konnte bestätigt werden. Ein maximales Bestimmtheitsmaß von etwa 0.67 wurde durch Anwendung der Gruppen *Energy*, *Harmony and Melody* und *Tempo and Rhythm* in Kombination mit den drei cepstralen Merkmalen erreicht. Insgesamt konnte die Vorhersage durch Hinzunahme jedes der drei cepstralen Merkmale weiter verbessert werden. Dies zeigt, dass MFCC, OBSC und CMRARE sich in ihrem Informationsgehalt bezüglich Emotionen nicht vollständig decken und beschreiben somit verschiedene Aspekte des Cepstrums.

Es ist unwahrscheinlich, dass die optimalen Einstellungen für MFCC, OBSC und CMRARE gefunden wurden. Eine genauere Evaluation der Werte erfordert jedoch mehr Zeit, da gerade die Extraktion der Merkmale ein zeitintensives Verfahren ist. Besonders das für Valence vielversprechende Merkmal CMRARE konnte nur oberflächlich evaluiert werden. Hier sind weitere Untersuchungen hinsichtlich höherer Koeffizientenanzahlen nötig, um das volle Potential des Merkmals repräsentativ vergleichen zu können. Für weitere Studien sollte der Einfluss verschiedener Vorverarbeitungen, darunter auch die angesprochene Zwischen-Onset Methode, sowie die Verwendung statistischer Werte, wie z.B. die Varianz, bei der Zusammenfassung anderer Merkmale mit kurzem Zeitfenster betrachtet werden.

Kapitel 7

Anhang

7.1 MFCC Evaluierung

1. Test: 5 MFCC Koeffizienten (Arousal)

Fenster / Mel bins	**256**	**512**	**1024**	**2048**	**4096**	**8192**
32	0.4603	0.463	0.463	0.4638	0.4674	0.4737
64	0.4737	0.4707	0.4695	0.4706	0.474	0.4802
128	0.4827	0.4817	0.4773	0.4779	0.4824	0.4887
256	0.491	0.4899	0.4881	0.4832	0.4865	0.4923
512	0.4954	0.4955	0.4947	0.4916	0.4886	0.4941
1024	0.4937	0.496	**0.4964**	0.4957	0.4946	0.4949

Tabelle 7.1: Bestimmtheitsmaße für Arousal mit 5 MFCC Koeffizienten.
(Ergebnisse basieren nicht auf zufälligen Folds)

1. Test: 10 MFCC Koeffizienten (Arousal)

Fenster / Mel bins	256	512	1024	2048	4096	8192
32	0.4586	0.461	0.4619	0.4635	0.4678	0.4752
64	0.4726	0.4698	0.4698	0.4725	0.4763	0.4833
128	0.4818	0.4802	0.4775	0.4794	0.4842	0.4913
256	0.4876	0.4876	0.4858	0.4839	0.4879	0.4943
512	0.4902	0.4906	0.4907	0.4883	0.489	**0.4947**
1024	0.4875	0.4895	0.4899	0.4909	0.4898	0.4947

Tabelle 7.2: Bestimmtheitsmaße für Arousal mit 10 MFCC Koeffizienten.
(Ergebnisse basieren nicht auf zufälligen Folds)

1. Test: 20 MFCC Koeffizienten (Arousal)

Fenster / Mel bins	256	512	1024	2048	4096	8192
32	0.4517	0.4566	0.4597	0.4629	0.4677	0.4743
64	0.4643	0.4653	0.4654	0.4685	0.4723	0.4787
128	0.4745	0.474	0.4695	0.4718	0.4777	0.4847
256	0.481	0.4825	0.479	0.4758	0.4804	**0.4871**
512	0.4828	0.4836	0.485	0.4817	0.4814	0.487
1024	0.4796	0.4804	0.4818	0.4839	0.4845	0.487

Tabelle 7.3: Bestimmtheitsmaße für Arousal mit 20 MFCC Koeffizienten.
(Ergebnisse basieren nicht auf zufälligen Folds)

1. Test: 5 MFCC Koeffizienten (Valence)

Fenster / Mel bins	256	512	1024	2048	4096	8192
32	0.0745	0.0739	0.073	0.076	0.0823	**0.0944**
64	0.0702	0.0741	0.072	0.0742	0.0804	0.0931
128	0.0638	0.0681	0.0712	0.0719	0.0784	0.0909
256	0.0595	0.0589	0.0602	0.0635	0.0718	0.0846
512	0.0548	0.0536	0.0509	0.0558	0.0661	0.0762
1024	0.0551	0.0501	0.0461	0.0466	0.0586	0.0678

Tabelle 7.4: Bestimmtheitsmaße für Valence mit 5 MFCC Koeffizienten. (Ergebnisse basieren nicht auf zufälligen Folds)

1. Test: 10 MFCC Koeffizienten (Valence)

Fenster / Mel bins	256	512	1024	2048	4096	8192
32	0.0552	0.0538	0.0535	0.0571	0.065	0.0807
64	0.0476	0.0561	0.0555	0.0591	0.0664	**0.081**
128	0.0468	0.048	0.0569	0.0592	0.0664	0.0799
256	0.0456	0.0408	0.0409	0.0512	0.0599	0.0739
512	0.0403	0.0375	0.033	0.0395	0.054	0.0653
1024	0.043	0.0338	0.0291	0.0303	0.0444	0.0573

Tabelle 7.5: Bestimmtheitsmaße für Valence mit 10 MFCC Koeffizienten. (Ergebnisse basieren nicht auf zufälligen Folds)

1. Test: 20 MFCC Koeffizienten (Valence)

Fenster / Mel bins	256	512	1024	2048	4096	8192
32	0.0442	0.0491	0.0511	0.0549	0.063	0.0766
64	0.0478	0.0598	0.0581	0.0615	0.068	**0.0813**
128	0.0426	0.0503	0.0538	0.0564	0.0648	0.0785
256	0.0344	0.0378	0.0363	0.0426	0.0548	0.0724
512	0.0271	0.0286	0.0226	0.0314	0.0487	0.0628
1024	0.0265	0.0207	0.0156	0.0168	0.0362	0.0518

Tabelle 7.6: Bestimmtheitsmaße für Valence mit 20 MFCC Koeffizienten.
(Ergebnisse basieren nicht auf zufälligen Folds)

2. Test: Evaluierung über Anzahl der Koeffizienten (Arousal)

MFCC's / Mel bins	2	3	4	5	6	7	8	9	10
512	0.4456	0.4863	0.4947	0.4947	0.4934	0.4922	0.491	0.4916	0.4907
1024	0.4125	0.4851	**0.4967**	0.4964	0.4945	0.4928	0.4906	0.4914	0.4899
2048	0.3783	0.4822	0.4964	0.4962	0.4932	0.4909	0.4893	0.4901	0.4884

Tabelle 7.7: Bestimmtheitsmaße für Arousal von 2 bis 10 MFCC's mit einem Fenster von 1024 Samples (23.2 ms).
(Ergebnisse basieren nicht auf zufälligen Folds)

2. Test: Evaluierung über Anzahl der Koeffizienten (Valence)

MFCC's / Mel bins	2	3	4	5	6	7	8	9	10
16	0.0795	0.0797	0.0854	0.0881	0.0824	0.0839	0.0862	0.0811	0.0803
32	0.0847	0.0848	0.0918	**0.0944**	0.0882	0.088	0.0854	0.0824	0.0807
64	0.0865	0.0869	0.0922	0.0931	0.0869	0.0857	0.0807	0.0821	0.081

Tabelle 7.8: Bestimmtheitsmaße für Valence von 2 bis 10 MFCC's mit einem Fenster von 8192 Samples (185.8 ms).
(Ergebnisse basieren nicht auf zufälligen Folds)

3. Test: Evaluierung über Fensterlänge und Mel bins für 4 MFCC's (Arousal)

Fenster / Mel bins	**256**	**512**	**1024**	**2048**	**4096**
256		0.4899			
512		0.4959			
1024	0.4947	0.4968	0.4967	0.4966	0.4965
2048	0.4949	**0.4974**	0.4964	0.4961	
4096		0.4961			

Tabelle 7.9: Bestimmtheitsmaße für Arousal mit 4 Koeffizienten.
(Ergebnisse basieren nicht auf zufälligen Folds)

3. Test: Evaluierung über Fensterlänge und Mel bins für 5 MFCC's (Valence)

Fenster / Mel bins	**2048**	**4096**	**8192**	**16384**	**32768**	**65535**
16					0.096	
32	0.076	0.0823	0.0944	0.104	0.1047	0.0972
64			0.0931	0.104	**0.1065**	0.1004
128					0.1043	

Tabelle 7.10: Bestimmtheitsmaße für Arousal mit 5 Koeffizienten.
(Ergebnisse basieren nicht auf zufälligen Folds)

7.2 OBSC Evaluierung

Fenster / F_{min}	256	512	1024	2048	4096	8192	16384	32768	65535
25				0.3241	0.3906	0.4345	0.4444	0.4288	0.401
27.5				0.3282	0.3914	0.4382	0.4477	0.4274	0.3999
50			0.3076	0.3539	0.4107	0.4442	0.4519	0.4368	0.4074
55			0.3	0.3594	0.4167	0.4459	**0.4533**	0.4344	0.4066
100		0.3056	0.3292	0.3679	0.4062	0.4345	0.4414	0.429	0.4051
110		0.308	0.3262	0.3651	0.4081	0.4345	0.442	0.4283	0.4053
200	0.2912	0.2794	0.275	0.3267	0.3713	0.4023	0.4128	0.4061	0.388
220	0.2825	0.2736	0.2767	0.326	0.3753	0.405	0.4184	0.4145	0.3981
400	0.2922	0.2601	0.2603	0.301	0.342	0.3821	0.4009	0.3984	0.3833

Tabelle 7.11: Übersicht der Bestimmtheitsmaße für verschiedene Fensterlängen und Frequenzbänder (Arousal).
(Ergebnisse basieren nicht auf zufälligen Folds)

Fenster / F_{min}	256	512	1024	2048	4096	8192	16384	32768	65535
25				-0.0237	-0.0044	0.0447	0.0895	0.1035	0.0772
27.5				-0.0168	-0.0061	0.0422	0.0909	0.1024	0.0754
50			-0.0137	-0.0123	0.009	0.0467	0.0864	0.1011	0.0778
55			-0.0133	-0.0014	0.0143	0.0488	0.0887	0.1011	0.0781
100		0.016	0.014	0.0178	0.0304	0.0583	0.0888	0.0974	0.0802
110		0.0207	0.021	0.03	0.0398	0.067	0.1005	**0.1072**	0.0886
200	0.0013	0.0049	-0.0071	-0.0006	0.0114	0.037	0.0588	0.0658	0.0521
220	-0.0013	0.0111	0.0022	0.0126	0.0247	0.0524	0.0763	0.0828	0.0649
400	0.0373	0.0155	0.0001	-0.0004	0.0123	0.0403	0.0639	0.0645	0.0505

Tabelle 7.12: Übersicht der Bestimmtheitsmaße für verschiedene Fensterlängen und Frequenzbänder (Valence).
(Ergebnisse basieren nicht auf zufälligen Folds)

Einstellung / F_{min}	16384 mean	32768 mean	16384 mean+var	32768 mean+var
5	0.158	0.1494	0.1768	0.1818
10	0.3856	0.3576	0.4019	0.3828
15	0.4315	0.4116	0.4428	0.433
20	0.4428	0.4215	0.4482	0.4342
25	0.4444	0.4288	0.4598	0.4555
27.5	0.4477	0.4274	0.464	0.4507
30	0.4495	0.4294	**0.4703**	0.4589
35	0.4475	0.4239	0.4597	0.4447
40	0.4485	0.4252	0.4522	0.4309
45	0.4549	0.4348	0.46	0.4534
50	0.4519	0.4368	0.4594	0.4559
55	0.4533	0.4344	0.4606	0.4496
60	0.4544	0.4367	0.4626	0.4555
65	0.452	0.4319	0.4615	0.4492
70	0.4502	0.4274	0.4581	0.4444
75	0.4487	0.4283	0.4528	0.4357
80	0.446	0.4256	0.4474	0.4286
85	0.4439	0.4274	0.4432	0.4305
90	0.4447	0.4295	0.4494	0.4472
95	0.4453	0.4326	0.4544	0.4509
100	0.4414	0.429	0.4458	0.4447
105	0.4454	0.4331	0.4484	0.4454
110	0.442	0.4283	0.4433	0.4364
200	0.4128	0.4061	0.4109	0.4187
220	0.4184	0.4145	0.4101	0.4138
400	0.4009	0.3984	0.3993	0.402

Tabelle 7.13: Bestimmtheitsmaße für verschiedene Frequenzbänder von 5 Hz bis 400 Hz (Arousal). Vergleich der vier Einstellungen mit den Zeitfenstern 16384 und 32768 Samples, sowie Hinzunahme der Varianz.
(Ergebnisse basieren nicht auf zufälligen Folds)

Einstellung / F_{min}	**16384 mean**	**32768 mean**	**16384 mean+var**	**32768 mean+var**
5	0.0571	0.0642	0.0599	0.0731
10	0.122	**0.1264**	0.1134	0.1201
15	0.0883	0.095	0.0844	0.0954
20	0.1085	0.121	0.1019	0.1158
25	0.0895	0.1035	0.0897	0.1066
27.5	0.0909	0.1024	0.0951	0.1088
30	0.0973	0.1023	0.101	0.1063
35	0.088	0.096	0.079	0.0906
40	0.1036	0.1137	0.0965	0.105
45	0.0987	0.1069	0.0997	0.1073
50	0.0864	0.1011	0.0882	0.1042
55	0.0887	0.1011	0.0926	0.1071
60	0.09	0.0955	0.0926	0.0974
65	0.0893	0.0955	0.0834	0.0933
70	0.084	0.0909	0.0707	0.0883
75	0.0896	0.1033	0.0837	0.0956
80	0.1029	0.11	0.0935	0.0972
85	0.0971	0.1035	0.0863	0.0897
90	0.0928	0.0962	0.0876	0.0947
95	0.0936	0.1024	0.0893	0.1046
100	0.0888	0.0974	0.0828	0.0989
105	0.0895	0.0982	0.0816	0.0987
110	0.1005	0.1072	0.0908	0.105
200	0.0588	0.0658	0.0482	0.0619
220	0.0763	0.0828	0.0625	0.0759
400	0.0639	0.0645	0.0551	0.0585

Tabelle 7.14: Bestimmtheitsmaße für verschiedene Frequenzbänder von 5 Hz bis 400 Hz (Valence). Vergleich der vier Einstellungen mit den Zeitfenstern 16384 und 32768 Samples, sowie Hinzunahme der Varianz.
(Ergebnisse basieren nicht auf zufälligen Folds)

Grad \ Fenster	110250	220500
3	0.434	0.4335
5	0.4513	0.4501
10	0.4687	**0.4712**

Grad \ Fenster	110250	220500
3	0.2547	0.2621
5	0.2523	0.2591
10	0.2778	**0.2897**

Tabelle 7.15: Bestimmtheitsmaße für Arousal (links) und Valence (rechts) mit CMRARE bei 22050 Hz Samplerate für Polynomgrade 3, 5 und 10

7.3 CMRARE Evaluierung

7.4 Auswahl nicht-cepstraler Merkmale

MRMR für Arousal

	Energy Merkmal	**Test-R^2**
1	RMS peak number in 3 seconds	0.2706
2	Zero-crossing rate 2048	0.4146
3	Root mean square 512	0.4751
4	Low energy 2048	0.476
5	RMS peak number above mean amplitude in 3 seconds	0.4752
6	Sub-band energy ratio 512 #2	0.4695
7	Sub-band energy ratio 512 #4	0.4841
8	Root mean square 2048	0.4942
9	Zero-crossing rate 1024	0.492
10	Low energy 512	0.4915
11	Sub-band energy ratio 1024 #2	0.4893
12	Root mean square 1024	0.4932
13	Sub-band energy ratio 1024 #4	0.4935
14	Sub-band energy ratio 2048 #1	0.5171
15	Low energy 1024	0.5195
16	Zero-crossing rate 512	0.5181
17	Sub-band energy ratio 512 #3	0.5041
18	Sub-band energy ratio 2048 #2	0.5003
19	Sub-band energy ratio 2048 #4	0.5131
20	Sub-band energy ratio 2048 #3	0.5154

	Timbre Merkmal	**Test-R^2**
1	Spectral flatness measure 2048 #1	0.3309
2	Distances in phase domain 1024	0.4707
3	Spectral brightness 1024	0.5054
4	Spectral crest factor 1024 #2	0.5062
5	Spectral irregularity 512	0.512
6	Spectral bandwidth 512	0.5101
7	Spectral crest factor 512 #1	0.5089
8	Spectral flatness measure 1024 #4	0.5072
9	Distances in phase domain 2048	0.5033
10	Spectral flatness measure 2048 #2	0.5027
11	Spectral skewness 512	0.506
12	Spectral irregularity 2048	0.5119
13	Spectral brightness 512	0.5114
14	Distances in phase domain 512	0.5071
15	Spectral flatness measure 1024 #1	0.5091
16	Spectral crest factor 512 #2	0.5109
17	Spectral centroid 512	0.5217
18	Spectral flatness measure 2048 #3	0.5259
19	Spectral irregularity 1024	0.5244
20	Spectral extent 2048	0.5355

Tabelle 7.16: MRMR-Rangfolge der ersten 20 *Energy* und *Timbre* Merkmale mit Bestimmtheitsmaßen (Arousal)

	Harmony and Melody Merkmal	**Test-R^2**
1	Harmonic change detection function 2048	0.311
2	Tristimulus 512 #1	0.4032
3	Inharmonicity 512	0.4872
4	Tristimulus 512 #2	0.5211
5	Strengths of minor keys 512 #4	0.5201
6	Harmonic change detection function 1024	0.5217
7	Chroma DCT-Reduced log Pitch #9	0.5204
8	Harmonic change detection function 4096	0.522
9	Tonal centroid vector 512 #6	0.5215
10	Inharmonicity 1024	0.5211
11	Chroma DCT-Reduced log Pitch #1	0.5208
12	Harmonic change detection function 512	0.5309
13	Tonal centroid vector 4096 #4	0.5301
14	Majorminor alignment 1024	0.5291
15	Strengths of major keys 512 #5	0.5275
16	Tonal centroid vector 4096 #3	0.5272
17	Strengths of minor keys 2048 #10	0.526
18	Number of different chords in 10s	0.5276
19	Inharmonicity 2048	0.527
20	Chroma DCT-Reduced log Pitch #3	0.5267

	Tempo and Rhythm Merkmal	**Test-R^2**
1	Estimated onset number per minute	0.2422
2	Characteristics of fluctuation patterns #3	0.2487
3	Rhythmic clarity	0.2968
4	Estimated beat number per minute	0.3066
5	Characteristics of fluctuation patterns #1	0.3028
6	Tempo based on onset times	0.3131
7	Five peaks of fluctuation curves #4	0.3127
8	Five peaks of fluctuation curves #5	0.3134
9	Five peaks of fluctuation curves #1	0.3176
10	Characteristics of fluctuation patterns #6	0.3269
11	Characteristics of fluctuation patterns #5	0.3866
12	Five peaks of fluctuation curves #3	0.3857
13	Five peaks of fluctuation curves #2	0.3861
14	Characteristics of fluctuation patterns #7	0.3919
15	Estimated tatum number per minute	0.3913
16	Characteristics of fluctuation patterns #4	0.3932
17	Characteristics of fluctuation patterns #2	0.3924

Tabelle 7.17: MRMR-Rangfolge der ersten 20 *Harmony and Melody* und aller *Tempo and Rhythm* Merkmale mit Bestimmtheitsmaßen (Arousal)

MRMR für Valence

	Energy Merkmal	**Test-R^2**
1	RMS peak number in 3 seconds	0.135
2	Low energy 512	0.1375
3	Zero-crossing rate 1024	0.1523
4	Root mean square 512	0.1495
5	RMS peak number above mean amplitude in 3 seconds	0.1516
6	Sub-band energy ratio 512 #4	0.1486
7	Low energy 2048	0.1594
8	Sub-band energy ratio 512 #2	0.16
9	Sub-band energy ratio 512 #3	0.1509
10	Low energy 1024	0.1474
11	Root mean square 2048	0.1999
12	Sub-band energy ratio 2048 #4	0.1978
13	Sub-band energy ratio 2048 #2	0.1967
14	Zero-crossing rate 512	0.1953
15	Sub-band energy ratio 2048 #3	0.1854
16	Root mean square 1024	0.1911
17	Sub-band energy ratio 1024 #1	0.1594
18	Sub-band energy ratio 1024 #4	0.1551
19	Sub-band energy ratio 1024 #2	0.1531
20	Sub-band energy ratio 1024 #3	0.1524

	Timbre Merkmal	**Test-R^2**
1	Spectral bandwidth 1024	0.1161
2	Spectral irregularity 512	0.1218
3	Spectral crest factor 2048 #1	0.1269
4	Spectral extent 2048	0.1359
5	Spectral crest factor 512 #3	0.1347
6	Distances in phase domain 512	0.137
7	Spectral crest factor 512 #1	0.1369
8	Spectral irregularity 2048	0.1431
9	Spectral crest factor 2048 #4	0.1479
10	Angles in phase domain 512	0.1451
11	Spectral skewness 512	0.1526
12	Spectral flatness measure 512 #2	0.1504
13	Spectral flatness measure 2048 #1	0.1484
14	Spectral flatness measure 2048 #4	0.1465
15	Spectral extent 1024	0.1437
16	Spectral irregularity 1024	0.1405
17	Distances in phase domain 1024	0.14
18	Spectral flatness measure 512 #1	0.143
19	Spectral flatness measure 2048 #3	0.1452
20	Spectral skewness 2048	0.1663

Tabelle 7.18: MRMR-Rangfolge der ersten 20 *Energy* und *Timbre* Merkmale mit Bestimmtheitsmaßen (Valence)

	Harmony and Melody Merkmal	**Test-R^2**
1	Harmonic change detection function 2048	0.1074
2	Strengths of minor keys 1024 #8	0.1053
3	Majorminor alignment 4096	0.1096
4	Inharmonicity 2048	0.1318
5	Number of chord changes in 10s	0.1494
6	Chroma DCT-Reduced log Pitch #4	0.1504
7	Strengths of minor keys 4096 #12	0.1516
8	Chroma DCT-Reduced log Pitch #2	0.1512
9	Harmonic change detection function 4096	0.1665
10	Chroma DCT-Reduced log Pitch #9	0.161
11	Tristimulus 512 #2	0.1635
12	Strengths of minor keys 512 #11	0.1619
13	Chroma DCT-Reduced log Pitch #6	0.1609
14	Tonal centroid vector 1024 #3	0.1588
15	Tristimulus 512 #1	0.1629
16	Tonal centroid vector 2048 #5	0.1661
17	Inharmonicity 1024	0.1655
18	Strengths of minor keys 1024 #3	0.1809
19	Tonal centroid vector 4096 #4	0.1793
20	Strengths of minor keys 512 #8	0.1781

	Tempo and Rhythm Merkmal	**Test-R^2**
1	Rhythmic clarity	0.149
2	Estimated onset number per minute	0.1921
3	Characteristics of fluctuation patterns #3	0.2166
4	Tempo based on onset times	0.2217
5	Characteristics of fluctuation patterns #6	0.2463
6	Estimated tatum number per minute	0.2483
7	Characteristics of fluctuation patterns #5	0.2894
8	Five peaks of fluctuation curves #5	0.2899
9	Five peaks of fluctuation curves #4	0.2871
10	Characteristics of fluctuation patterns #1	0.2875
11	Estimated beat number per minute	0.2919
12	Characteristics of fluctuation patterns #4	0.2927
13	Five peaks of fluctuation curves #2	0.2945
14	Five peaks of fluctuation curves #1	0.2977
15	Five peaks of fluctuation curves #3	0.2951
16	Characteristics of fluctuation patterns #2	0.3047
17	Characteristics of fluctuation patterns #7	0.3028

Tabelle 7.19: MRMR-Rangfolge der ersten 20 *Harmony and Melody* und aller *Tempo and Rhythm* Merkmale mit Bestimmtheitsmaßen (Valence)

Abbildungsverzeichnis

Literaturverzeichnis

[1] Mirex 2007: Audio music mood classification. `http://www.music-ir.org/mirex/wiki/2007:Audio_Music_Mood_Classification`. Aufgerufen am 26.3.2016.

[2] Moodswings. `http://music.ece.drexel.edu/mssp/`. Aufgerufen am 8.2.2016.

[3] Vincent Akkermans, Joan Serrà, and Perfecto Herrera. Shape-based spectral contrast descriptor. In *Proceedings of the 6th Sound and Music Computing Conference*, 2009.

[4] Jesús Piedrafita Arilla. Multiple linear regression.

[5] Jean-Julien Aucouturier and Francois Pachet. Improving timbre similarity : How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1:1, 2004.

[6] Luke Barrington, Douglas Turnbull, Damien O'Malley, and Gert Lanckriet. User-centered design of a social game to tag music. *ACM KDD Workshop on Human Computation*, 2009.

[7] Bruce P. Bogert, Michael J.R. Healy, and John W. Tukey. The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, volume 15, 1963.

[8] Judith C. Brown. Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, 89:1:425–434, 1991.

[9] Geoffrey L. Collier. Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, 35:1, 2007.

[10] Manuel Davy. An introduction to statistical signal processing and spectrum estimation. In Anssi Klapuri and Manuel Davy, editors, *Signal Processing Methods for Music Transcription*. Springer, 2006.

[11] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the 2003 IEEE Bioinformatics Conference*, 2003.

[12] Antti Eronen. *Signal Processing Methods for Audio Classification and Music Content Analysis.* dissertation, Tampere University of Technology, 2009.

[13] Alf Gabrielsson. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 33:3, 2002.

[14] Debalina Ghosh, Depanwita Sarkar Debnathand, and Saikat Bose. A comparative study of performance of fpga based mel filter bank and bark filter. *International Journal of Artificial Intelligence and Applications*, 3:2, 2012.

[15] Patrick Gomez and Brigitta Danuser. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7:2:377–87, 2007.

[16] Stephen Hainsworth. Beat tracking and musical metre analysis. In Anssi Klapuri and Manuel Davy, editors, *Signal Processing Methods for Music Transcription.* Springer, 2006.

[17] Byeong-jun Han, Seungmin Rho, Roger Dannenberg, and Eenjun Hwang. Smers: Music emotion recognition using support vector regression. In *Proceedings of the 8th International Conference on Music Information Retrieval*, 2009.

[18] Christopher Harte and Mark Sandler. Detecting harmonic change in musical audio. In *Proceedings of the 1st Audio and Music Computing for Multimedia Workshop*, pages 21–26, 2006.

[19] Kate Henver. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48:2, 1936.

[20] Karl Kristoffer Jensen. Timbre models of musical sounds. Technical Report 99:7, University of Copenhagen, 1999.

[21] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings of the IEEE International conference on Multimedia and Expo*, volume 1, 2002.

[22] Patrik N. Juslin. Cue utilization in communication of emotion in music performance: relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26:6:797–813, 2000.

[23] Patrik N. Juslin and Petri Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33:3, 2004.

[24] Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proceedings of the 11th International Conference on Music Information Retrieval*, pages 255–266, 2010.

[25] Jochen Krimphoff, Stephen McAdams, and Suzanne Winsberg. Caractérisation du timbre des sons complexes. ii analyses acoustiques et quantication psychophysique. *Journal de Physique IV*, 04:C5):pp. C5–625–C5–628, 1994.

[26] Olivier Lartillot. Mirtoolbox 1.4 user's manual. Technical report, Finnish Centre of Excellence in Interdisciplinary Music Research and Swiss Center for Affective Sciences, 2012.

[27] Olivier Lartillot and Petri Toiviainen. Mir in matlab (ii): A toolbox for musical feature extraction from audio. In *Proceedings of the 8th International Conference on Music Information Retrieval*, page 127–130, 2007.

[28] Edith L. M. Law, Luis von Ahn, Roger B. Dannenberg, and Mike Crawford. Tagatune: A game for music and sound annotation. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 361–364, 2007.

[29] Dan Liu, Lie Lu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 14:1, 2006.

[30] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval*, 2000.

[31] Michael I. Mandel and Daniel P. W. Ellis. A web-based game for collecting music metadata. *Journal Of New Music Research*, 37:151–165, 2008.

[32] Rainer Martin and Anil Nagathil. Cepstral modulation ratio regression (cmrare) parameters for audio signal analysis and classification. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[33] Rainer Martin and Anil Nagathil. Digital filters and spectral analysis. In Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Günter Rudolph, editors, *Music Data Analysis: Foundations and Applications*. CRC Press, 2016. to appear.

[34] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[35] Cory McKay and Ichiro Fujinaga. jmir: Tools for automatic music classification. In *Proceedings of the International Computer Music Conference*, pages 65–68, 2009.

[36] Martin Mckinney and Jeroen Breebaart. Features for audio and music classification. *Proceedings of the International Symposium on Music Information Retrieval*, pages 151–158, 2003.

[37] Ingo Mierswa and Katharina Morik. Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58:2-3, 2005.

[38] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 935–940, 2006.

[39] Fabian Mörchen, Alfred Ultsch, Mario Nöcker, and Christian Stamm. Databionic visualization of music collections according to perceptual distance. In *Proceedings of the 6th International Conference on Music Information Retrieval*, page 396–403, 2005.

[40] Meinard Müller. *Information Retrieval for Music and Motion.* Springer-Verlag, 2007.

[41] Meinard Müller and Sebastian Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Conference on Music Information Retrieval*, pages 215–220, 2011.

[42] Anil Nagathil and Rainer Martin. Signal-level features. In Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Günter Rudolph, editors, *Music Data Analysis: Foundations and Applications.* CRC Press, 2016. to appear.

[43] Alan V. Oppenheim and Ronald W. Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 2004.

[44] Tae H. Park. *Time-Domain Signal Processing I.* World Scientific Publishing Company, 2009.

[45] Bryan Pellom. Automatic speech recognition: From theory to practice. Technical report, Department of Computer Science Center for Spoken Language Research University of Colorad, 2004.

[46] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 27:8, 2005.

[47] Rudolf Rasch and Reiner Plomp. The perception of musical tones. In Diana Deutsch, editor, *The Psychology of Music.* Academic Press, INC., 2013.

[48] Thomas Rossing, Richard Moore, and Paul Wheeler. *Hearing.* Addison-Wesley, 2001.

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

[50] Günther Rötter and Igor Vatolkin. Emotions. In Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Günter Rudolph, editors, *Music Data Analysis: Foundations and Applications*. CRC Press, 2016. to appear.

[51] Erik Schmidt, Douglas Turnbull, and Youngmoo Kim. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the 10th International Conference on Music Information Retrieval*, 2010.

[52] Erik M. Schmidt, Matthew Prockup, Jeffery Scott, Brian Dolhansky, Brandon G. Morton, and Youngmoo E. Kim. Relating perceptual and feature space invariances in music emotion recognition. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval*, pages 534–542, 2012.

[53] Mohammad Soleymani, Michael N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *CrowdMM '13 Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 1–6, 2013.

[54] Robert Thayer, Robert Newman, and Tracey McClain. Self-regulatoin of mood: Strategies for changing a bad mood, raising energy, and reducing tension. *Journal of Personality and Social Psychology*, 67:5, 1994.

[55] Wolfgang Theimer, Igor Vatolkin, and Antti Eronen. Definitions of audio features for music content description. Technical Report TR08-2-001, Faculty of Computer Science, Technische Universiät Dortmund, 2008.

[56] Douglas Turnbull, Ruoran Liu, Luke Barrington, and Gert Lanckriet. A game-based approach for collecting semantic annotations of music. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 535–538, 2007.

[57] George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organised Sound*, 4:3:pp. 169–175, 2000.

[58] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10:5:pp. 293–302, 2002.

[59] Igor Vatolkin. *Improving Supervised Music Classification by Means of Multi-Objective Evolutionary Feature Selection.* dissertation, Faculty of Computer Science, Technische Universiät Dortmund, 2013.

[60] Igor Vatolkin, Wolfgang Theimer, and Martin Botteck. Amuse (advanced music explorer) - a multitool framework for music data analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval*, page 33–38, 2010.

[61] Luis von Ahn. Games with a purpose. *Computer*, 39:6:92–94, 2006.

[62] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, and Liming Chen. What is the best segment duration for music mood analysis? In *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, 2008.

[63] Yi-Hsuan Yang and Homer H. Chen. Emotion recognition. In *Music Emotion Recognition*, chapter 2.2. CRC Press, 2011.

[64] Yi-Hsuan Yang and Homer H. Chen. Emotion recognition. In *Music Emotion Recognition*, chapter Dimensional Approach. CRC Press, 2011.

[65] Yi-Hsuan Yang and Homer H. Chen. *Music Emotion Recognition.* CRC Press, 2011.

[66] Kelly H. Zou, Kemal Tuncali, and Stuart G. Silverman. Correlation and simple linear regression. *Radiology*, 227:3, 2003.