Technische Universität Berlin
Fakultät I – Geistes- und Bildungswissenschaften
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation

Masterarbeit

Ein Algorithmus zur Bestimmung der rhythmischen Ähnlichkeit von Musikstücken

Nico Lehrbach

[redacted]

2. Juni 2017

1. Gutachter: Prof. Dr. Stefan Weinzierl
2. Gutachter: Dr. Holger Kirchhoff

Zusammenfassung

In dieser Arbeit wird die Rhythmusähnlichkeit als ein Parameter für den Vergleich und die Klassifikation von Musik untersucht. Bestehende Ähnlichkeitsmaße wie die Rhythmic Patterns und die Onset Patterns werden miteinander und mit einem davon abgeleiteten eigenen Ansatz, den Bar Patterns, verglichen. Zudem wird eine Beschreibung der Rhythmen auf Basis von Rhythmusfamilien in den Vergleich mit einbezogen, diese erfordern Kenntnis über die Positionen der Taktanfänge. Für die Evaluation liefert ein durchgeführter Hörversuch quantitative, empirische Referenzwerte der rhythmische Ähnlichkeit von kurzen Musikauschnitten.

Die Ergebnisse des Hörversuchs zeigen, dass die Aufgabe, Musikausschnitte auf rhythmische Ähnlichkeit zu bewerten, nicht völlig losgelöst von anderen, insbesondere klanglichen Parametern behandelt werden kann. Dennoch lassen sich konsistente Daten beobachten. Der musikalische Hintergrund zeigt einen signifikanten Einfluss auf die Streuung der Bewertung rhythmischer Ähnlichkeit zwischen verschiedenen Teilnehmenden.

Bei der Evaluation der algorithmischen Ähnlichkeitsmaße zeigen insbesondere die Onset Patterns und die Bar Patterns eine hohe Korrelation mit den empirischen Daten. Diese Ergebnisse können auch bei einer Klassifikationsaufgabe auf dem Ballroom Datensatz bestätigt werden. Eine Kombination der Onset- und Bar Patterns verbessern die Ergebnisse für beide Datensätze. Die rhythmische Ähnlichkeit basierend auf den Rhythmusfamilien liefert für den empirischen Datensatz vergleichbare Korrelationskoeffizienten und erzielt bei der Klassifikationsaufgabe bessere Ergebnisse.

Abstract

Audio-based music similarity estimation is a common task in Music Information Retrieval. This work investigates the rhythmic similarity of short music excerpts, seen as one component of music similarity. Derived from existing measures of music similarity like the Rhythmic Patterns and the Onset Patterns, a new approach is presented, the Bar Patterns. A comparision of these measures with a rhythmic similarity measure based on rhythmic families is performed. In order to determine rhythmic families, the knowledge of bar positions is required. A survey is carried out in order to gain a quantitative, empirical ground truth of the rhythmic simiarity of short music excerpts presented to participants.

The results of the survey indicate that an investigation of only rhythmic aspects of music excerpts suffers from the influence of other aspects, e.g. tonal parameters. However, consistent data can still be observed. The variance of given similarity ratings of same pairs of songs depends significantly on the musical background of the participants.

Onset Patterns, Bar Patterns and the measure based on rhythmic families show high correlations with the empirical data. A further evaluation based on genre classification on the Ballroom Dataset confirms these results. A combination of Onset Patterns and Bar Patterns improves results on both datasets.

Danksagung

Mein großer Dank gebührt Herrn Prof. Dr. Stefan Weinzierl und zplane.development für das Ermöglichen dieser Arbeit. Insbesondere Tim Flohrer und Dr. Holger Kirchhoff standen mir jederzeit mit Rat, Tat und den richtigen Fragen zur Seite. Desweiteren danke ich Dr. Jochen Steffens und Athanasios Lykartsis für die Unterstützung insbesondere bei der Planung des Hörversuches.

Fakultät |
Geisteswissenschaften

Prüfungsausschüsse

Studiendekan und Vorsitzender

Prof. Dr. Friedrich Steinle

Eidesstattliche Erklärung

Ist jeder an der TU Berlin verfassten schriftlichen Arbeit eigenhändig unterzeichnet beizufügen!

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen.

Titel der schriftlichen Arbeit

VerfasserIn/VerfasserInnen*

Name | Vorname | Matr.-Nr.

Betreuende/r DozentIn

Name | Vorname

Mit meiner Unterschrift bestätige ich, dass ich über fachübliche Zitierregeln unterrichtet worden bin und verstanden habe. Die im betroffenen Fachgebiet üblichen Zitiervorschriften sind eingehalten worden.

Eine Überprüfung der Arbeit auf Plagiate mithilfe elektronischer Hilfsmittel darf vorgenommen werden.

Ort, Datum | Unterschrift**

*Bei Gruppenarbeiten sind die Unterschriften aller VerfasserInnen erforderlich.
**Durch die Unterschrift bürgen Sie für den vollumfänglichen Inhalt der Endversion dieser schriftlichen Arbeit.

Inhaltsverzeichnis

1. Einleitung **11**

2. Grundlagen **15**
2.1. Physikalische Grundlagen des Schalls 15
2.2. Psychoakustische Grundlagen 16
2.2.1. Tonhöhenwahrnehmung 17
2.2.2. Lautstärkewahrnehmung 18
2.2.3. Räumliche Position 19
2.2.4. Klangfarbe 19
2.2.5. Gruppieren 19
2.2.6. Zeitliche Verarbeitung 22
2.3. Grundlagen der digitalen Signalverarbeitung 22
2.3.1. Fouriertransformation 23
2.3.2. Kurzzeitfouriertransformation 24
2.3.3. Chroma Differenz Funktion 24
2.3.4. Distanzmaße 24
2.3.5. Korrelation 25
2.3.6. Rhythmogramm 25
2.3.7. Viterbi-Algorithmus 26
2.4. Verwendete Datensätze 26

3. Rhythmus **29**
3.1. Wahrnehmung von Rhythmus 29
3.1.1. Puls 30
3.1.2. Gruppieren 30
3.2. Rhythmusähnlichkeit 32
3.2.1. Symbolische Ähnlichkeit 33
3.2.2. Ähnlichkeit auf Basis des Audiomaterials 36
3.3. Abgeleiteter Ansatz für den Rhythmusvergleich 37

4. Algorithmus **41**
4.1. Rhythmusdarstellung durch Bar Patterns 41
4.1.1. Vorverarbeitung 41
4.1.2. Rhythmogramm 42
4.1.3. Tatum 43
4.1.4. Taktlänge 46
4.1.5. Rhythmusdarstellung 48
4.1.6. Vergleich von Bar Patterns 49

4.2. Rhythmusfamilien 51
4.2.1. Zeitlich Auflösung 52
4.2.2. Rhythmus der Instrumente 52
4.2.3. Rhythmus in Frequenzbändern 52
4.2.4. Vergleich der Rhythmusfamilien 53

5. Empirische Untersuchung **55**
5.1. Operationalisierung 55
5.1.1. Datenerfassung 56
5.1.2. Musikauswahl 56
5.2. Versuchsaufbau und -ablauf 58
5.3. Versuchsergebnisse 59
5.3.1. Datenbereinigung 59
5.3.2. Stichprobe 60
5.3.3. Auswertung 60
5.3.4. Auswertung nach Gruppen 63

6. Evaluation **69**
6.1. Datensatz des Hörversuchs 70
6.2. Ballroom Datensatz 74

7. Zusammenfassung und Ausblick **79**

A. Anhang **83**

Literatur **87**

1. Einleitung

Soll Musik charakterisiert werden, so wird im Allgemeinen zunächst das Genre des entsprechenden Stückes genannt. Die Klassifikation in Genres und bei feinerer Unterscheidung in Subgenres ist weit verbreitet, sie basiert allerdings auf vielen verschiedenen Aspekten wie z.B. der Instrumentierung, Rhythmik, harmonischen Eigenschaften, auch Tempi oder Textinhalten. Musikstücke gleicher Genres weisen gleiche oder ähnliche Eigenschaften in verschiedenen Parametern auf, auch zwischen (Sub-)Genres können Verwandschaften und Ähnlichkeiten durch Vergleich einzelner musikalischer Eigenschaften formuliert werden. Es liegen allerdings nicht immer eine klare Definition und deutliche Abgrenzung vor, Musikstücke können verschiedenen Genres gleichzeitig angehören und es lassen sich Musikstücke unterschiedlicher Genres finden, die sich in bestimmten Aspekten ähnlicher sind, als manche Musikstücke gleichen Genres.

Die automatische Klassifikation und der automatische Vergleich von Musik muss nicht zwangsläufig in Genres arbeiten. Für ein präzises Vorgehen ist in jedem Fall ein getrenntes Erfassen der einzelnen Parameter erforderlich, insbesondere der klanglichen, rhythmischen und formalen Eigenschaften der Musikstücke. Dies ermöglicht dann eine individuelle Gewichtung der betrachteten Parameter und eine Anpassung der Vergleiche an verschiedene Aufgabenstellung. So können durch einen ausgereiften Musikvergleich beispielsweise Musikstücke in Genres eingeteilt werden, Musikstücke nach harmonischer Komplexität sortiert werden oder in einer automatisierten DJ-Software eine Folge von Musikstücken gefunden werden, die rhythmisch zueinander passen und eine Ähnlichkeit aufweisen.

In dieser Arbeit soll der rhythmische Aspekt von Musikähnlichkeit untersucht und ein Algorithmus entwickelt werden, der die rhythmische Ähnlichkeit bewerten kann. Dies wird als ein Teilaspekt der Musikähnlichkeit aufgefasst und kann einen Beitrag zur genaueren (algorithmischen) Beschreibung und dem Vergleich von Musik liefern. Das ganz allgemeine Vorgehen für Ähnlichkeitsmaße ist das Gewinnen einer aussagekräftigen Repräsentation und die Entwicklung einer entsprechenden Metrik (Hampton, 1998, S. 4). Für die Rhythmusähnlichkeit bedeutet dies, dass aus dem Musikstück zunächst eine

Beschreibung des Rhythmus bzw. eine repräsentative Rhythmusdarstellung extrahiert werden muss, auf die dann eine passende Ähnlichkeitsmetrik angewendet wird.

Gerade das Konzept der Ähnlichkeit birgt jedoch Probleme. In früheren Untersuchungen zeigte sich die Bewertung von Ähnlichkeit in ganz verschiedenen Bereichen der Wahrnehmung als flexibel und mit großen individuellen Unterschieden behaftet (vgl. Goldstone, 1994). Die individuellen Unterschiede wurden auch in einfachen Untersuchungen zur Rhythmusähnlichkeit beobachtet, dennoch zeigte sich auch eine gewisse Konsistenz in der Ähnlichkeitsbewertung (Cao et al., 2014; Gómez-Marin et al., 2015). Die Entwicklung eines Maßes für rhythmische Ähnlichkeit scheint somit möglich, die Überprüfung und Auswertung bedarf jedoch besonderer Aufmerksamkeit.

Ein verbreiteter Ansatz zur Überprüfung von Vergleichsalgorithmen basiert auf einer Klassifikation. Es werden dabei Musikstücke durch einen Vergleichsalgorithmus in Gruppen eingeteilt, die einander ähnliche Stücke enthalten. Mit der Annahme, dass die Klassifikation z.B. in Genres nach ähnlichen Parametern vorgenommen wird, liefert die Übereinstimmung der Gruppierungen ein Maß für die Qualität des Vergleichsalgorithmus (z.B. McKinney und Breebaart, 2003). Gelegentlich wird auch eine Gruppierung nach Interpreten als Referenz gewählt (z.B. Mandel und Ellis, 2005)). Mit z.B. dem Ballroom Datensatz (vgl. Kapitel 2.4) existiert auch ein Datensatz, der eine Gruppierung nach überwiegend rhythmischen Merkmalen kategorisiert und damit für die Evaluation von rhythmischen Deskriptoren und Vergleichsmetriken deutlich besser geeignet ist.

Eine Kategorisierung aufgrund von Ähnlichkeit erfordert jedoch neben der Repräsentation und Metrik den zusätzlichen Schritt der Einteilung in Gruppen. Das zunächst metrische Maß wird in ein binäres (zu einer Gruppe gehörend oder nicht) reduziert, dabei gehen Informationen über Ähnlichkeiten zwischen einzelnen Elementen verloren. Ein Teil dieser Arbeit ist daher die Erstellung eines Datensatzes, der eine empirisch abgefragte Rhythmusähnlichkeit enthält. In einem empirischen Hörversuch wird dafür die Bewertung der paarweisen rhythmischen Ähnlichkeit von Musikausschnitten erfasst.

Die Evaluation des hier entwickelten Ansatzes für den Rhythmusvergleich findet im Vergleich mit vorhandenen Verfahren statt, insbesondere der Fluctuation Patterns und der Onset Patterns (vgl. Kapitel 3.2.2). Auch die Theorie der Rhythmusfamilien von Cao et al. (2014) wird angepasst und in den Vergleich einbezogen. Neben dem empirisch gewonnen Datensatz wird eine Klassifikation auf dem Ballroom Datensatz zur Überprüfung vorgenommen.

Im ersten Teil dieser Arbeit werden die hier relevanten Grundlagen der Akustik, der digitalen Signalverarbeitung und der Psychoakustik kurz zusammengefasst. Es folgt ein

kurzer Überblick über den aktuellen Stand der Rhythmusforschung sowie vorhandene Ansätze des Rhythmusvergleiches. Anschließend wird der hier verwendete Ansatz zum rhythmischen Vergleich zweier Musikausschnitte abgeleitet.

In der zweiten Hälfte (ab Kapitel 4) werden die algorithmische Umsetzung des Rhythmusvergleiches sowie der angesetzte Hörversuch beschrieben und ausgewertet. Die Evaluation des eigenen Rhythmusvergleiches erfolgt auf zwei Datensätzen im Vergleich mit den bisher verfügbaren Verfahren.

2. Grundlagen

Die automatische und algorithmische Analyse von Musik arbeitet mit Verfahren der digitalen Signalverarbeitung. Die Auswahl der Verfahren richtet sich sowohl nach den physikalischen Eigenschaften des Schalls als auch nach den Verarbeitungsmethoden in der menschlichen Schallwahrnehmung, die in der Psychoakustik beschrieben werden. Im Folgenden werden die im weiteren Verlauf der Arbeit vorausgesetzen Grundlagen kurz erläutert.

2.1. Physikalische Grundlagen des Schalls

Wird eine Schallquelle angeregt, so gerät sie in eine schwingende Bewegung und gibt diese an die sie umgebende Luft weiter. Innerhalb der Luft erfolgt eine Weiterleitung der Schwingungen in Form vom Druck- und Dichteänderungen, die sich wellenförmig ausbreiten und vom menschlichen Gehör als Schallereignis wahrgenommen werden können (Möser, 2009). Die Stärke der Luftdruckänderungen wird üblicherweise mit dem Effektivwert des Wechselschalldrucks als *Schalldruck* p in der Einheit Pascal (Pa) angegeben, bzw. in der Audiotechnik in logarithmischer Skalierung als Schalldruckpegel

$$L_p = 10 \cdot log_{10} \left(\frac{p^2}{p_0^2} \right)$$

mit dem Bezugswert $p_0 = 2 \cdot 10^{-5}$ Pa in der Pseudoeinheit dBSPL. Bei energiebasierter Betrachtung liefert die *Schallintensität* I die Schallenergie, die pro Zeit und Fläche übertragen wird.

Innerhalb der menschlichen Hörgrenzen gilt in der Luft als übertragendes Medium das Prinzip der Linearität (Möser, 2009). Das bedeutet insbesondere, dass sich die Schalldrücke zweier Schallquellen einfach addieren und somit die Summe zweier Schallereignisse genau ihre Summe ergibt. Im Umkehrschluss folgt aus der Linearität ebenfalls, dass ein Schallereignis als Summe beliebig vieler Teilschallwellen aufgefasst werden kann. Ein aus der Mathematik lange bekanntes Verfahren zur Zerlegung von periodischen Funktionen ist die Fouriertransformation, die auch aus psychoakustischer Sicht

sinnvoll ist. Eine Schallwelle wird dabei in eine Summe von Sinusschwingungen zerlegt, die durch ihre Frequenz (Tonhöhe) und Amplitude (Lautstärke), sowie Phasenlage zueinander vollständig beschrieben sind. Die Verteilung der beteiligten Frequenzen bildet das *Spektrum* eines Klangs. Jeder natürlich vorkommende Klang enthält hierbei eine Vielzahl beteiligter Frequenzen verschiedener Amplituden, die grob in tonale und rauschhafte Klänge eingeteilt werden können. Tonale Klänge bestehen aus mehreren diskreten Frequenzen, die sich aus einem Grundton (tiefster Frequenzanteil) und mehreren Obertönen (Vielfache der Grundfrequenz) zusammensetzen. Das Spektrum rauschhafter Klänge hingegen weist Energieanteile in ausgedehnten Frequenzbereichen (Frequenzbändern) auf.

2.2. Psychoakustische Grundlagen

Akustische Information erreichen das menschliche Gehör über das Trommelfell, auf das winzige Luftdruckänderungen als Funktion der Zeit auftreffen. Unter anderem sorgen die Haarzellen des Corti-Organs im Innenohr für eine Transformation der Schalldruckänderungen in ein elektrisches Nervensignal, dessen neuronale Verarbeitung im auditiven Cortex in der Großhirnrinde erfolgt (Weinzierl, 2008).

Das menschliche Ohr reagiert auf Luftdruckänderungen im Bereich von etwa $2 \cdot 10^{-5}$ Pa (Hörschwelle) bis zu 200 Pa (Schmerzgrenze), bei beteiligten Teilschwingungen im Frequenzbereich von etwa 20 Hz bis 20.000 Hz (Möser, 2009). Der Lautstärkebereich deckt dabei etwa sieben Größenordnungen ab, der Frequenzbereich drei. Die Wahrnehmung solch großer Wertebereiche wird durch eine logarithmische Empfindlichkeit erreicht, für merkliche Änderungen muss bei hohen Lautstärken bzw. Frequenzen eine größere Änderung erfolgen als bei niedrigen.

Bei den wahrgenommenen Luftdruckänderungen handelt es sich um eine einzige, ungerichtete physikalische Größe. Dennoch ist der Mensch in der Lage, sehr genaue Informationen über die Zusammensetzung des akustischen Signals zu extrahieren: es können verschiedene Schallquellen getrennt erkannt und im Raum lokalisiert, sowie Informationen über die Schallquelle und den Raum gewonnen werden. Auch komplexe Informationen wie Musik und Sprache werden dekodiert. Dies alles erfordert einen komplexen und fein abgestimmten Analyseprozess.

Nach Bregman (1994) ist eine grundlegende Fähigkeit des Gehörs die *auditory scene analysis (ASA)*. Der ständig eintreffende Strom akustischer Informationen wird durch *simultanes* und *sequentielles Gruppieren* in einzelne auditive Objekte zerlegt, die sich

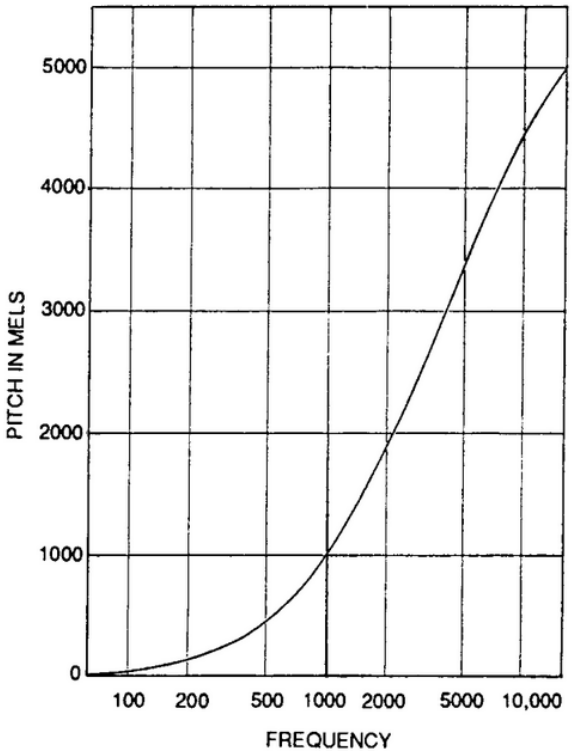

Abbildung 2.1.: Tonheitskurve der Mel Skala aus Houtsma (1995).

sowohl zeitlich als auch in Frequenzbereichen überlappen können. Die Gruppierung erfolgt dabei nach Tonhöhe, Lautstärke, Klangfarbe und räumlicher Position.

2.2.1. Tonhöhenwahrnehmung

Der erste Analyseschritt im menschlichen Ohr ist die Frequenzanalyse. Die genaue Position der Transformation des eintreffenden (Teil-)Schalls in Nervensignale auf der etwa 32 mm langen Oberfläche des Corti-Organs im Innenohr ist abhängig von der Frequenz, es handelt sich um eine Frequenz-Ort-Transformation (Rasch und Plomp, 1999). Die Wahrnehmung der Tonhöhe hängt neben der Frequenz geringfügig auch von weiteren Parametern wie der Dauer und dem Lautstärkeverlauf des Klangs ab (Houtsma, 1995). Zudem ist der Zusammenhang zwischen wahrgenommenener Tonhöhe (der *Tonheit*) und der Frequenz kein linearer, wie empirische Untersuchungen zeigen. So entstand durch die Aufgabenstellung, Tonpaare aus reinen Sinustönen zu finden, deren Tonhöhen dem Verhältnis „doppelt so hoch“ bzw. „halb so hoch“ entsprechen, die Mel Skala, die die Tonfrequenz (in Hz) der Tonheit (in Mel) zuordnet (Houtsma, 1995), siehe Abbildung 2.1.

Vielfach durchgeführte Hörversuche mit Maskierungseffekten zeigen, dass die akustische Wahrnehmung in Frequenzbändern erfolgt. Für diese Untersuchung wird bestimmt, ab welcher Lautstärke ein Sinuston gegenüber einem schmalbandigen Rauschen gleicher Mittenfrequenz wahrgenommen werden kann. Mit steigender Bandbreite bei gleichbleibender Energiedichte steigt diese Lautstärke, allerdings nur bis zu einer bestimmten Bandbreite des Rauschens, welche genau einem Frequenzband der auditiven Wahrnehmung entspricht. Eine weitere Steigerung der Bandbreite hat keinen weiteren Effekt. Durch Experimente lassen sich 24 Frequenzbänder bestimmen, deren Breite bis etwa 500 Hz konstant ist und danach mit steigender Frequenz zunimmt. Die Frequenzbänder werden in der Frequenzgruppenskala mit der Einheit Bark zusammengefasst. Es zeigt sich, dass die Frequenzgruppenskala und die Mel Skala einander mit der Umrechnung 1 bark = 100 Mel sehr gut entsprechen.

Trotz der festgestellten nichtlinearen Zusammenhängen zwischen Frequenz und Tonheit ist für die Tonhöhe und das harmonische Empfinden die streng logarithmische (gleichstufige) chromatische Tonleiter der westlichen Musik ausschlaggebend. Der Tonvorrat von zwölf Halbtönen wird dabei in konstanten Verhältnissen innerhalb einer Oktave (Verdopplung der Frequenz) untergebracht, als Referenzwert gilt der Kammerton a mit meist 440 Hz.

2.2.2. Lautstärkewahrnehmung

Die physikalische Beschreibung der Stärke des Schallereignisses als Effektivwert des Wechselschalldrucks bzw. als Schalldruckpegel ist nur ein Parameter für die wahrgenommene Lautstärke, die *Lautheit*. Das Lautstärkeempfinden ist stark frequenzabhängig, so reagiert das menschliche Gehör auf Frequenzen zwischen 500 Hz und 5000 Hz viel empfindlicher als auf Frequenzen am Rand des Hörspektrums (Weinzierl, 2008, S.54). Zudem findet insbesondere die Lautstärkewahrnehmung in Frequenzbändern statt, breitbandige Schallsignale gleicher Intensität werden lauter wahrgenommen, je mehr Frequenzbänder sie überstreichen (Weinzierl, 2008, S.61). Diese Effekte wurden in empirischen Untersuchungen quantitativ erfasst und bilden die Basis für die Lautheitsskala *Sone*, welche die wahrgenommene Lautheit von Schall angibt. Für sehr kurze Schallereignisse zeigt sich weiter eine Zunahme der Lautheit mit steigender Dauer des Schallereignisses (Plack und Carlyon, 1995). *Maskierungseffekte* sorgen dafür, dass Teiltöne von lauteren Teiltönen sowohl zeitlich (bei genügend kleiner Differenz unabhängig von der Reihenfolge) als auch im Frequenzbereich (innerhalb eines auditiven Frequenzbandes) überdeckt werden (Moore, 1995).

Ein Modell zur Beschreibung und Berechnung der Lautheit eines Schallsignals, das die genannten psychoakustischen Effekte berücksichtigt und Ergebnisse liefert, die den Ergebnissen empirischer Untersuchungen gut entsprechen, liefern Zwicker und Fastl (2006).

2.2.3. Räumliche Position

Für die räumliche Positionierung einer Schallquelle sind in erster Linie Laufzeit- und Lautstärkeunterschiede zwischen den beiden Ohren, sowie Verzerrungen des Schallsignals durch die asymmetrische Form der Ohrmuscheln verantwortlich (Zwicker und Fastl, 2006, Kap.15). In dieser Arbeit wird ausschließlich mit Monosignalen gearbeitet, die Richtungswahrnehmung spielt hier keine Rolle.

2.2.4. Klangfarbe

Praktisch alle natürlich vorkommenden Schallereignisse sind komplexe Klänge, die verschiedene Frequenzanteile enthalten. Bei tonalen Klängen bildet die Struktur bzw. Gewichtung der Obertöne, bei rauschhaften und gemischten Klängen die Energieverteilung in den verschiedenen Frequenzbereichen die Grundlage für die Klangfarbenwahrnehmung (Risset und Wessel, 1999).

2.2.5. Gruppieren

Bregman (1994) unterscheidet bei der auditory scene analysis zwischen simultanem und sequentiellem Gruppieren. Diese werden im Folgenden getrennt erläutert, sind in der Wahrnehmung jedoch nicht unabhängig voneinander, wie Experimente zeigen (Darwin, 2005). Den Vorgang des simultanen Gruppierens nennt er *auditory source separation*, das sequentielle Gruppieren wird mit *auditory stream segregation* bezeichnet. Als Basis dienen in beiden Fällen universelle, nicht auf die auditorische Wahrnehmung beschränkte Gruppierungsprinzipien, die insbesondere in der visuellen Wahrnehmung eindrücklich darstellbar sind und als Gestaltwahrnehmung bezeichnet werden (Deutsch, 1999). Schon im frühen 20. Jahrhundert wurden eine Reihe von einfachen Regeln formuliert, nach denen das Bilden von Gruppen beobachtet werden kann, eine Illustration für die visuelle Wahrnehmung zeigt Abbildung 2.2. Wird von Gruppieren gesprochen, so kann dies sowohl durch Zusammenfassen von Elementen der gleichen Gruppe als auch durch Trennung von Elementen verschiedener Gruppen geschehen.

Nähe Objekte geringen Abstands voneinander werden eher zu einer Gruppe zusammengefasst, als Objekte größeren Abstands. In Abbildung 2.2a bilden sich vier Gruppen bestehend aus zwei Punkten.

Ähnlichkeit Es werden die Objekte zusammengefasst, die mehr gemeinsame Parameter haben. So bilden sich in Abbildung 2.2b Gruppen aus gefüllten und aus leeren Kreisen.

Gute Fortführung Elemente, die einander in einer bestimmten Richtung folgen, erscheinen zusammengehörend. Die Wahrnehmung zweier Linien AB und CD in Abbildung 2.2c ist die Folge.

Gemeinsames Schicksal Objekte, welche die gleiche Änderung eines Parameters erfahren, werden in der Wahrnehmung gruppiert.

Vertrautheit Wir tendieren dazu, Gruppen zu formen, die uns bekannt erscheinen.

Geht ein Schallereignis von einer einzelnen Schallquelle aus, so erklingen alle beteiligten Schallanteile gleichzeitig. Die Synchronität der Klanganteile ist ein wichtiges Element für die Aufteilung des eintreffenden Schalls in verschiedene auditive Objekte (Bregman, 1994) und kann den Gruppierungsregeln der (zeitlichen) Nähe und des gemeinsamen Schicksals zugeordnet werden. Schon kurze Differenzen in den Anfängen von Klängen und Klanganteilen ab 15 - 30 ms erzeugen die Wahrnehmung zweier verschiedener Schallquellen (Bregman, 1994).

Tonale Klänge und Klanganteile, wie sie in der Natur oft vorkommen, weisen ein Obertonspektrum bestehend aus der Grundfrequenz und ihren ganzzahligen Vielfachen auf. Dieses Spektrum ist dem Gehör wohlbekannt und führt nach dem Prinzip der Vertrautheit zu einer Gruppierung und als Folge daraus zu der Wahrnehmung eines einzelnen Tons mit spezifischer Klangfarbe (Deutsch, 1999). Enthält das Signal Frequenzen, die als Obertöne zu einem oder weiteren Grundtönen gehören, so können diese als getrennte Töne wahrgenommen werden (Bregman, 1994).

Zusammen mit der hier nicht weiter behandelten räumlichen Lokalisation gehören diese beiden Vorgänge zur simultanen Gruppierung, da die Schallquellen praktisch sofort identifiziert werden können und nicht erst im Laufe einiger Zeit perzeptiv herausgebildet werden. Letzteres ist bei der sequentiellen Gruppierung zu beobachten, hier wird eher von *auditory streams* (Strömen) gesprochen, die per auditory stream segregation voneinander getrennt werden.

Dem Prinzip des gemeinsamen Schicksals entspricht eine Gruppierung aufgrund von gleicher zeitlicher Veränderungen der Frequenz (z.B. beim Vibrato) oder der Lautstärke.

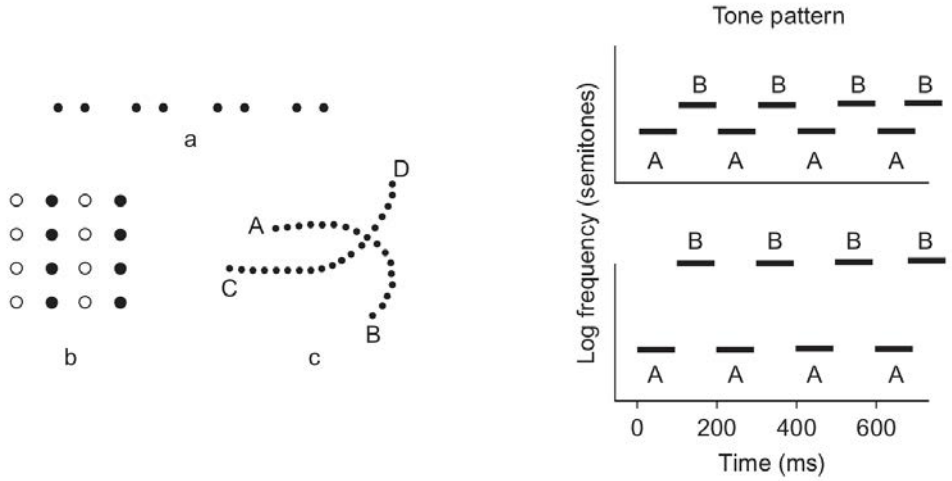

Abbildung 2.2.: **Links:** Illustration der Gestaltprinzipien Nähe (a), Ähnlichkeit (b) und gute Fortfürung (c). **Rechts:** Beispiel für die Gruppierung nach der Nähe der Tonhöhe. Beides aus Deutsch (1999)

Die Klanganteile, die eine parallele Bewegung ausführen, können zusammengefasst wahrgenommen werden. Dieser Effekt kann zwar nachgewiesen werden (Bregman et al., 1985; Darwin und Ciocca, 1992), ist jedoch nicht besonders stark und wird kontrovers diskutiert (Deutsch, 1999). So verändert sich die Obertonstruktur mancher Instrumente im Verlaufe des Tons, die verschiedenen Lautstärkeverläufe der einzelnen Obertöne führt jedoch nicht zu einer Aufteilung des Klangs in verschiedene auditive Objekte (Deutsch, 1999). Bei Frequenzanteilen, die nicht in einem harmonischen Frequenzverhältnis (Obertonstruktur) stehen, konnten Teilnehmende eines Hörversuchs von Carlyon und McAdams (1992) zudem nicht zwischen kohärenten und nicht-kohärenten Frequenzmodulationen unterscheiden.

Stärker, sich jedoch jedoch erst nach längeren Zeiträumen ausbildend, ist die Trennung nach Tonhöhennähe. Hörende gruppieren schnelle, sequenzielle Tonfolgen aus verschiedenen Tonhöhenregionen dem Prinzip der guten Fortführung folgend nach den Regionen und können so parallel verlaufende Tonfolgen oder Melodien wahrnehmen, wie in Abbildung 2.2 (rechts) nach Deutsch (1999) illustriert ist. Dieses Prinzip fand insbesondere in der Barockmusik bei Kompositionen für einstimmige Instrumente Anwendung, wo dennoch der Eindruck verschiedener Melodielinien entstehen kann.

Die hier genannten Beispiele der auditory source separation, der auditory stream segregation und der Gruppierungen bilden nur den für diese Arbeit relevanten Teil des aktuellen Forschungsstandes ab, für einen Überblick siehe Deutsch (1999).

2.2.6. Zeitliche Verarbeitung

Akustische Ereignisse, die weniger als etwa 5 ms auseinanderliegen, werden als gleichzeitig wahrgenommen (Pöppel, 1997). Sie fallen in ein Integrationsfenster, das in einem Block verarbeitet wird, die Länge des Intergrationsfensters ist jedoch stark abhängig von der Art des eintreffenden Schallsignals. Die 5 ms sind dabei die Mindestlänge des Integrationsfensters und wurden ermittelt als der Zeitabstand, der nötig ist, um kurze Klickgeräusche als zwei verschiedene Ereignisse wahrnehmen zu können. Abhängig von vielen Faktoren wie Länge, Lautstärke oder Art der akustischen Ereignisse kann der erforderliche Zeitabstand um vieles größer sein. Auch die Reihenfolge zweier verschiedener Klänge kann erst bei deutlich längeren Abständen erkannt werden. Um zwei verschiedene Ereignisse innerhalb einer komplexeren akustischen Struktur wie z.B. Musik gut voneinander trennen zu können werden erforderliche Zeitabstände von mindestens 100 ms angegeben (Fischinger, 2009, S.48f).

Eine obere Grenze in der zeitlichen Verarbeitung wird in der Literatur als psychische Präsenzzeit angegeben. Sie umfasst die maximale Dauer, innerhalb der akustische Ereignisse als zusammengehörend erfasst werden können, z.B. eine Folge von Tönen als Melodie wahrgenommen wird. Je nach genauer Definition und Quelle finden sich Angaben zwischen 2 s und 5 s (vgl. Fischinger, 2009, S.50).

2.3. Grundlagen der digitalen Signalverarbeitung

Die digitale Signalverarbeitung erfordert zunächst eine digitale Darstellung des Schallsignals, also des Schalldruckes in Abhängigkeit der Zeit. Aufgrund der endlichen Speicherkapazität von Computern muss das kontinuierliche Signal diskretisiert werden, sowohl im Zeit- als auch im Wertebereich. Für den Zeitbereich bedeutet dies eine Abtastung des Signals mit einer endlichen Anzahl an Abtastzeitpunkten (*Samples*) pro Zeiteinheit. Ein gängiger Wert ist eine Abtastrate (sampling frequency *fs*) von 44.100 Hz. Die Diskretisierung des Wertebereichs führt zu einer endlichen Anzahl an Amplitudenstufen, eine typische Samplingtiefe von z.B. 16 Bit führt zu $2^{16} = 65536$ möglichen Werten. Gemäß dem Nyquist-Theorem entspricht die höchste darstellbare Frequenz (Nyquist-Frequenz f_{Ny}) der halben Abtastrate (Weinzierl, 2008).

2.3.1. Fouriertransformation

Um ein Zeitsignal $x(t)$ in den Frequenzbereich $X(f)$ zu transformieren, also in eine Summe aus Sinusschwingungen verschiedener Frequenzen und Amplituden zu zerlegen, bzw. die Rücktransformation vorzunehmen, liefert die Fouriertransformation folgende Gleichungen:

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} x(t)e^{-2\pi ift}dt \\ x(t) &= \int_{-\infty}^{\infty} X(f)e^{2\pi ift}df \end{aligned} \tag{2.1}$$

mit i^2 = -1 und $e^{it} = cos(t) + i \cdot sin(t)$.

Diese Gleichungen geben die Fouriertransformation für ein kontinuierliches, im Allgemeinen komplexes Signal $x(t)$ an. Für in der digitalen Signalverarbeitung vorkommende diskrete Signale $x[n]$ mit N Samples Länge lauten die Transformationsgleichungen dann:

$$\begin{aligned} X(f) &= \sum_{n=0}^{N-1} x[n]e^{-2\pi in\frac{f}{fs}} = X_R(f) + iX_I(f) \\ x(t) &= \sum_{f=0}^{F-1} X[f]e^{2\pi if\frac{f}{fs}}. \end{aligned} \tag{2.2}$$

Von Interesse ist oftmals besonders das Betragsspektrum

$$|X(f)| = \sqrt{X_R^2(f) + X_I^2(f)}, \tag{2.3}$$

das für Frequenzen bis zur Nyquistfrequenz berechnet werden kann. Die diskrete Fouriertransformation führt so zu einer Zerlegung des Signals in $N/2 + 1$ Sinusfunktionen, deren Basisfrequenzen mit linearem Abstand zwischen 0 Hz und der maximalen Frequenz f_{Ny} liegen. Die Frequenzauflösung ist abhängig von der Signallänge und über den gesamten Frequenzbereich konstant.

Mit der schnellen Fourier Transformation (fast fourier transform, FFT) existieren für die Berechnung der diskreten Fouriertransformation sehr effiziente Algorithmen (Weinzierl, 2008), sofern die Länge des Signals einem Produkt kleiner Primfaktoren (in besten Fall einer 2er Potenz) entspricht.

2.3.2. Kurzzeitfouriertransformation

Meist ist nicht die Frequenzdarstellung des gesamten Signals von Interesse, sondern die Veränderung der spektralen Zusammensetzung. Für diesen Fall wird das Signal in (sich überlappende) Blöcke unterteilt und jeder Block einzeln transformiert. Dieses Vorgehen wird als Kurzzeit-Fourier-Transformation (short-time Fourier transform, *STFT*) bezeichnet, als Resultat ergibt sich ein *Spektrogramm*. Passt ein Vielfaches der jeweiligen spektralen Grundfrequenz nicht genau in einen Block (was im Allgmeinen der Fall ist), so kommt es zu Verschmierungen im Frequenzbereich (Leck-Effekt), was durch eine geringere Gewichtung der äußeren Samples des Blocks (*Fensterung*) abgeschwächt wird. Die STFT ist durch die Größen *blocksize* (Blocklänge), *hopsize* (Differenz zwischen den ersten Samples zweier benachbarter Blöcke) und die Art der Fensterung charakterisiert.

2.3.3. Chroma Differenz Funktion

Die Tonalität und Harmonie ist nicht unmittelbar aus dem Spektrum oder Spektrogramm eines Musiksignal abzulesen, da spektrale Anteile eines Tons über mehrere Oktaven verteilt sein können. Aus diesem Grund schlägt Leman (1995) eine Reduktion des Spektrums auf einen zwölfdimensionalen Vektor vor. Jeder Eintrag entspricht der Signalenergie für jeden der zwölf Halbtöne (oktavlagenunabhängig, daher besser Tonklassen, *pitch class* oder *chroma*) westlicher Musik, die über den Mittelwert aller entsprechenden Frequenzbins des Spektrums bestimmt wird. Angewendet auf das Spektrogram ergibt sich ein *Pitch Class Profile* (Fujishima, 1999) oder Chromagramm, das Auskunft über den Verlauf der Tonalität gibt. Für die *Chroma Differenz Funktion* (CDF) wird die Manhattan Distanz zweier aufeinanderfolgender Tonklassenvektoren bestimmt, wodurch Änderungen in der Tonalität angezeigt werden.

2.3.4. Distanzmaße

Wie bei der Chroma Differenz Funktion sind häufig Abstände oder Distanzen zwischen mehrdimensionalen Vektoren von Interesse. Die oben genannte *Manhattan Distanz* ist definiert als die Summe der absoluten Differenz der Einzelkoordinaten. Ein weiteres häufig verwendetes Abstandsmaß ist die *Euklidische Distanz*; hier werden die Differenzen der Einzelkoordinaten quadratisch aufsummiert und anschließend die Wurzel gezogen (Cha, 2007).

2.3.5. Korrelation

Die Korrelation beschreibt die Ähnlichkeit zweier Signale bei verschiedenen Verschiebungen zueinander. Für endliche, diskrete Signale $x[i]$ und $y[i]$ der Längen l_x und l_y werden dazu für jede mögliche Verschiebung η der Signale zueinander die Produkte aller übereinanderliegenden Samples aufsummiert (Lerch, 2012, S.25).

$$r_{xy}(\eta) = \sum_{i=max(0,-\eta)}^{min(l_x,l_y-\eta)} x[i] \cdot y[i+\eta] \tag{2.4}$$

Sind $x[i]$ und $y[i]$ verschieden voneinander, so handelt es sich um die *Kreuzkorrelation*, die Korrelation eines Signales mit sich selber nennt sich *Autokorrelation*. Letztere zeigt Periodizitäten in dem Signal auf.

Die Korrelationsfunktion wird üblicherweise normalisiert mit dem Faktor

$$\lambda_c = \frac{1}{\sqrt{\left(\sum_{i=0}^{l_x} x^2[i]\right)\left(\sum_{i=0}^{l_y} y^2[i]\right)}}. \tag{2.5}$$

Insbesondere bei der Autokorrelation und der Kreuzkorrelation von Signalen etwa gleicher Länge steigt die Anzahl der Summanden, die eine Multiplikation mit Null enthalten (die Signale überlappen sich nur teilweise), was zu einer Dreiecksform der Korrelationsfunktion mit dem Maximum bei einer Verschiebung von $\eta = 0$ führt. Eine Division durch die Anzahl der sich überlappenden Samples kann diese Gewichtung kompensieren.

2.3.6. Rhythmogramm

Wie auch die Fouriertransformation kann die Autokorrelationsfunktion eines Signals in fortschreitenden, sich überlappenden Blöcken bestimmt werden. Analog zur Kurzzeitfouriertransformation wäre dies eine Kurzzeitautokorrelation, K. Jensen (2004) spricht von einem Rhythmogramm. Das Rhythmogramm stellt den zeitlichen Verlauf der lokalen Selbstähnlichkeit dar, zeigt also die Veränderung von Periodizitäten an.

2.3.7. Viterbi-Algorithmus

Eine häufige Vorgehensweise in der algorithmischen Analyse von Audiomaterial ist es, zunächst Wahrscheinlichkeiten für eine zu bestimmende Größe zu verschiedenen Zeitpunkten zu berechnen und erst in einem späteren Schritt die Entscheidung über den tatsächlichen Wert zu treffen. Ein Beispiel wäre es, in regelmäßigen Abständen das Tempo eines Musikstückes abzuschätzen, indem für einen jeweils lokalen Ausschnitt (Block) die Wahrscheinlichkeit für verschiedenen Tempi berechnet wird. Für den Gesamtverlauf des Tempos würde dann für jeden Block ein Tempo ausgewählt werden. Das ausgewählte Tempo hängt hierbei von den Tempowahrscheinlichkeiten des jeweiligen Block ab, jedoch auch von dem ausgewählten Tempo des vorherigen Blocks, da das wahrgenommene Tempo sich erst mit etwas Zeit etablieren kann und somit Temposprünge nicht beliebig auftreten.

Eine mathematische Beschreibung dieses Problems liefern Markow-Ketten, die durch eine endliche Anzahl von Zuständen (z.B. diskrete Tempi) sowie Übergangswahrscheinlichkeiten zwischen allen möglichen Zuständen beschrieben werden (Behrends, 2000). Da jedoch die einzelnen Zustände nicht direkt beobachtet werden können, sondern nur eine vom Zustand abhängige Wahrscheinlichkeitsverteilung („Emissionen“) wird hier von einem *hidden markov model* gesprochen (Rabiner und Juang, 1986). Einen optimalen Pfad, also eine Abfolge von Zuständen, die sowohl die lokale Wahrscheinlichkeit als auch die Übergangswahrscheinlichkeit berücksichtigt, findet der Viterbi-Algorithmus (Forney, 1973).

2.4. Verwendete Datensätze

Neben der Erhebung eines eigenen Datensatzes wird in dieser Arbeit mit zwei weiteren Datensätzen gearbeitet. Der *GTZAN Music Genre Dataset*[1] (GTZAN) ist ein frei verfügbarer Datensatz, der häufig für die automatische Genreklassifizierung genutzt wird. Er besteht aus 1000 etwa 30-sekündigen Musikausschnitten, die in zehn Genres mit jeweils 100 Tracks eingeteilt sind: Blues, Klassik, Country, Disco, HipHop, Jazz, Metal, Pop, Reggea, und Rock. Der gesamte Datensatz wurde nachträglich manuell mit Informationen zu Tempo, Beats, Taktanfängen und Swing-Ratio versehen (Marchand et al., 2015), und dient somit als Referenzdatensatz für die automatische Extraktion von Tempo, Taktlängen und Taktanfängen.

[1] zu finden auf http://marsyas.info/downloads/datasets.html

Der *Ballroom Datensatz*[2] bietet 698 Musikausschnitte von 30 Sekunden Länge, die nach Standardtänzen gruppiert sind (Gouyon et al., 2006). Die genauen Gruppen und Anzahl der Beispiele sind in Tabelle 2.1 aufgeführt. Es wird angenommen, dass eine Gruppierung nach Tänzen eine bessere rhythmische Gruppierung liefert, als eine Gruppierung nach allgemeinen Musikgenres. Der Ballroom Datensatz wird daher als zusätzlicher Datensatz für die Rhythmusähnlichkeit verwendet. Auch für den Ballroom Datensatz sind manuelle Annotationen des Tempos und der Takte verfügbar[3] (Krebs et al., 2013).

Tanz	Anzahl	Tanz	Anzahl
Cha-Cha-Cha	111	Samba	86
Jive	60	Tango	86
Quickstep	82	Wiener Walzer	65
Rumba	98	Langsamer Walzer	110

Tabelle 2.1.: Tänze und Anzahl der Beispiele des Ballroom Datensatzes.

[2]verfügbar auf http://mtg.upf.edu/ismir2004/contest/tempoContest/node5.html
[3]https://github.com/CPJKU/BallroomAnnotations

3. Rhythmus

Rhythmus ist ein Begriff, der Anwendung in der Beschreibung vieler alltäglicher Dinge findet, sei es der Herzschlag, die Gezeiten, das Leben in einer Stadt. So vielseitig wie die Verwendung des Begriffs, so vielseitig sind auch wissenschaftliche Definitionsversuche (Fischinger, 2009). Allen gemein ist jedoch der Bezug auf die Ordnung zeitlicher Abläufe. So kann auch der Rhythmus in Musik als ihre zeitliche Strukturierung aufgefasst werden. Dort müssen allerdings Grenzen gezogen werden, da auch die Form z.B. eines Sinfoniesatzes eine zeitliche Struktur beschreibt, für die der Begriff Rhythmus nicht verwendet wird[1]. Ebenso wird die zeitliche Feinstruktur, das Mikrotiming, nicht zwangsläufig mit dem Rhythmus verknüpft. Die Abgrenzung und genaue Verwendung des Begriffs Rhythmus in dieser Arbeit soll sich an den Vorgängen der Wahrnehmung von Rhythmus orientieren.

3.1. Wahrnehmung von Rhythmus

Die menschliche Wahrnehmung ist spezialisiert darauf, Veränderungen in der Umwelt wahrzunehmen. So ziehen Bewegungen oder plötzliche Geräusche in einer sonst ruhigen Umgebung schnell die Aufmerksamkeit auf sich. Der zugrunde liegende Mechanismus aus neurophysiologischer Sicht ist die Adaption. Erreicht ein gleichbleibender Ton das Ohr, so gibt es eine Nervenreaktion, die zunächst sehr stark ist, jedoch praktisch sofort absinkt. Das Absinken erfolgt zunächst sehr schnell und flacht dann ab, ist jedoch über lange Zeiträume zu beobachten (Moore, 2012, S.50). Beim plötzlichen Beenden des Tons ist allerdings kein besonderer Anstieg zu beobachten.

Auch für die Wahrnehmung von Musik gilt, dass plötzliche Änderungen einen starken Wahrnehmungsreiz erzeugen. Diese werden meist durch die Tonanfänge (*Onsets*) der beteiligten Instrumente hervorgerufen. Nach Zwicker und Fastl (2006) erzeugen lokale Maxima in der Lautheitsfunktion nach dem von ihnen entwickelten Modell einen

[1] Es gibt tatsächlich Arbeiten, in denen eine solch weite Definition von Rhythmus verwendet wird (Cooper und Meyer, 1963), eine engere Definition ist jedoch verbreiteter und besser mit der Benutzung im allgemeinen Sprachgebrauch vereinbar.

rhythmisch relevanten Wahrnehmungsreiz, solange diese einen bestimmten Schwellenwert überschreiten und mindestens 120 ms auseinanderliegen.

Eine Rhythmuswahrnehmung entsteht aus mehreren aufeinanderfolgenden akustischen Ereignissen bzw. rhythmisch relevanten Wahrnehmungsreizen. Nach Bruhn (2000) und Snyder (2000, S.159) etabliert sich bereits bei zwei aufeinanderfolgenden Ereignissen innerhalb der psychischen Präsenszeit ein Rhythmus (vgl. Fischinger, 2009, S.23f). Rhythmus wird dabei als „psychologisches Konstrukt“ (Bruhn, 2000) gesehen, dem zwei Mechanismen zugrunde liegen. Die *beat induction* beschreibt die Fähigkeit des Menschen, aus einer komplexen Folge akustischer Ereignisse (z.B. Musik) einen gleichmäßigen Puls abzuleiten (Honing, 2012), in den die Ereignisse eingebettet werden. Durch weiteres *Gruppieren* oder Hierarchisieren werden diese zu größeren Einheiten zusammengefasst, die ein Metrum, eine schwer-leicht Gewichtung aufweisen (Bruhn, 2000).

3.1.1. Puls

Das Ableiten eines gleichmäßigen Pulses ist eine möglicherweise angeborene (Honing, 2012) und bis auf wenige Ausnahmen nur bei Menschen beobachtete Fähigkeit (Patel et al., 2009), die außerhalb der aktiven Wahrnehmung stattfindet (Honing, 2012) und nicht einmal zwangsläufig eine periodische Abfolge von akustischen Ereignissen erfordert (Clarke, 1985). Dennoch spielen auch persönliche Präferenzen bei der Selektion eines Pulses aus komplexen rhythmischen Signalen eine Rolle. Handel (1993) präsentierte den Teilnehmenden einer empirischen Studie Polyrhythmen verschiedener Konfigurationen (zwei oder drei isochrone Patterns gleichzeitig mit jeweils zwei, drei, fünf oder sieben Schlägen pro Zeiteinheit) und Tempi, zu denen sie mit dem Finger gleichmäßig klopfen sollten. Die gewählte Klopfgeschwindigkeit, der wahrgenommene Puls, zeigte Präferenzen für Tempi und Klangfarben, die sich zwischen den Teilnehmenden jedoch unterschieden. Insgesamt bewegten sie sich meist zwischen 200 ms und 800 ms Intervalldauer (entspricht 75 bis 300 Schlägen pro Minute (*bpm*)).

3.1.2. Gruppieren

Das zeitliche Zusammenfassen zu Gruppen und Strukturen hingegen wird als ein gelernter Vorgang gesehen. Bruhn (2000) benennt für die zeitliche Strukturierung die Notwendigkeit eines kulturspezifischen Referenzsystems, zum Beispiel das Taktsystem der westlich-mitteleuropäischen Musik, timeline-pattern in Teilen afrikanischer Musik oder die Gliederung durch spezifische Instrumente (z.B. Gong) in südostasiatischer Musik. Er formuliert also klar die Orientierung an bekannten Strukturen.

Abbildung 3.1.: Beispiel der Wahrnehmung einer *loud rest* auf der ersten Zählzeit des zweiten Taktes bei Interpretation des Rhythmus in gerader Zählzeit (linke Seite) nach Honing (2012)

Nach Clarke (1999) und Huron und Royal (1996) bilden subjektive Akzentuierungen die Grundlage für das Bilden von Wahrnehmungseinheiten. Die Autoren unterscheiden zwischen *phenomenal accents* (Akzente durch physikalische Parameter wie Lautstärke, Klangfarbe oder Dauer eines Tones), *structural accents* (Akzente durch bekannte Strukturen wie z.B. eine Kadenz), *metrical accents* (Akzente durch zeitliche Fortführung bestehender Strukturen) und *melodic accents* (Akzente durch besondere Töne innerhalb einer Melodie). Die Erweiterung der Akzentuierungswahrnehmung über physikalische Parameter hinaus auf eine subjektive, kognitive Ebene wird durch Erfahrungen und Experimente gestützt. So wird eine Reihe von kurzen, identischen akustischen Ereignissen (z.B. Klickgeräuschen), die in zeitlich immer gleichem Abstand erklingen, von Hörenden in Gruppen eingeteilt, die durch wahrgenommene (aber physikalisch nicht vorhandene) Akzente gestützt werden (Handel, 1993). Die Einteilung erfolgt nur innerhalb gewisser Tempi, davon abhängig werden meist Gruppen aus zwei, drei oder vier Elementen gebildet (Fraisse, 1982). Ein weiterer Effekt ist die Wahrnehmung von *loud rests* („lauten Pausen“) durch Synkopierungen (vorziehen einer betonten Zählzeit auf eine unbetonte), ein Beispiel hierfür nach Honing (2012) zeigt Abbildung 3.1. Von einem Drumcomputer gespielt klingen beide Beispiele gleich. Die in westlicher Musik übliche Präferenz für gerade Zählzeiten führt im linken Beispiel zu einer leeren Eins auf dem zweiten Takt, einer lauten Pause. Diese Wahrnehmung verschwindet bei subjektiver Einteilung in drei-achtel-Gruppen (vgl. auch Honing (2012)).

Deutsch (1999) argumentiert auch bei zeitlicher Gruppierung mit den Prinzipien der Gestaltwahrnehmung. Das stärkste Gruppierungsmerkmal sei die zeitliche Nähe, sehr kleine zeitliche Verschiebungen eines Tones um etwa 30 ms nach hinten führten zur Bildung von zusammenhängenden Phrasen auf dem verschobenen Ton beginnend. Auch das Prinzip der Tonhöhennähe ist auf die zeitliche Strukturierung anwendbar, in der Tonhöhe ähnliche Bereiche können zu Phrasen zusammengefasst werden.

Ausgehend vom erfassten Puls der Musik und den Gruppierungs- (Deutsch, 1999) bzw. Akzentwahrnehmungsvorgängen (Clarke, 1999) formuliert Todd (1994) ein hierarchisches Strukturmodell, das sowohl in zeitlich kleinere als auch größere Bereiche entwickelt wird. Die Grundannahme ist, dass alle rhythmisch relevanten Wahrnehmungsreize

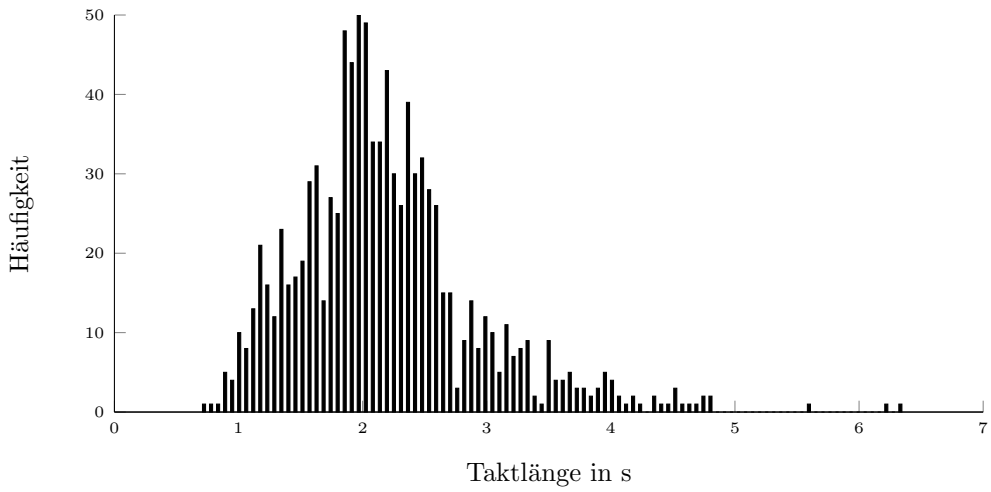

Abbildung 3.2.: Histogramm über die mittleren Taktlängen der 1000 Musikausschnitte des GTZAN-Datensatzes

in ein isochrones Raster aus Vielfachen des Pulses „zurechtgehört“ werden. Diese Quantisierung wurde in Experimenten ausführlich untersucht und bestätigt (Povel, 1981). Der kleinste zeitliche Abstand zwischen zwei Rhythmusereignissen ist das *Tatum*, dessen Elemente zu nächsthöheren Rhythmuseinheiten gruppiert werden, die durch eine stärkere Akzentuierung hervorgehoben sind. Dieser Vorgang wird wiederholt, bis sich verschiedene, hierarchisch angeordnete Rhythmusebenen bilden, darunter der Puls, in westlicher Musik der *Takt*, bis hin zu größeren Formen wie musikalischen Phrasen. Die einzelnen Ebenen unterscheiden sich um kleine ganzzahlige Faktoren voneienader, meist um den Faktor zwei, drei, vier oder sechs. Diese Struktur entspricht sehr gut der Darstellung des Rhythmus in westlicher Notation in Form von Viertelnoten, Achtelnoten, Triolen und in selteneren Fällen n-tolen. Jede Rhythmusebene weist ein eigenes Metrum auf, nach Bruhn (2000) ist für die Rhythmuswahrnehmung die metrische Struktur einer Rhythmusebene entscheidend, die etwa dem Zeitrahmen der psychischen Präsenszeit entspricht. Eine Analyse des GTZAN-Datensatzes zeigt, dass diese Rhythmusebene sehr gut den Takten der Musik entspricht, wie an der Verteilung der Taktlängen zu erkennen ist (Abbildung 3.2).

3.2. Rhythmusähnlichkeit

Die Auffassung von Rhythmus als kognitives Konstrukt mit interindividuell verschiedener Wahrnehmung (z.B. des Pulses bei polyphonen Rhythmen) lässt einen objektiven Vergleich von Rhythmen verschiedener Musikstücke als diffizile Aufgabe erscheinen.

Dennoch zeigen Experimente zur Rhythmusähnlichkeit, dass sich trotz individueller Unterschiede in vielen Fällen durchaus konsistente Bewertungen ergeben (Gómez-Marín et al., 2015; Gómez-Marin et al., 2015; Cao et al., 2014).

Die Ideen zur Beschreibung von rhythmischer Ähnlichkeit können unterschieden werden in solche, die mit notierten, also symbolisch vorliegenden Rhythmen arbeiten, und solche, die Audiomaterial als Ausgangspunkt verwenden, also eine algorithmische Extraktion der rhythmischen Merkmale vornehmen. In allen Fällen werden zunächst Abstände (Distanzen) der Rhythmen berechnet, die als Kehrwert der Ähnlichkeit betrachtet werden.

3.2.1. Symbolische Ähnlichkeit

Um aus einem notiert vorliegenden Rhythmus ein (einzahliges) Abstandsmaß zu gewinnen, muss die Notation in einer Form vorliegen, die die Anwendung eines Distanzmaßes erlaubt. Toussaint (2004) zeigt die gängigen Rhythmusdarstellungen (Abbildung 3.3), von denen insbesondere die letzten beiden eine Abstandsberechnung zulassen. Die binäre Notation (7.) ordnet den Rhythmus auf einem isochronen Gitter an (beruhend auf dem Tatum), Einheiten mit einem Onset erhalten eine 1, alle anderen eine 0. Die Intervalllängennotation (8.) basiert auf dem gleichen Gitter, gibt für jeden Onset jedoch den Abstand bis zum nächsten Onset an, also die „Länge“ der rhythmischen Ereignisse.

Aus einer Reihe von möglichen Distanzmaßen zeigten drei Maße bei der Ähnlichkeitsberechnung von südamerikanischen Rhythmen gute Ergebnisse (Toussaint, 2004):

Die *Swap Distance* ist auf zwei Rhythmen in binärer Notation anwendbar und bestimmt die Anzahl der ungleichen rhythmischen Einheiten. Sie kann somit nur zwischen Rhythmen bestimmt werden, die eine Darstellung gleicher Länge aufweisen. Eine allgemeinere Variante für Darstellungen, die den Rhythmus nicht binär mit 1 und 0 darstellen, sondern z.B. die Onsets mit ihrer Intensität zwischen 0 und 1 beschreiben, ist die *Manhattan Distance*, welche die Differenzen zwischen allen Elementen gleicher Position aufsummiert. Für die binäre Darstellung liefern Swap Distance und Manhattan Distance die gleichen Abstände.

Ein Abstandsmaß für die Intervalllängendarstellung ist die *Chronotonic Distance*. Dazu wird die kompakte Intervalllängendarstellung zunächst wieder auf das Tatumgitter erweitert, indem jede Ziffer ihrem Wert entsprechend oft wiederholt wird. Für die Rhythmusdarstellung bedeutet dies, dass ein Onset, dessen „Dauer“ der Länge vom vier Tatums entspricht, dem also drei Einheiten Pause folgen, mit 4444 dargestellt

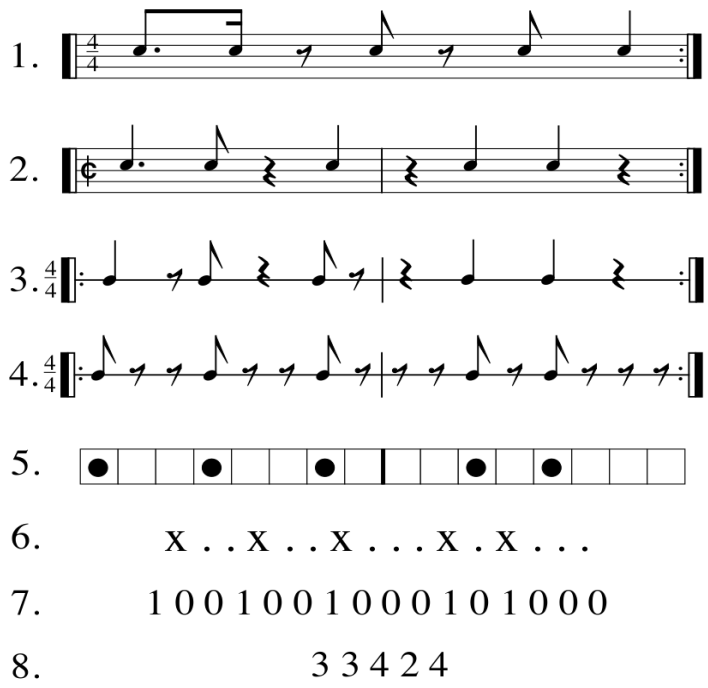

Abbildung 3.3.: Verschiedene Notationsformen des *Son-Clave* Rhythmus aus Toussaint (2004)

wird, ein Onset der Dauer von drei Schlägen mit 333. Für den Son-Clave Rhythmus aus Abbildung 3.3 ergibt sich also

```
3   3   4    2 4
3333334444224444
```

Von zwei Rhythmen dieser Darstellung wird nun wieder die Manhattan Distance berechnet.

Die *Edit Distance* ist ein Abstandsmaß, das in verschiedenen Experimenten eine gute Beschreibung wahrgenommener Ähnlichkeit zeigte (Mongeau und Sankoff, 1990). Gezählt wird hier die Anzahl der Veränderungen (*edits*), die nötig sind, um einen Rhythmus in den anderen zu überführen. Diese Veränderungen können das Ändern eines Elements sein, aber auch das Einfügen oder Löschen eines Elements. Der Vergleich von Rhythmen mit verschieden langen Darstellungen ist hierbei also möglich.

Cao et al. (2014) entwickelten die Theorie der *Rhythmusfamilien*, nach der zur Ähnlichkeitsbestimmung zwischen Rhythmen nicht immer das gesamte Pattern berücksichtigt werden muss. Grundlage ist wieder das rhythmische Gitter mit Einheiten des Tatums (oder einer anderen niedrigen Rhythmusebene), das jedoch auf eine höhere rhythmische Ebene um den Faktor zwei reduziert wird (eine ternäre Unterteilung des Pulses wird in der Arbeit nicht behandelt). Dabei kann jedes Element der höheren Ebene einen

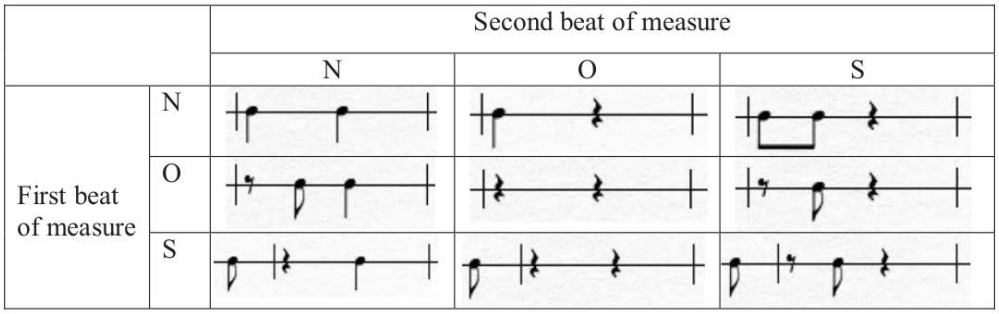

Abbildung 3.4.: Alle neun möglichen Rhythmusfamilien eines 2/4-Taktes aus Cao et al. (2014)

von drei Werten annehmen. Ein Onset (N, note) liegt vor, wenn auf der Zählzeit ein Onset erklingt, unabhängig von auf der niedrigeren rhythmischen Ebene benachbarten Zählzeiten. Eine Pause (0) ergibt sich, wenn die Zählzeit, sowie die auf niedrigerer Ebene vorherige Zählzeit keinen Onset aufweist. Ist die Zählzeit leer, doch der Schlag davor nicht, sprechen Cao et al. (2014) von einer Synkope (S). Die Konstruktion aller möglichen Rhythmusfamilien eines 2/4-Taktes zeigt Abbildung 3.4. Auf diese Rhythmusfamilien angewendete Abstandsmaße wie die Edit Distance oder (nach ersetzen der Symbole N, 0, S durch die Werte 1, 0, -1) die Manhattan Distance zeigen in Versuchen starke Korrelationen mit von Menschen bewerteten Abständen zwischen Rhythmen (Cao et al., 2014; Gómez-Marin et al., 2015).

Für die genannten Ansätze zum Rhythmusvergleich gilt, dass der Rhythmus in notierter Form vorliegen muss. Es werden nur die Tonanfänge berücksichtigt und jeder Ton wird nur aufgrund seiner rhythmischen Position gewichtet. Instrumentierung und somit Klang, Harmonie und Tonalität spielen keine Rolle. Dementsprechend basieren die Hörversuche zur Untersuchung der Qualität der Ähnlichkeitsbeschreibung auch jeweils nur auf Rhythmusbeispielen, die mit perkussiven Klickgeräuschen einstimmig und ohne dynamische Variation erzeugt wurden. Ein rhythmischer Vergleich zwischen Rhythmen in verschiedenen Taktarten ist insbesondere bei Beteiligung ternärer Taktarten problematisch. Je nach Abstandsmetrik ist ein Vergleich eventuell nicht möglich (Rhythmusfamilien) oder erfordert den größten gemeinsamen Teiler der beteilgten Tatums (Chronotonic Distance, Swap Distance).

3.2.2. Ähnlichkeit auf Basis des Audiomaterials

Wie Esparza et al. (2016) zusammenfasst, sind sich aktuelle Verfahren zur kompakten Beschreibung des Rhythmus von digitalem Audiomaterial im Grundsatz sehr ähnlich. Häufig wird eine rhythmussensitiv reduzierte Darstellung des Audiosignals (*Novelty Function*) abschnittsweise auf Periodizitäten untersucht, welche zu einer Gesamtdarstellung zusammengefasst werden. Da eine Tempoabhängigkeit der Rhythmusdarstellung einen Rhythmusvergleich in den meisten Fällen dominieren würde, findet im Allgemeinen ein Zwischenschritt statt, um eine tempoinvariante Form zu erhalten.

Verschiedene Verfahren zur Extraktion einer Novelty Function werden in Gouyon (2005) und Pampalk (2006) ausführlich diskutiert und verglichen. Die meisten aktuellen Ansätze verwenden jedoch einfach die positiven Änderungen der Energieverlaufskurve des Audiosignals oder die Lautheitskurve nach Zwicker und Fastl (2006), da beide sehr empfindlich auf Tonanfänge (onsets) reagieren, die als rhythmusdominierend angesehen werden (vgl. Gruhne et al., 2009; Holzapfel und Stylianou, 2011)). Diese wird meist in verschiedenen Frequenzbändern ermittelt und dann aufsummiert.

Die Beschreibung des Rhythmus eines Musikstückes oder -ausschnittes über das *beat histogram* ist sehr verbreitet, es zeigt die Intensität verschiedener Periodizitäten in der Musik, diese entsprechen meist den verschiedenen hierarchischen Rhythmusebenen. Die Novelty Function wird dazu in sich überlappende Blöcke zwischen 3 s (Tzanetakis und Cook, 2002) und 5 s (Gruhne et al., 2009) eingeteilt. Für jeden Block wird die Autokorrelationsfunktion berechnet. Über diejenigen Verschiebungen, die in den jeweiligen Blöcken die höchsten Ausschläge der Autokorrelationsfunktion ergeben, wird ein Histogramm erstellt. Manche Autoren schlagen Transformationen vor, um ein tempoinvariantes Beat Histogramm zu erhalten. Foote und Uchihashi (2001) normalisieren die Verschiebungen auf das Tempo des Musikstückes. J. H. Jensen et al. (2009) hingegen bevorzugen eine logarithmische Skala der Verschiebungen. Letztere behält zwar eine Tempoabhängigkeit, jedoch führen gleiche Rhythmen in verschiedenen Tempi zu gleichen Beat Histogrammen, die lediglich horizontal verschoben sind (J. H. Jensen et al., 2009). Um den Abstand oder die Ähnlichkeit zweier Beat Histogramme zu bestimmen, können sowohl Merkmale extrahiert werden, die das Beat Histogramm beschreiben (Tzanetakis und Cook, 2002; Foote und Uchihashi, 2001), oder Distanzmaße wie die Euklidische oder die Kosinusdistanz berechnet werden (Holzapfel und Stylianou, 2011; J. H. Jensen et al., 2009).

Dixon et al. (2004) unterteilen die Novelty Function in Abschnitte, die genau den Takten der Musik entsprechen. Eine Mittelung über alle Takte (mit gleichem Rhythmus) führt zu einem repräsentativen *Rhythmic Pattern*, das die Onsets und Gewichtungen

innerhalb eines Taktes anzeigt, zu sehen in Abbildung 3.5b. Die Rhythmic Patterns sind tempoinvariant, da sie auf Taktlängen normiert sind, erfordern jedoch die genaue Erkennung der Takte und Taktanfänge. Ist diese gegeben, so ist die Rhythmusbeschreibung über Rhythmic Patterns zusammen mit der Tempoinformation anderen Ansätzen überlegen, jedoch liefern auch die besten aktuellen Algorithmen noch keine ausreichend verlässliche Takterkennung (Esparza et al., 2016). Auch die Rhythmic Patterns werden über das Euklidische Distanzmaß verglichen.

Pampalk (2001) wählte für die Novelty Function einen multidimensionalen Ansatz. Anstatt die Energieverlaufskurven der verschiedenen Frequenzbänder zusammenzufassen, analysiert er die Lautheitskurven (nach Zwicker und Fastl (2006)) in 20 Frequenzbändern, deren Mittenfrequenzen gleiche Abstände auf der Barkskala aufweisen. Diese werden in Blöcke von 6 s unterteilt und jeweils per Fouriertransformation auf Periodizitäten (Fluktuationen) zwischen 0 Hz und 10 Hz untersucht. Über die Blöcke wird der Median berechnet. Das Ergebnis ist eine zweidimensionale Darstellung des Rhythmus mit der Fluktuationsintensität (farblich kodiert) in Abhängigkeit der Fluktuationsfrequenz (x-Achse) und der Frequenzbänder (y-Achse), die *Fluctuation Patterns* (Abb. 3.5a). Die Abstände liefert das euklidische Distanzmaß.

Pohle et al. (2009) erreichen durch geringfügige Änderungen eine nach Ihren Angaben verbesserte Darstellung der Fluctuation Patterns, die sie *Onset Patterns* nennen. Die Lautheitskurve wird auf Bereiche steigender Lautheit (onsets) beschränkt, statt der Barkskala werden Frequenzbandeinteilungen in Halbtonabständen gewählt und die Fluktuationen werden auf einer logarithmischen Skala untersucht (Abb. 3.5c). Letzteres führt wieder dazu, dass sich verschiedene Tempi nur durch eine horizontale Verschiebung bei jedoch gleicher Patternstruktur bemerkbar machen. Für Abstandmaße wird die euklidische Distanz oder eine parametrische Beschreibung über 16 Merkmale (*onset coefficients*) herangezogen.

3.3. Abgeleiteter Ansatz für den Rhythmusvergleich

Der in dieser Arbeit entwickelte und untersuchte Ansatz der Darstellung von Rhythmen kombiniert Elemente der vorgestellten bekannten Verfahren, das Ergebnis ist das *Bar Pattern*. Ausgehend vom hierarchischen Strukturmodell (Todd, 1994), der in der westlichen Musik dominierenden Notation und Strukturierung der Musik und den guten Ergebnissen der Rhythmic Patterns bilden Takte die Basis für die Beschreibung des Rhythmus der Musikstücke. Als Takt wird hierbei diejenige rhythmische Struktur (aus dem hierarchischen Strukturmodell) bezeichnet, die am besten in die empirische

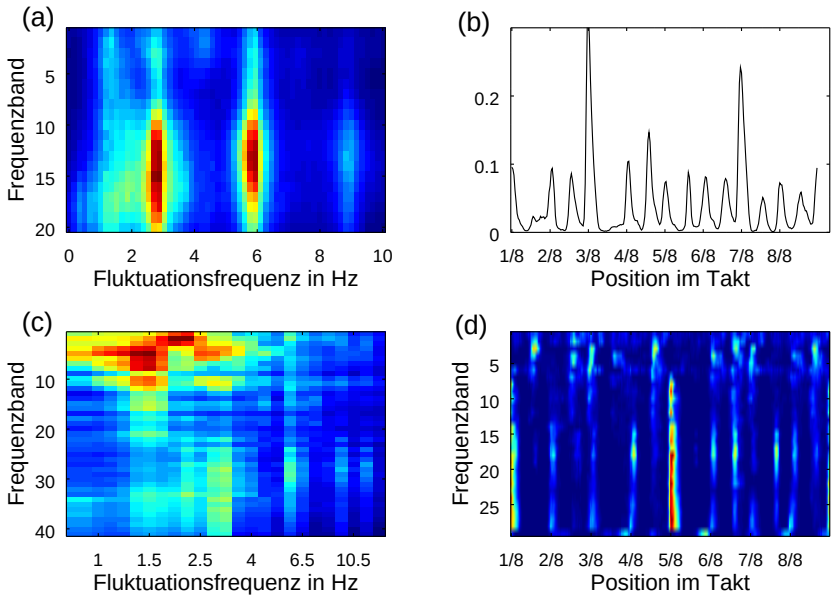

Abbildung 3.5.: Fluctuation Pattern (a), Rhythmic Pattern (b), Onset Pattern (c) und Bar Pattern (d) des Songs „Tippie-Toes“ von The Meters im direkten Vergleich. Das Rhythmic Pattern ist manuell auf einen Taktanfang ausgerichtet.

Verteilung der Taktlängen passt (vgl. Kapitel 3.1.2), also die stärkste Struktur nahe der Länge von 2 Sekunden. In einigen Fällen wird diese Taktbestimmung möglicherweise um den Faktor 2 oder 0,5 von der Notation der Komposition abweichen. Diese Takte liegen innerhalb der psychischen Präsenszeit und können als Ganzes erfasst werden, es sind also Art und Reihenfolge der rhythmischen Ereignisse innerhalb eines Taktes relevant.

Als rhythmische Ereignisse werden zunächst zeitliche Änderungen aller grundlegenden akustischen Wahrnehmungsparameter (vgl. Kapitel 2.2.1 bis 2.2.4) abgesehen von der räumlichen Lokalisation betrachtet. Änderungen der Lautheit werden über die Energieverlaufskurve erfasst, im Unterschied zu den Rhythmic Patterns und in Anlehnung an die Fluctuation Patterns in mehreren Frequenzbändern entlang der Barkskala. Änderungen der Tonhöhenwahrnehmung zeigt eine Chroma Differenz Funktion an. Die Änderungen der Klangfarbe spielen bei den hier betrachteten kurzen Musikausschnitten keine Rolle und werden nicht weiter berücksichtigt. Die mehrdimensionale Novelty Function, die in Takte eingeteilt wird, besteht also aus dem Energieverlauf in mehreren Frequenzbändern sowie der CDF.

Für die Darstellung des Gesamtrhythmus des Musikstückes wird der Median über

die taktweise Darstellung der Novelty Function gebildet. So etablieren sich diejenigen rhythmischen Ereignisse, die wiederholt an der gleichen Position innerhalb eines Taktes auftreten und das rhythmische Gerüst bilden.

Um die Schwierigkeit der Takterkennung zu reduzieren, werden die Bar Patterns als Basis für den Rhythmusvergleich in zwei verschiedenen Versionen untersucht. Als Referenz dienen Bar Patterns auf der Grundlage von manueller Annotation der Takte mit korrekter Taktlänge und korrekter Ausrichtung (die erste Zählzeit des Taktes ist am Anfang des Patterns). Als zweite Variante werden die Bar Patterns mit automatisch extrahierter Taktlänge, jedoch ohne korrekte Ausrichtung auf den Taktanfang verwendet. Der Unterschied zu der annotierten Version besteht hierbei lediglich in einer zyklischen Verschiebung entlang der x-Achse.

Die klassischen Abstandsmaße, die sich durch Übereinanderlegen der Bar Patterns ergeben würden, können im Fall der zyklischen Verschiebung natürlich nicht angewendet werden. Als Ähnlichkeitsmaß wird hier ein Abstandsmaß auf Basis einer Kreuzkorrelation verwendet.

4. Algorithmus

Neben den bekannten und in Kapitel 3.2.2 vorgestellten Verfahren zum Rhythmusvergleich werden in dieser Arbeit zwei Ansätze besonders betrachtet. Zum Einen wird der Vergleich zweier Rhythmusdarstellungen nach den Überlegungen aus Kapitel 3.3, dem Bar Pattern, untersucht. Die Extraktion der Bar Patterns ist im folgenden Abschnitt 4.1 beschrieben. Zum Anderen wird in Abschnitt 4.2 das Prinzip der Rhythmusfamilien auf komplexere Audiosignale erweitert, um auf dieser Basis die Rhythmen von Musik vergleichen zu können.

4.1. Rhythmusdarstellung durch Bar Patterns

Die verschiedenen Schritte, die vorgenommen werden, um das Bar Pattern nach den Ideen aus Kapitel 3.3 aus einem Musiksignal zu extrahieren, sind in einem Flussdiagramm in Abbildung 4.1 schematisch dargestellt.

4.1.1. Vorverarbeitung

Liegt das Audiosignal als Stereosignal vor, so wird es zunächst durch Mittelung über die Kanäle in ein Monosignal umgewandelt. Eine Kurzzeitfouriertransformation mit einer Blocklänge von 46,4 ms (2048 samples bei einer Abtastrate von 44,1 kHz) und einer Sprunglänge von 11,6 ms (512 Samples bei einer Abtastrate von 44,1 kHz) führt zu einem Spektrogramm, das durch sich überlappende Dreiecksfilter entlang der Barkskala zu Frequenzbereichen zusammengefasst wird. Die eigentliche Aufteilung der Barkskala in 24 Frequenzbänder wurde zugunsten einer feineren Auflösung im Bassbereich aufgegeben. Die Barkskala wird ab einer Frequenz von 400 Hz mit 18 Frequenzbändern bis 11.000 Hz bedient, darunter wurden 8 Frequenzbänder entlang der Mel Skala zwischen 20 Hz und 400 Hz gewählt. Die Energieverlaufskurve der 26 Frequenzbänder werden um die Chroma Differenz Funktion erweitert, zusammen bilden sie so die 27-dimensionale Novelty Function als Basis der folgenden Schritte.

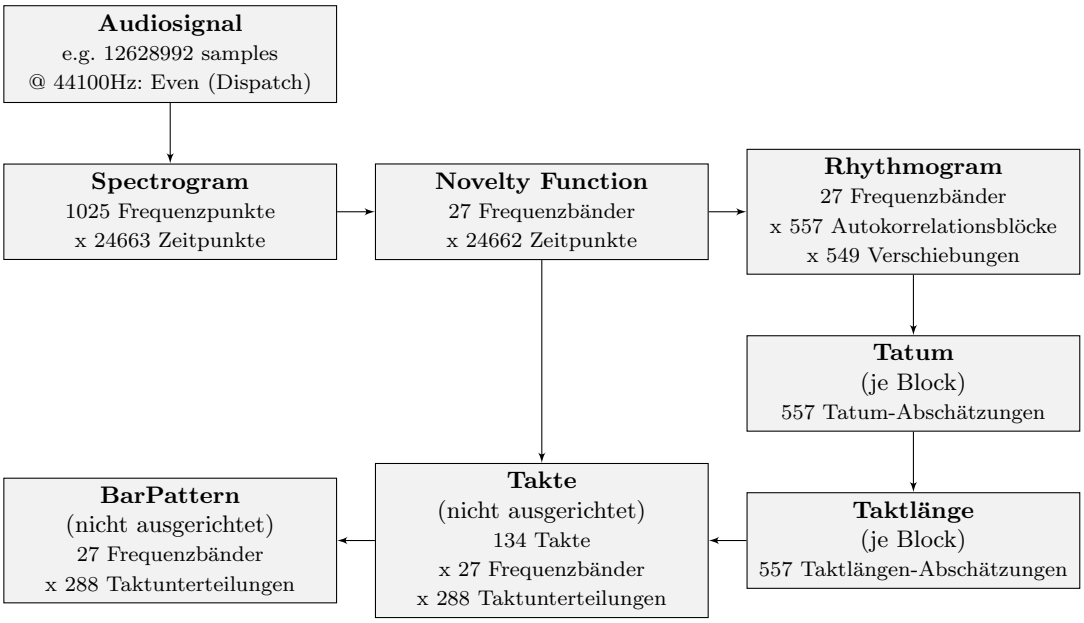

Abbildung 4.1.: Flussdiagramm mit den einzelnen Zwischenstufen während der Extraktion des Bar Pattern. Die Größenangaben zeigen den Datenumfang jedes Zwischenschrittes am Beispiel des Songs „Even“ von Dispatch.

4.1.2. Rhythmogramm

Der erste Analyseschritt ist die Berechnung des Rhythmogramms für jede Dimension der Novelty Function, also für jedes Frequenzband und die CDF. Da für das Rhythmogramm Periodizitäten im Bereich des Tatums, des Tempos und der Taktlängen interessant sind, wird eine Blocklänge von etwa 8,5 Sekunden gewählt (732 Samples der Novelty Function). So liegen mehrere Takte in einem Block, dennoch findet die Analyse in einem kurzen lokalen Bereich statt. Um kleine Änderungen im Tempo schnell detektieren zu können wird eine kleine Sprunglänge von etwa 0,5 Sekunden (44 Samples der Novelty Function) gewählt. Da die Autokorrelationsfunktion bei größer werdenden Verschiebungen immer weniger Signalanteile miteinander vergleicht, wird das Ergebnis in diesem Bereich auch mit der in Kapitel 2.3.5 beschriebenen Korrektur ungenau. Daher werden für das Rhythmogramm nur Verschiebungen bis 6 Sekunden berücksichtigt. Die 27 auf diese Weise gewonnenen Rhythmogramme werden gemittelt, um die folgende Analyse mit einem einzigen Rhythmogramm durchführen zu können, zudem wird jede Zeile des Rhythmogramms mittelwertbefreit. Ein Beispiel ist dargestellt in Abbildung 4.2.

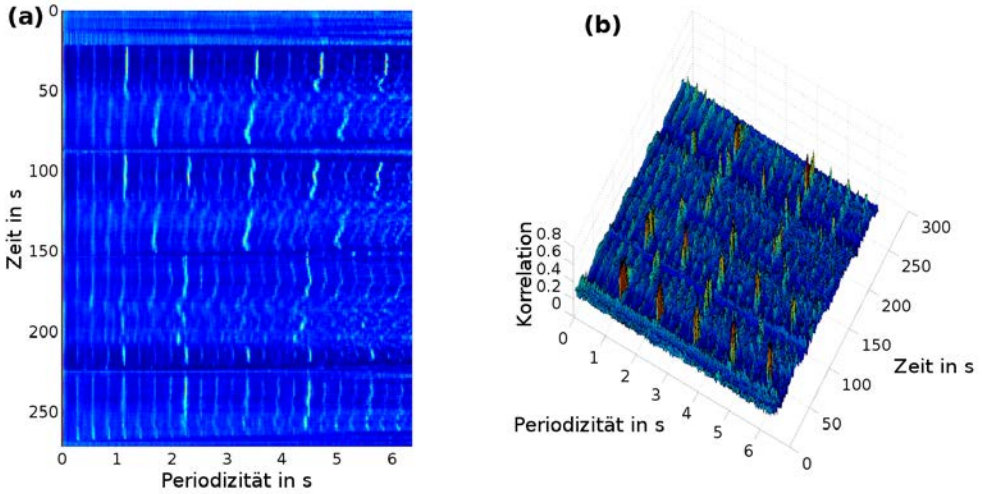

Abbildung 4.2.: 2D (a) und 3D (b) Darstellung des Rhythmogramms des Songs „Even“ von Dispatch. Rötlichere Farben stehen für höhere Werte der Autokorrelation.

4.1.3. Tatum

Das Rhythmogramm zeigt Periodizitäten im Audiosignal zwischen ca. 25 ms (Verschiebung um 2 Samples der Novelty Function) und 6 Sekunden. Ein sich gleichmäßig wiederholendes (akustisches) Ereignis im Quellsignal wird in der Autokorrelationsfunktion einen hohen Wert bei einer Verschiebung um die zeitliche Differenz zweier dieser Ereignisse hervorrufen, jedoch auch bei allen Vielfachen dieser Verschiebung (wenn das erste Ereignis mit dem dritten, vierten und weiteren Ereignissen zusammenfällt). Es bietet sich zur genauen Untersuchung des Tempos also eine Analyse einer kleinen rhythmischen Einheit, z.B. des Tatums, an, da zu dessen genauer Bestimmung eine Vielzahl an Peaks der Autokorrelation herangezogen werden können.

Die genaue Definition sieht das Tatum als kleinste in einem Musikausschnitt vorkommende rhythmische Einheit. Bei manchen Musikstücken ist diese Zuordnung nicht ganz eindeutig. Um etwaigen Interpretationsspielraum zu umgehen, wird mit dem Tatum im Folgenden die stärkste rhythmische Einheit zwischen 360 bpm (6 Hz) und 750 bpm (12,5 Hz) bezeichnet. Dieser Bereich deckt etwas mehr als eine Oktave (um den Faktor zwei auseinanderliegend) ab, so dass jedes Tatum oder ein ganzzahliger Teiler davon erfasst werden kann. Zudem liegen die Teiler gängiger Tempi um 120 bpm nicht in den Randbereichen. In einigen Fällen wird sich das tatsächliche Tatum von diesem Tatum um einen kleinen ganzzahligen Faktor unterscheiden. Dennoch ist das hier ermittelte Tatum ein gemeinsamer Teiler aller weiteren hierarchischen Rhythmusebenen.

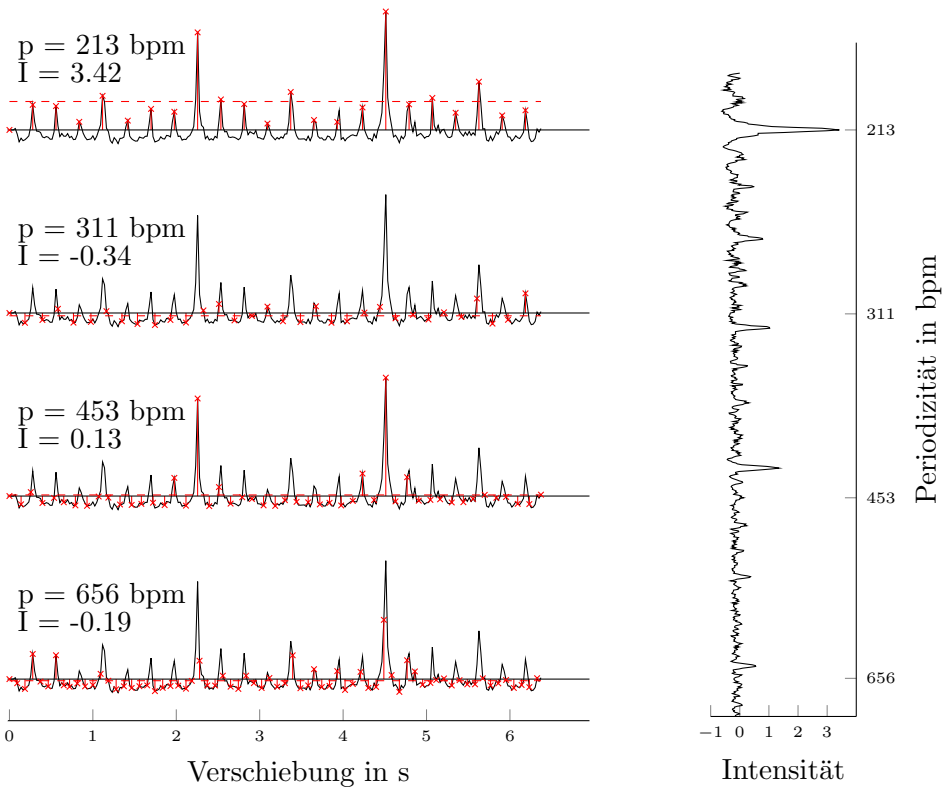

Abbildung 4.3.: Bestimmung der Tatumsintensität (I) für einen Block des Rhythmogramms, dargestellt für vier mögliche Tatums (Periodizitäten, p). Die Vielfachen des jeweiligen Tatums sind rot markiert, über die Werte an diesen Stellen wird gemittelt. Die sich ergebende Intensität in Abhängigkeit der Periodizität (rechts, auch für die anderen, nicht weiter abgebildeten Periodizitäten) ist nur in dieser Darstellung vertikal angeordnet, in den Darstellungen in Abbildung 4.4 (a) und (b) entspricht dies einer Zeile.

Zunächst wird für jeden Block der Kurzzeitautokorrelation (also für jede Zeile des Rhythmogramms) die „Intensität“ jedes möglichen Tatums in 1 bpm-Schritten von 190 bpm (entspricht in der Autokorrelation einer Verschiebung um 27,2 Samples) und 750 bpm (6,9 Samples) ermittelt, das ergibt 560 Werte. Dazu wird für jedes mögliche Tatum der Mittelwert der Autokorrelationsfunktion an den Stellen der Verschiebung und allen seinen Vielfachen ermittelt[1]. Nicht ganzzahlige Verschiebungen werden gerundet, durch Berücksichtigung aller Vielfachen wird eine genügend hohe Auflösung erreicht. Abbildung 4.3 zeigt dieses Vorgehen für einen Block und vier zu überprüfende Tatums beispielhaft.

[1]Dies entspricht pro Tatum einer Abtastung des Autokorrelationsblocks mit einer der Frequenz des Tatums entspechenden Abtastrate mit anschließender Mittelung.

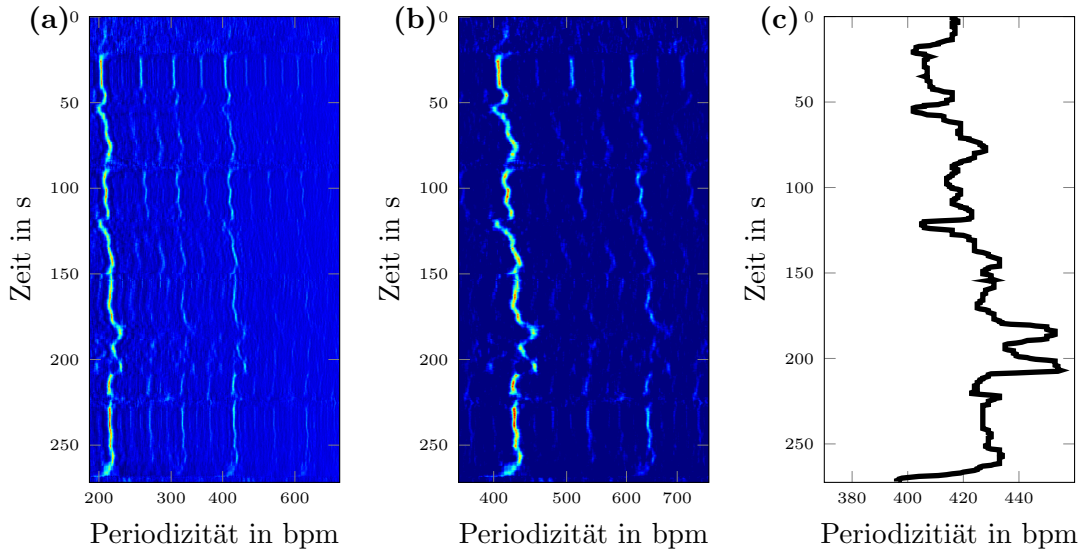

Abbildung 4.4.: Tatumselektion: Rohdaten der Intensität möglicher Tatums (a), nachbearbeitete Intensität (b) und ausgewähltes Tatum (c).

Die so gewonnene Tatumsintensität ist in Abbildung 4.4(a) dargestellt. In einem Zwischenschritt werden nun Periodizitäten, die sich um kleine ganzzahlige Faktoren f unterscheiden, gewichtet (mit dem Gewicht w) zusammengefasst, das Ergebnis zeigt Abbildung 4.4(b). Mit der Definition der Autokorrelation aus Kapitel 2.3.5 gilt für die rhythmische Intensität I in Abhängigkeit der Periodizität p in Samples:

$$I(p) = \sum_i \left(w_i \cdot \frac{1}{N_i} \cdot \sum_{k=1}^{N_i} r_{xx}(f_i \cdot k \cdot p) \right)$$

Auf dem GTZAN-Datensatz führen die Faktoren $f_i = \{0{,}25,\ 0{,}33,\ 0{,}5,\ 1,\ 2,\ 3,\ 4\}$ mit den dazu gehörenden Gewichten $w_i = \{0{,}8,\ 0{,}5,\ 0{,}8,\ 1,\ 0{,}8,\ 0{,}5,\ 0{,}8\}$ zu den besten Ergebnissen.

Dieser Verlauf der Tatumsintensität kann als *hidden markov model* aufgefasst werden, bei dem für jeden Block die Emissionen (Tatumsintensität) der nicht bekannten Zustände (Tatums) beobachtet werden können. Für die Auswahl des expliziten Tatums einer jeden Zeile wird also der Viterbi-Algorithmus (vgl. Kapitel 2.3.7) eingesetzt. Die Wahrscheinlichkeit eines jeden Zustandes ist durch die Tatumsintensität gegeben, diese wird so normiert, dass die Summe gleich eins ist. Die Übergangswahrscheinlichkeit zwischen zwei Tatums sinkt quadratisch mit der Differenz der Tatums in bpm.

Der optimale Pfad durch den Verlauf der Tatumsintensität liefert ein Tatum für jeden Block. Auf diese Weise kann Tempoänderungen gefolgt werden, wenn die Tatumsinten-

sität das impliziert, ein Tempo jedoch auch beibehalten werden, wenn die Änderungen in der Intensitätsfunktion nur kurz vorliegen. Abbildung 4.4(c) zeigt den im obigen Beispiel ausgewählten Tatumsverlauf. Es handelt sich auch hier um den Song „Even“ von Dispatch in einer Live-Version. Gut sind die Tempoänderungen zu erkennen, die an den Wechseln von den Strophen in die Refrains auftreten (nach etwa 50 Sekunden und nach etwa 120 Sekunden), sowie die langsame Tempozunahme über das gesamte Stück. Zu beachten ist, dass die Periodizität in den Abbildungen 4.3 und 4.4 in Schlägen pro Minute (bpm) angegeben ist, bei der Darstellung des Rhythmogramms allerdings in Sekunden. Die beobachtete Tempozunahme zeigt sich in beiden Abbildungen durch einen Verlauf in entgegengesetzte Richtung.

Die Leistungsfähigkeit dieser Bestimmung des Tatums wurde mit dem GTZAN-Datensatz überprüft, als korrekt wird ein ermitteltes Tatum dann gewertet, wenn das Verhältnis des annotierten Tempos zum mittleren Tatum ein ganzzahliges ist (+/- 0,0625). Die Ergebnisse der Auswertung sind in Tabelle 4.1 aufgeführt, bei rhythmisch dominierten Musikrichtungen liegt die korrekte Abschätzung bei deutlich über 90 %, zumeist bei mindestens 95 %.

4.1.4. Taktlänge

Die Taktlänge ist nach dem hierarchischen Strukturmodell ein ganzzahliges Vielfaches des Tatums. Tempoänderungen und -variationen werden durch das Tatum erfasst, eine Änderung des Verhältnisses Taktlänge zu Tatum tritt nur bei Taktänderungen auf, ansonsten ist dieses Verhältnis konstant. Dieser Idee folgend wird die Taktlänge hier in Vielfachen des Tatums ermittelt.

Dazu wird das Rhythmogramm zeilenweise (jeder Block der Autokorrelation) auf die Vielfachen des jeweiligen Tatums reduziert, um Rundungsfehler und Ungenauigkeiten zu kompensieren wird der Mittelwert der drei dem exakten Wert am nächsten liegenden Verschiebungen bestimmt. Das entstehende reduzierte Rhythmogramm wird mit einem Gewichtungsfaktor für die jeweiligen Taktlängen multipliziert, der der Verteilung der Taktlängen nachempfunden ist (Abbildung 3.2 in Kapitel 3.1.2). Das Ergebnis ist in Abbildung 4.5(b) zu sehen. Für die Auswahl der Taktlängen eines jeden Blockes wird wieder der Viterbi-Algorithmus herangezogen. Im Unterschied zur Bestimmung des Tatums sind die Übergangswahrscheinlichkeiten zwischen zwei Taktlängen nicht von der Sprungweite abhängig, sondern für von Null verschiedene Werte konstant. Ein Sprung entspricht hier einem Taktwechsel, ein Wechsel der Taktlänge von beispielsweise 16 Tatums auf 15 Tatums sollte einem Wechsel auf 12 Tatums keineswegs vorgezogen werden.

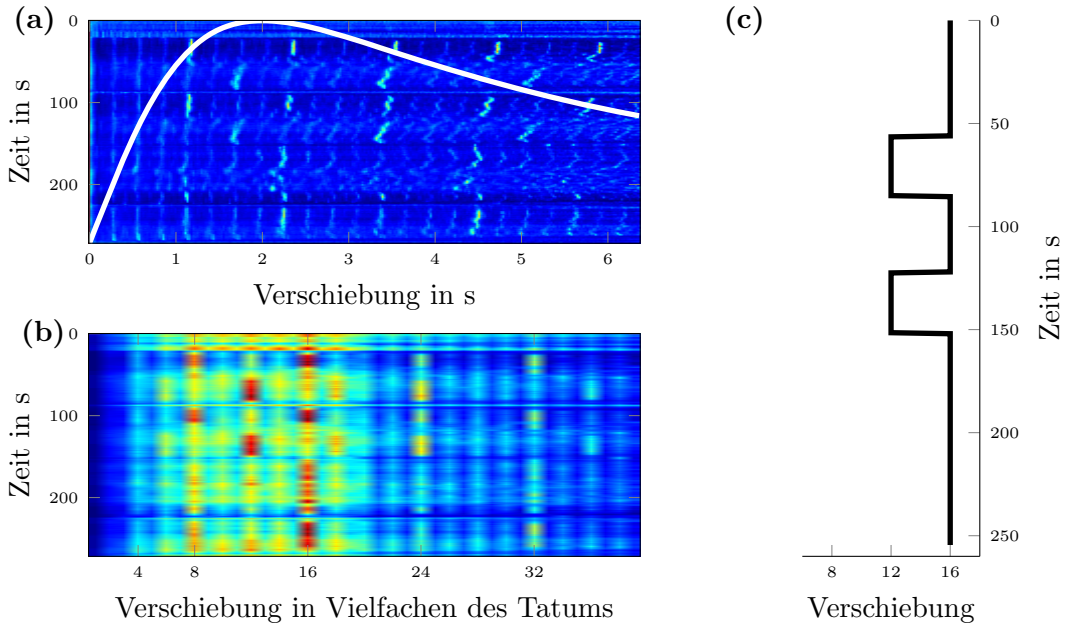

Abbildung 4.5.: Taktlängenabschätzung: (a) Rhythmogram mit Gewichtungsfaktor für Taktlänge, (b) reduziertes Rhythmogram in Vielfachen des Tatums, (c) ausgewählte Taktlänge in Vielfachen des Tatums.

Das Ergebnis ist eine Taktlängenabschätzung innerhalb des Musikstückes in Schritten der Rhythmogramm-Schrittweite, also alle 0,5 Sekunden in Vielfachen des Tatums, bei Multiplikation mit dem jeweiligen Tatum in Sekunden, dargestellt in Abbildung 4.5, wieder für den bereits bekannten Song von Dispatch. Der Wechsel von einem 4/4-Takt zu einem 3/4-Takt in beiden Refrains wird korrekt erfasst.

Eine Analyse dieser Taktlängenerkennung wurde wiederum mit dem GTZAN Datensatz durchgeführt. Dazu wurde die abgeschätzte Taktlänge an jedem annotierten Taktanfang mit der annotierten Taktlänge verglichen. Als korrekt wird eine Taktlänge dann angesehen, wenn die Abweichung weniger als 1/16 der annotierten Taktlänge beträgt. In Tabelle 4.1 sind, zusammen mit den korrekt erkannten Tatums, der Anteil der Musikausschnitte aufgeführt, bei denen tatsächlich alle erkannten Taktlängen korrekt sind. In manchen Fällen konnte beobachtet werden, dass Änderungen der Taktlänge etwas spät erfasst wurden, bei starken Änderungen kam es so an der Position des annotierten Taktanfangs noch zu einer Fehleinschätzung der Taktlänge. In der Folge gibt es einige Beispielsongs, in denen lediglich einzelne Taktlängen nicht korrekt bestimmt wurden. Für den hier durchgeführten Rhythmusvergleich war diese Ungenauigkeit nicht relevant und wurde nicht weiter untersucht, in der dritten Spalte der Tabelle 4.1 ist daher aber zusätzlich der Gesamtanteil an korrekt erkannten Taktlängen über alle Songs der

Genre	Tatum korrekt	Songs mit 100% korrekt erkannten Taktlängen	korrekte Taktlängen total (in %)
Blues	90	83	91
Klassik	58	34	59
Country	96	95	98
Disco	99	96	99
Hiphop	98	100	100
Jazz	82	66	89
Metal	90	86	96
Pop	97	94	98
Reggea	99	96	99
Rock	94	92	97

Tabelle 4.1.: Auswertung der Tatums- und Taktlängenerkennung auf dem GTZAN Datensatz, aufgeteilt nach Genres. Jedes Genre enthält 100 Songs. Die Abweichung im Genre Hiphop und Reggea (korrekte Taktlänge bei inkorrektem Tatum) rührt von einer ungenauen Annotation des Tempos einiger Stücke her, tatsächlich sind in beiden Genres alle ermittelten Tatums korrekt.

jeweiligen Genres aufgeführt. Da Fehleinschätzungen um den Faktor 2 bei solcher Art der Abschätzung von Periodizitäten häufig auftreten (vgl. Gouyon et al., 2006), wurden auch Taktlängen, die Abweichungen um den Faktor 2 oder 0,5 zeigten, als korrekt eingestuft.

4.1.5. Rhythmusdarstellung

Liegt die Abschätzung der Taktlängen vor, so kann die Novelty Function in Takte eingeteilt werden, die sich weder überlappen, noch Lücken zwischenenander aufweisen. Jeder Takt beginnt also direkt im Anschluss an den vorherigen Takt, als jeweilige Taktlänge wird die abgeschätzte Taktlänge an der Taktanfangsposition gewählt. Der erste Takt beginnt direkt am Anfang des Musikausschnittes.

Da bei verschiedenen Taktlängen die entstehenden Takte eine unterschiedliche Anzahl an Samples (der Energieverlaufskurve) aufweisen, werden sie auf eine einheitliche Länge resampled. Hier wurde eine Länge von 288 Samples gewählt ($2^5 \cdot 3^2$), da so sowohl eine fünffache Halbierung (32tel Noten) als auch eine zweifache Drittelung (9er Takt) exakt darstellbar ist. Um zudem geringfügige Variationen zu kompensieren wird im Anschluss eine leichte Glättung durch Faltung mit einem Gaussfenster der Länge 9 Samples durchgeführt.

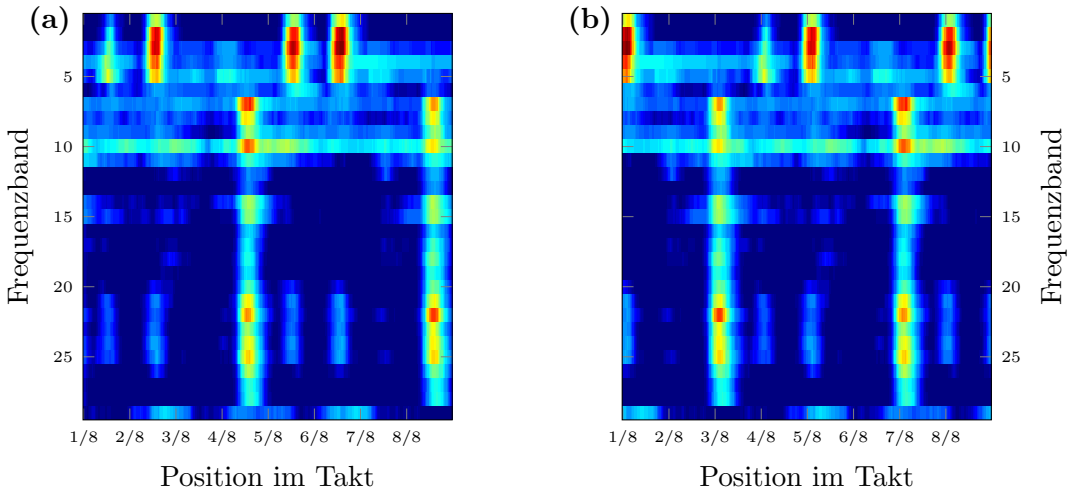

Abbildung 4.6.: Bar Pattern der Strophen aus dem Song „Even“ von Dispatch. Extrahierte Version (a) und manuell auf den Taktanfang ausgerichtet (b). Die unterste Zeile besteht jeweils aus der CDF.

Diese Takte sind im Allgemeinen nicht auf einen Downbeat ausgerichtet, der Taktanfang entspricht also nicht einem musikalischen Taktanfang. Bei korrekt erkannter Taktlänge sind die einzelnen Takte jedoch alle um den gleichen Betrag verschoben.

Um die rhythmisch repräsentative Darstellung eines Musikausschnittes, das *Bar Pattern*, zu erhalten, wird der Median über alle Takte berechnet[2]. So werden akustische Ereignisse erfasst, die sich in mehreren Takten in den gleichen Frequenzbändern und an der gleiche Position innerhalb der Takte befinden, also genau die rhythmisch relevanten Ereignisse. Auch das Bar Pattern ist nur dann auf einen Taktanfang ausgerichtet, wenn der Beginn des Musikausschnittes genau mit einem Taktanfang zusammenfällt, im Allgemeinen also (zirkular) verschoben. Abbildung 4.6 zeigt das Bar Pattern der rhythmisch konstanten Strophen des Songs „Even“ in der extrahierten Version (a), sowie manuell auf den Taktanfang ausgerichtet (b).

4.1.6. Vergleich von Bar Patterns

Werden in einem weiteren Schritt die Bar Patterns korrekt auf die Taktanfänge ausgerichtet, so können zwei Bar Patterns mit klassischen Distanzmaßen wie der Euklidischen Distanz, der Manhattan Distanz oder dem Mittelwert der zeilenweisen Korrelationskoeffizienten verglichen werden. Alle drei Methoden liefern ähnliche Ergebnisse, für die

[2] Bei längeren Musikausschnitten muss eventuell eine Segmentierung in rhythmisch ähnliche Bereiche vorgenommen werden, dieser Aspekt wird hier nicht behandelt.

Auswertung wurde daher lediglich der Korrelationskoeffizient verwendet.

Ohne gleiche Ausrichtung der Bar Patterns liefern die genannten Distanzmaße jedoch Werte, die in erster Linie von der Verschiebung der Bar Patterns zueinander abhängig sind. Um dennoch auf den noch fehleranfälligen Schritt der Downbeaterkennung zu verzichten kann ein Abstandmaß über die Kreuzkorrelation zweier nicht ausgerichteter Bar Patterns ermittelt werden. Dazu wird die zeilenweise Kreuzkorrelation bestimmt. Die Verschiebung der Bar Patterns gegeneinander muss dabei zirkular geschehen, damit bei jeder Verschiebung 288 Samples übereinander liegen. Dies wird gelöst, indem ein Bar Pattern horizontal verdoppelt wird und bei der Kreuzkorrelation nur die mittleren 288 Verschiebungen betrachtet werden. Ausgewählt wird dann diejenige Verschiebung, die im Mittel über alle Frequenzbänder den höchsten Wert ergibt. Dieser Mittelwert ist gleichzeitig das Maß für die Ähnlichkeit der Bar Patterns. Der Vorgang ist in Abbildung 4.7 dargestellt. Das Bar Pattern des Songs „Even“ (a) wird mit dem verdoppelten Bar Pattern des Songs „Tippi Toes“ (b) verglichen, (c) zeigt das Ergebnis der zeilenweisen Kreuzkorrelation. Der Mittelwert über alle Frequenzbänder (und CDF) sowie das markierte Maximum sind in (d) zu sehen.

Im Vergleich zum Korrelationskoeffizienten der korrekt ausgerichteten Bar Patterns wird durch die Kreuzkorrelation die Distanz möglicherweise unterschätzt (die Ähnlichkeit überschätzt). Der Wert der Korrelation bei korrekter Ausrichtung ist einer der 288 möglichen Verschiebungen der Kreuzkorrelation, er wird nur dann nicht ausgewählt, wenn bei einer anderen Verschiebung eine höhere Korrelation auftritt.

Die automatisch extrahierte Taktlänge unterscheidet sich gelegentlich um den Faktor 2 von der tatsächlichen Taktlänge. Beim Vergleich zweier Musikausschnitte kann dieser Fehler bei keinem, bei einem oder bei beiden Beispielen vorliegen, in letzterem Fall möglicherweise in entgegengesetzte Richtung. Dieser mögliche Fehler wird mit einigem Aufwand kompensiert. Für jedes Stück werden zwei Bar Patterns bestimmt, eines mit der ausgewählten Taktlänge, das zweite mit einer alternativen Taktlänge. Diese ist doppelt so groß bei ausgewählten Taktlängen von maximal 2,3 Sekunden und halb so groß bei darüber liegenden Taktlängen (so wird die Verteilung der Taktlängenhäufigkeit auch bei der alternativen Taktlänge gut bedient). Die Ähnlichkeit wird dann für alle vier Kombinationen der Bar Patterns bestimmt und die ähnlichste ausgewählt.

Da bei beiden Abstandmaßen (Korrelation der ausgerichteten Bar Patterns sowie Kreuzkorrelation der nicht ausgerichteten Patterns) für jede Dimension der Novelty Function, also für jedes Frequenzband und für die CDF ein einzelner (Kreuz-)Korrelationskoeffizient bestimmt wird, kann später eine Gewichtung der einzelnen Frequenzbänder vorgenommen werden. Durch die Anpassung der Gewichte an den Datensatz können dieje-

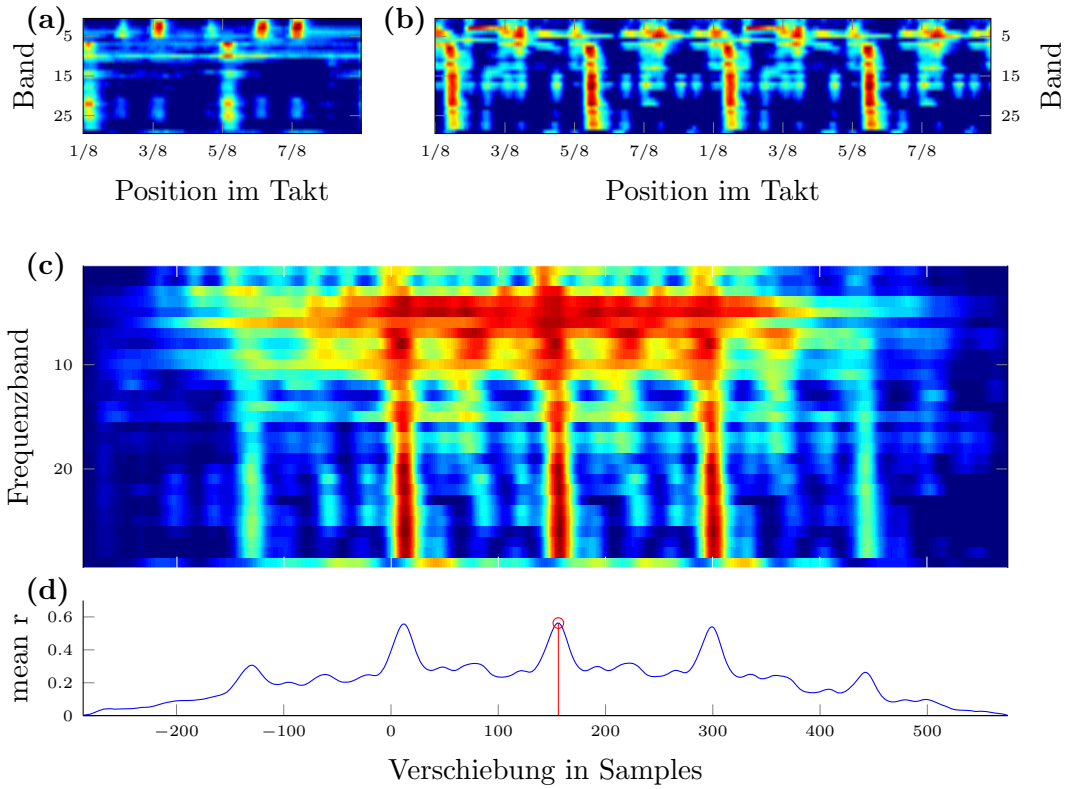

Abbildung 4.7.: Kreuzkorrelation zweier Bar Patterns: Einfaches Bar Pattern (a) und verdoppeltes Bar Pattern (b), Ergebnis der Kreuzkorrelation (c) und Mittelwert der Kreuzkorrelation (d). Die Markierung zeigt das Maximum der Korrelation, der Korrelationswert an dieser Stelle wird als Ähnlichkeit zweier Bar Patterns betrachtet.

nigen Frequenzbänder identifiziert werden, die für die Rhyrhmusähnlichkeit besonders dominant sind. Für diesen Fall werden zusätzlich zu den „direkten“ Ähnlichkeiten (gleiche Gewichtung aller Dimensionen) die Ähnlichkeiten der einzelnen Frequenzbänder als 27-dimensionaler Vektor extrahiert.

Die Korrelation beschreibt eine Ähnlichkeit, der Korrelationskoeffizient r kann Werte zwischen 0 und 1 annehmen. Ein Distanzmaß ergibt sich in diesem Fall aus $1 - r$.

4.2. Rhythmusfamilien

Neben den Bar Patterns liegt in dieser Arbeit ein besonderes Augenmerk auf der Theorie der Rhythmusfamilien, die in einfachen Experimenten bereits eine Vorhersage der

Rhythmusähnlichkeit ermöglichte. Die Zuordnung von Rhythmen in Rhythmusfamilien und deren Vergleich nach Cao et al. (2014) (vgl. Kapitel 3.2.1) basiert auf Rhythmen aus gleichen Klängen ohne Lautstärkeabstufungen. Ein Versuch der Erweiterung des Modells auf Musiksignale wird in diesem Abschnitt vorgenommen. Dabei werden die Rhythmusfamilien auf Basis einzelner Instrumente oder in Frequenzbändern bestimmmt. Für die Auswertung beider Verfahren wurde jeweils auf manuelle Annotation zurückgegriffen, um Fehler in der Vorverarbeitung auszuschließen.

4.2.1. Zeitlich Auflösung

Die Rhythmusfamilien können auf verschiedenen rhythmischen Einheiten basieren. Da immer zwei Grundschläge gemeinsam betrachtet werden, um eine Einteilung in die drei Kategorien Onset, Synkope oder Pause vorzunehmen, wird die Anzahl der Schläge in der Darstellung der Rhythmusfamilien halbiert. Hier werden die drei Unterteilungen in 16tel Noten mit 8 Werten ($3^8 = 6561$ mögliche Rhythmusfamilien), 8tel Noten (81 Rhythmusfamilien) und 4tel Noten mit lediglich 2 Werten (9 Rhythmusfamilien) untersucht.

4.2.2. Rhythmus der Instrumente

Die naheliegendste Methode der Modellerweiterung ist es, die Rhythmusfamilien für die einzelnen Instrumente zu bestimmen und diese dann zu kombinieren. Hier wurden lediglich die Tonanfänge der Instrumente berücksichtigt, die Tonlängen spielten keine Rolle. Der vorliegende (annotierte) Takt wurde dazu Instrumentenweise in die gewünschten Notenwerte aufgeteilt. Lag innerhalb dieses Bereichs ein Tonanfang, so wurde diese Zählzeit als Zählzeit mit Onset gewertet und die Rhythmusfamilien nach dem bekannten Verfahren ermittelt. Für den Vergleich solcher mehrdimensionaler Rhythmusfamilien müssen jedoch Musikausschnitte mit den gleichen Instrumenten vorliegen, oder es muss eine Zuordnung getroffen werden, welche Instrumente miteinander verglichen werden. Im hier verwendeten Datensatz ist die Instrumentengleichheit gegeben.

4.2.3. Rhythmus in Frequenzbändern

Liegt keine Annotation der Tonanfänge der einzelnen beteiligten Instrumente vor, so kann das Bar Pattern als Basis für die Rhythmusfamilien fungieren, eine korrekte Taktanfangsausrichtung vorausgesetzt. Dazu wird die positive Energieänderung in den Bereichen, die zu den jeweiligen Notenwertunterteilungen gehören, aufsummiert. Da die

Energiewerte der einzelnen Notenwerte nun kontinuierlich sind, muss die Unterscheidung zwsichen Onset, Stille und Synkope angepasst werden. Als Grenzwert zur Unterscheidung zwischen Onset und Stille wurden mit dem 0.9-fachen des Mittelwertes in jedem Frequenzband die besten Ergebnisse erzielt. Eine Synkope liegt dann vor, wenn die Summe der Energieänderung innerhalb des Notenwertes einer leichteren Zählzeit ebenfalls um den Faktor 0.9 höher ist, als die in der darauffolgenden schwereren Zählzeit. So kann für jedes Frequenzband eine Rhythmusfamilie ermittelt werden, der Vergleich zweier solcher mehrdimensionalen Rhythmusfamilien erfordert nicht mehr eine Zuordnung der Instrumente.

4.2.4. Vergleich der Rhythmusfamilien

Wie auch in Cao et al. (2014) und Gómez-Marin et al. (2015) wurde für den Vergleich von Rhythmusfamilien die Manhattan Distance verwendet. Dafür wurde auch hier die Zuordnung Synkope = -1, Stille = 0 und Onset = 1 gewählt. Der Vergleich des gesamten Rhythmus liefert einen Einzahlwert („direkte“ Distanz). Auch hier wurde zusätzlich der band- und instrumentenweise Vergleich zur späteren Anpassung der Gewichte vorgenommen. Dieser liefert jeweils Vektoren mit 27 (26 Frequenzbänder und CDF) bzw. 5 (Instrumente) einzelnen Werten.

5. Empirische Untersuchung

Während die Wahrnehmung von Rhythmus umfangreich untersucht wurde, ist über die Kriterien für die Bewertung und insbesondere den Vergleich von Rhythmen verschiedener Musikstücke deutlich weniger bekannt. Rhythmus wird als kognitives Konstrukt aufgefasst, auf vielen Ebenen der Vorverarbeitung (z.B. Wahrnehmung des Pulses oder Gruppieren zu größeren Einheiten) zeigen sich interindividuelle Unterschiede, die sich auch in der Bewertung der Ähnlichkeit einfacher Rhythmen niederschlägt (vgl. Kapitel 3.2). Über den Vergleich von ganzen Musikauschnitten auf rhythmische Ähnlichkeit liegen bisher noch keine empirischen Daten vor. Die für eine Ähnlichkeitsbewertung relevanten Parameter und deren Gewichtung sind also auf empirischer Basis nicht bekannt.

Die Überprüfung algorithmisch gewonnener Ähnlichkeitsmaße von Musikausschnitten muss jedoch mit einem validen Datensatz durchgeführt werden. Es bieten sich zur Generierung dieses Datensatzes mehrere Möglichkeiten. Eine manuelle Gruppierung von Musikausschnitten nach rhythmischer Ähnlichkeit von ein oder mehreren Personen liegt nahe, kann aber lediglich die Rhythmuswahrnehmung und -beurteilung der erstellenden Personen abbilden. Ein weiter verbreiteter Ansatz, Musik in Gruppen ähnlicher Rhythmen einzuteilen, ist die übliche Charakterisierung durch Genres. Dabei wird davon ausgegangen, dass Musikstücke gleichen Genres ähnliche Rhythmen aufweisen und Musikstücke verschiedener Genres sich auch rhythmisch unterscheiden. Allerdings basiert die Genreeinteilung auf vielen verschiedenen Parametern, insbesondere auch der Instrumentierung und harmonischen Eigenschaften.

Ein Teil dieser Arbeit ist daher die Generierung eines empirischen Datensatzes mit bewerteten Rhythmusähnlichkeiten von Musikstücken (*ground truth*). Diese dienen als Basis für die Evaluation der automatisch extrahierten Abstände.

5.1. Operationalisierung

Die Ähnlichkeit beschreibt zunächst eine Beziehung zwischen zwei Elementen (Lin et al., 1998). Eine Ähnlichkeit zwischen mehreren Elementen kann als Menge aller paar-

weisen Ähnlichkeiten aufgefasst werden. Für den empirischen Hörversuch wird also die Ähnlichkeit aller möglichen Paare von Elementen erfasst.

Eine übliche Darstellung ist in Form einer quadratischen Ähnlichkeitsmatrix A, in der die Ähnlichkeit für jedes Paar erfasst ist. Für jedes Element existiert eine Spalte und eine Zeile, die Ähnlichkeit zwischen dem Element i und dem Element j steht in der Position A_{ij}. Da die Ähnlichkeit nicht von der Reihenfolge abhängig ist, ist die Matrix symmetrisch, es gilt $A_{ij} = A_{ji}$. Alternativ kann von Distanzen und einer Distanzmatrix D gesprochen werden, bei Normierung der Ähnlichkeit auf Werte zwischen 0 und 1 gilt für die Distanzmatrix $D = 1 - A$.

5.1.1. Datenerfassung

Aus einem Datenpool von 28 kurzen Musikausschnitten werden zufällig Songpaare gewählt, die auf einer vierstufigen Skala auf ihre rhythmische Ähnlichkeit zwischen „gar nicht ähnlich“, „etwas ähnlich“, „ähnlich“ und „sehr ähnlich/gleich“ bewertet werden sollen. Zum beschleunigten Ablauf des Versuches werden Frageblöcke gebildet und dazu jeweils ein (zufällig gewählter) Referenzsong mit fünf (ebenfalls zufällig gewählten) Vergleichssongs verglichen, Abbildung 5.1 zeigt die Oberfläche des Hörversuchs. Das mehrfache Auftreten eines Vergleichssongs pro Block wird unterbunden, die erneute Auswahl des Referenzsongs ist jedoch möglich. Insgesamt werden sieben Blöcke à fünf Vergleichssongs mit jeweiligem Referenzsong präsentiert. Somit werden pro Versuchsdurchgang 35 Songpaare abgefragt. Für die Generierung von qualitativen Daten wird am Ende des Versuchs für drei feste Songpaare jeweils die Benennung der Parameter, die zur Bewertung berücksichtigt werden, abgefragt.

Neben den üblichen statistischen Angaben zum Erfassen der Stichprobe (Alter, Geschlecht) werden zusätzlich zu Beginn der Umfrage die rhythmisch relevante Aktivität (Spielen eines Instruments, eines Rhythmusinstruments, in Band/Orchester/Chor, Tanzen von Standard oder Turniertänzen) abgefragt, um einen möglichen Einfluss von rhythmischem Training erfassen zu können. Zudem wird die Wiedergabequelle der Songbeispiele erhoben (Smartphone/Laptop-Lautsprecher, einfache Kopfhörer, einfache externe Boxen, hochwertige Kopfhörer, HiFi-Lautsprecher).

5.1.2. Musikauswahl

Um den Fokus der Untersuchung auf das rhythmische Grundmuster (Pattern) der Musikstücke zu richten, weisen die präsentierten Musikausschnitte große Ähnlichkeit in

vielen Parametern auf. Die Ausschnitte haben alle eine Länge von acht Sekunden, in den letzten zwei Sekunden findet ein linearer Fade-out statt. Sie stehen alle im 4/4-Takt, beginnen ohne Verzögerung mit einer schweren Eins im ersten Takt und präsentieren schon im ersten Takt das relevante rhythmische Pattern. Da alle Stücke zudem im etwa gleichen Tempo um 120 bpm gespielt sind, ergibt sich eine Spieldauer von vier Takten. Jeder Musikausschnitt beinhaltet nur eine Schlagzeugspur, einen Bass und eine (Rhythmus-)Gitarre. Die Musikstücke wurden aus den RealTracks von BandInABox®[1] generiert, es handelt sich um von Musikern eingespielte Musik von hoher klanglicher wie musikalischer Qualität.

Ein intern durchgeführter Vorversuch zeigte für Musik verschiedenen Genres die Tendenz, die rhythmische Ähnlichkeit im Mittel als sehr gering (häufigste Antwort: gar nicht ähnlich) zu bewerten. Daher wurde bei der Auswahl der Ausschnitte auf eine grundsätzliche Vergleichbarkeit der rhythmischen Patterns geachtet. Die Musikstücke sind grob in drei, etwas feiner in fünf rhythmisch motivierte Kategorien einteilbar. Eine Gruppe weist starke Betonungen auf den geraden Achteln zwischen den vier Zählzeiten auf (Gruppe „Offbeats“), entweder gespielt von der Gitarre (oft im Ska zu finden) oder der SnareDrum (Country). Eine zweite Gruppe bilden Stücke mit lateinamerikanischen Bossa-Rhythmen („Bossa“. Der größte Teil der Musik ist dem Rock- und Country-Genre zuzuordnen. Hier bilden Rhythmen mit einer starken Betonung auf dem vierten Achtel (Zählzeit 2+, daher Gruppe „Rock2+“), Rhythmen des Funk („Funk“ und nicht weiter spezifizierte Rock-Rhythmen („div. Rock“) die Untergruppen. Die Übergänge zwischen den Gruppen sind fließend gewählt, einige Beispiele können in zwei Gruppen eingeteilt werden. Zur genaueren Betrachtung der Rhythmen findet sich im Anhang (Anhang A ab Seite 83) die Darstellung aller 28 Songausschnitte in Schlagzeug-Tabulatur-Notation, erweitert um die Rhythmen der Gitarre und des Bass.

Wenn diese Gruppierung auch nach rhythmischen Kriterien vorgenommen wurden, so ist dennoch eine klangliche Ähnlichkeit innerhalb der Gruppen gegeben. Insbesondere die Stücke mit lateinamerikanischen Rhythmen gruppieren sich ebenfalls aufgrund des ähnlichen Gitarrenklangs, der sich von der anderen Stücken abhebt. Dennoch wurde innerhalb der Gruppen auch auf eine klangliche Variabilität geachtet.

Es wird erwartet, dass sich diese Vorgruppierung auch in den Ergebnissen des Hörversuches widerspiegelt.

[1] www.bandinabox.com

5.2. Versuchsaufbau und -ablauf

Um ein subjektives Maß wie die bewertete Ähnlichkeit von Rhythmen zu erfassen, bieten Online-Hörversuche eine Möglichkeit, viele Versuchspersonen mit wenig Aufwand zu erreichen. Mit LimeSurvey[2] steht ein umfangreiches, quelloffenes und durch seine Nutzbarkeit in einem Browser orts- und plattformunabhängiges Tool mit zeitgemäßem Erscheinungsbild zur Vefügung, das für diese Studie genutzt wurde. Die Umfrage lief auf einem Server von zplane.development, der Zugriff war von jedem internetfähigen Endgerät mit aktuellem Browser und Soundwiedergabe möglich. Die Umfrage stand in deutscher und in englischer Sprache zur Verfügung.

Bei der Einladung der VersuchsteilnehmerInnen über verschiedene Emailverteiler wurde darauf hingewiesen, dass eine Unterbrechung des etwa 20-minütigen Versuchs jederzeit möglich ist, sowie die Anonymität der Teilnehmenden gewahrt bleibt. Die Teilnahme erfolgte freiwillig und unentgeltlich.

Auf der ersten Seite der Umfrage werden die TeilnehmerInnen neben einem erneuten Hinweis zum Datenschutz über die genaue Aufgabenstellung mit folgendem Wortlaut informiert:

> In dieser Umfrage geht es um rhythmische Ähnlichkeiten verschiedener Musikausschnitte.
>
> Dazu werden auf den folgenden acht Seiten jeweils ein Referenzstück und fünf Vergleichstücke von jeweils etwa acht Sekunden Länge präsentiert. Für jedes der Vergleichsstücke soll die rhythmische Ähnlichkeit mit dem jeweiligen Referenzstück bewertet werden. Die Musikausschnitte können so oft angehört werden, wie nötig.
>
> Jeder der Ausschnitte ist im 4/4-Takt, das Tempo ist in etwa gleich und auch eine ähnliche Instrumentierung wurde gewählt. Der Fokus soll auf den Rhythmus gelegt werden und das spontane Empfinden, wie ähnlich sich die Rhythmen sind. Klangliche Eigenschaften sollen so gut wie möglich ignoriert werden.

Die zweite Seite der Umfrage erfasst die Angaben zu rhythmischem Training, Wiedergabequelle sowie Alter und Geschlecht. Letztere Angaben wurden als freiwillig deklariert. Die Oberfläche der nächsten sieben Seiten zur Bewertung der rhythmischen Ähnlichkeit ist beispielhaft in Abbildung 5.1 zu sehen. Das Abspielen jedes Beispielausschnitts kann über die Steuerung des Audio-Players beliebig oft gestartet und pausiert werden. Bei Unsicherheit der Ähnlichkeitsbewertung für ein Songpaar kann die Vorgabe „keine Antwort“ beibehalten werden.

[2] www.limesurvey.org

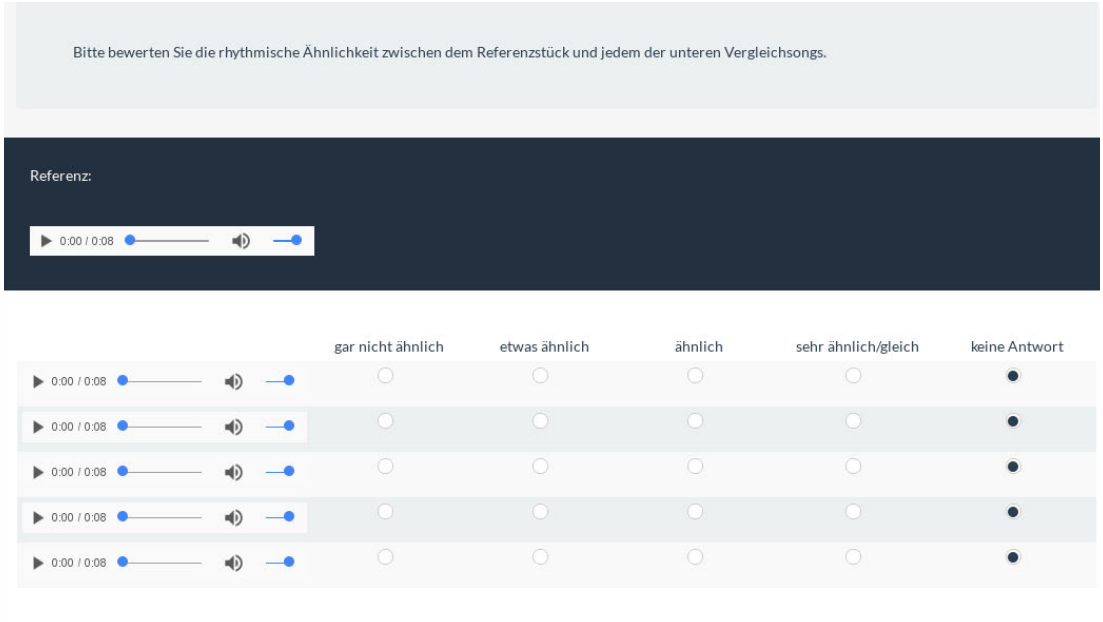

Abbildung 5.1.: Oberfläche zur Bewertung der rhythmischen Ähnlichkeit im Online-Hörversuch

Der Hörversuch endet nach der Abfrage der gewählten Parameter der Rhythmusähnlichkeit mit dem Hinweis, dass aufgrund der zufälligen Auswahl der Musikausschnitte ein mehrfache Teilnahme möglich sei. Für diesen Fall wird in dem Browser ein anonymes Cookie gesetzt, das eine eindeutige Kennung enthält. Bei einem erneuten Durchlauf kann durch Einlesen des Cookies eine Zusammenführung der Datensätze erfolgen.

5.3. Versuchsergebnisse

Insgesamt wurde 126 Zugriffe auf die Onlineumfrage registriert, von denen 97 Teilnehmende die erste Seite mit allgemeinen Informationen ausfüllten und jeweils zwischen 3 und 70 (zweimalige komplette Teilnahme) Songpaare bewerteten. Die Auswahl „keine Antwort“ wurde ignoriert.

5.3.1. Datenbereinigung

Für die Auswertung wurden nur die Antworten von Versuchspersonen berücksichtigt, die mindestens zehn Songpaare bewerteten, also 2 Blöcke à 5 Bewertungen bearbeiteten, da erst dann von einer ernsthaften Teilnahme ausgegangen wird.

Der verbleibende Datensatz wurde nach zwei Kriterien weiter bereinigt. Teilnehmende, die bei dem Vergleich zweier identischer Songausschnitte nicht die Antwort „sehr ähnlich/gleich“ auswählten, wurden aus dem Datensatz entfernt (2 Versuchspersonen). Desweiteren wurden Teilnehmende aus der Auswertung herausgenommen, deren Antworten zu viele Ausreißer enthielten. Für jedes Songpaar gelten als Ausreißer diejenigen Bewertungen, die um mehr als die doppelte Standardabweichung vom Mittelwert abweichen. Zwei Teilnehmende wiesen in mehr als 20 % der gegebenen Antworten Ausreißer auf und wurden aus dem Datensatz entfernt.

5.3.2. Stichprobe

Nach dieser Bereinigung lagen insgesamt 2789 Bewertungen von 86 verschiedenen VersuchsteilnehmerInnen im Alter zwischen 18 und 66 Jahren (Durchschnitt 32,7 Jahre, 14 mal keine Angabe) vor. 17 Personen gaben das weibliche Geschlecht an, 63 das männliche, eine Person Sonstiges und 5 mal gab es keine Angabe. Etwa 70 % der Teilnehmenden gaben an, ein Instrument zu spielen, knapp 31 % spielen Schlagzeug. Etwas mehr als die Hälfte der Versuchspersonen spielen mit anderen Menschen zusammen Musik (51 %), zu den TänzerInnen zählen sich 15 %.

Die meisten Teilnehmenden hörten die Musik über eingebaute Lautsprecher (32 %), höherwertige externe Lautsprecher nutzten knapp 17 %, die Wiedergabe über Hifi-Lautsprecher wurde sechs mal angegeben (ca. 7 %). Der Rest teilte sich auf einfache (21 %) und höherwertige (23 %) Kopfhörer auf.

5.3.3. Auswertung

Im Folgenden werden nur die Vergleiche verschiedener Songs betrachtet, die Vergleiche identischer Songs wurden ausschließlich als Filterkriterium verwendet. Es ergeben sich $28 \cdot 27 = 756$ Paare, die sich wegen der Symmetrie auf 378 verschiedene Songpaare reduzieren. Aufgrund der zufälligen Auswahl der Songpaare ergibt sich keine gleichmäßig verteilte Häufigkeit der Bewertung einzelner Paare. Im Mittel wurde jedes der Songpaare 7,1 mal bewertet, die genaue Verteilung ist in Abbildung 5.2 gezeigt. Insgesamt gibt es 33 Songpaare mit drei oder weniger Bewertungen, 14 Songpaare mit nur einer oder zwei Bewertungen. Dies sind jedoch weniger als 10 % (zwei oder weniger: 4 %) der Gesamtzahl aller Songpaare und wird nicht gesondert berücksichtigt.

Ein persönliches Feedback einiger Teilnehmenden zeigte, dass die Bewertung rhythmischer Ähnlichkeit eine anspruchsvolle Aufgabe ist. Die zu beachtenden Parameter seien

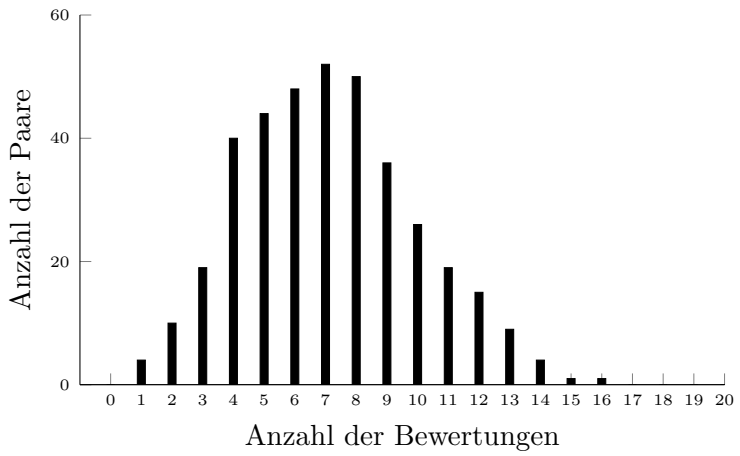

Abbildung 5.2.: Histogram über die Anzahl der erhaltenen Bewertungen pro Songpaar

sehr vielseitig, müssten für verschiedene Songpaare variiert werden; oftmals herrschte Unsicherheit, ob am Ende des Versuchsdurchlaufs tatsächlich konsistente Antworten gegeben wurden. Auch ein qualitativer Blick auf die benannten Ähnlichkeitsparameter im letzten Versuchsteil zeigt eine Vielzahl an gegebenen Antworten. Die Bandbreite reicht von rein klanglichen Beschreibungen wie dem Gitarrenklang über Tempounterschiede bis zu einer differenzierten Beschreibung des Schlagzeuggrooves.

So zeigen sich trotz der Bereinigung des Datensatzes innerhalb der Songpaare sehr verschiedene Antworten. Bei knapp 20% der Paare (75) wurde sowohl die Antwort „sehr ähnlich/gleich“ als auch die Antwort „gar nicht ähnlich“ gegeben. Mit einer großen Varianz musste aufgrund der Komplexität der Aufgabenstellung und dem Mangel einer eindeutigen Rhythmusdefinition gerechnet werden. Aufgrund der großen Anzahl an Bewertungen zeigen sich in den mittleren Bewertungen dennoch konsistente Ergebnisse. So liegt die mittlere Varianz innerhalb der Songpaare bei 0.07, sie ist damit signifikant kleiner als die Varianz einer zufälligen Bewertung ($p < .001$).

Auch die mittlere Ähnlichkeitsbewertung folgt der im Vorfeld vorgenommenen Gruppierung der ausgewählten Musikausschnitte, wie Abbildung 5.3 zeigt. Sehr klar heben sich die (meisten) Stücke der Offbeat-Gruppe sowie die Stücke, die allein dem Funk zugeordnet werden, von den anderen Stücken ab. Eng gruppieren sich auch ein Großteil der Stücke aus der Gruppe Rock2+ sowie einige der Stücke im Bossa-Rhythmus. Die Übergänge zwischen den Gruppen wurden in der Studie allerdings anders bewertet, als im Vorfeld. Interessant ist der sechste Song, der von den Versuchsteilnehmern eher dem Rock als der Offbeat-Gruppe zugeordnet wurde. Dieser Song zeigt starke Offbeats, die allerdings nur von der Gitarre gespielt werden, während das Schlagzeug einen geraden

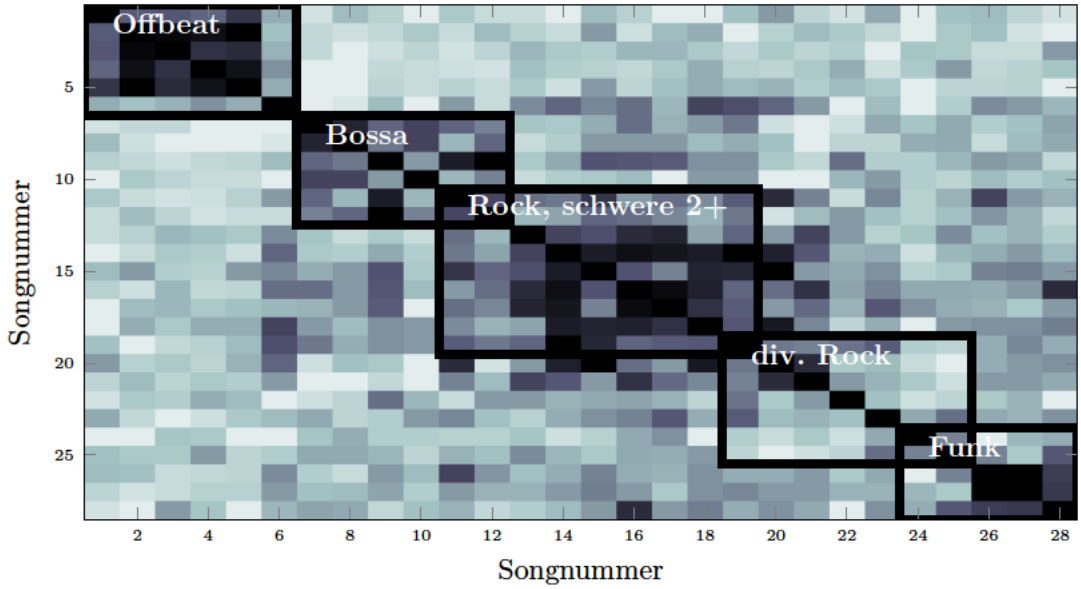

Abbildung 5.3.: Bewertete Ähnlichkeit aller Songpaare sortiert nach vordefinierten Gruppen. Dunkle Felder stehen für rhythmisch ähnliche Paare.

Rockbeat spielt (Bass Drum auf Eins und Drei, Snare Drum auf Zwei und Vier). In allen anderen Stücke der Offbeat-Gruppe werden die Schläge zwischen den Zählzeiten auch von der Snare Drum gespielt.

Die Songbeispiele mit nicht weiter spezifizierten Rock-Rhythmen zeigen keine starke Gruppierung. Die ersten Songs dieser Gruppe (19 bis 21) werden als ähnlich zu den Songs der Gruppe mit Betonung auf der Zählzeit 2+ bewertet, die Beispiele Nr. 22-25 wurden aber als zu keinem der anderen Songs ähnlich bewertet. Diese Ergebnisse decken sich mit dem Eindruck im Vorfeld, diese Rhythmen keiner Gruppe zuzuordenen und unter „diverse“ Rock-Rhythmen zusammenzufassen.

An den Beispielen der Gruppe Funk zeigt sich allerdings, dass einige Antworten wahrscheinlich auch von klanglichen Merkmalen beeinflusst wurden. Die Songs Nr. 26 und 27 sind klanglich praktisch identisch, unterscheiden sich aber im Rhythmus des Basses und des Schlagzeuges. Dieses Paar wurde jedoch immer mit „sehr ähnlich“ bewertet. Die Songs Nr. 25 und 28 hingegen unterscheiden sich klanglich stark, insbesondere im Klang der Gitarre und der Präsenz des Basses, die rhythmische Notation ließe jedoch auf eine rhythmische Ähnlichkeit schließen. Bei diesem Paar sind alle vier Antwortmöglichkeiten zur Bewertung der Ähnlichkeit vertreten.

Gruppe	Anzahl Paare	mittlere Varianz	Anzahl Bewertungen	mittlere Bewertung
Instrument				
Ja	358	0.067	1983	0.323
Nein	200	0.070	806	0.318
Schlagzeug				
Ja	233	0.050*	898	0.278*
Nein	354	0.073*	1891	0.342*
Band				
Ja	321	0.075*	1468	0.335
Nein	302	0.062*	1321	0.306
Gesamt	373	0.070	2789	0.322

Tabelle 5.1.: Varianz und Mittelwerte der gegebenen Antworten nach Gruppen sortiert. Die Varianz wurde über Paare mit mindestens zwei gegebenen Antworten bestimmt, der Mittelwert über alle gegebenen Antworten. Mit * versehene Angaben unterscheiden sich innerhalb der Gruppierung signifikant voneinander ($p < 0.01$).

5.3.4. Auswertung nach Gruppen

Durch die Abfrage der musikalischen Aktivität der Versuchspersonen können die Antworten in jeweils komplementären Gruppen ausgewertet werden. Drei Gruppenpaare wurden gebildet, basierend auf folgenden Angaben:

- „Ich spiele aktiv ein Instrument (auch Gesang oder Turntables)“ („Instrument“)
- „Ich spiele aktiv Schlagzeug oder Percussioninstrumente“ („Schlagzeug“)
- „Ich mache mit anderen Menschen Musik (Orchester, Chor, Band, Kammermusik, Jammen)“ („Band“)

Jede Versuchsperson ist in jedem Gruppenpaar einer Gruppe (Ja oder Nein) zugeordnet, die Gruppen haben verschiedene Größen. Eine Aufteilung nach der Tanzaktivität führte aufgrund zu weniger TänzerInnen nicht zu aussagekräftigen Ergebnissen. Tabelle 5.1 zeigt die Verteilung innerhalb der Gruppenpaare und die mittlere abgegebene Bewertung aller Gruppenmitglieder. Zudem ist die mittlere Varianz der Bewertung jedes Songpaares nach Gruppen sortiert aufgeführt, dabei wurden nur Songpaare berücksichtigt, die mindestens zwei Bewertungen erhalten haben. Die Bewertung wurde

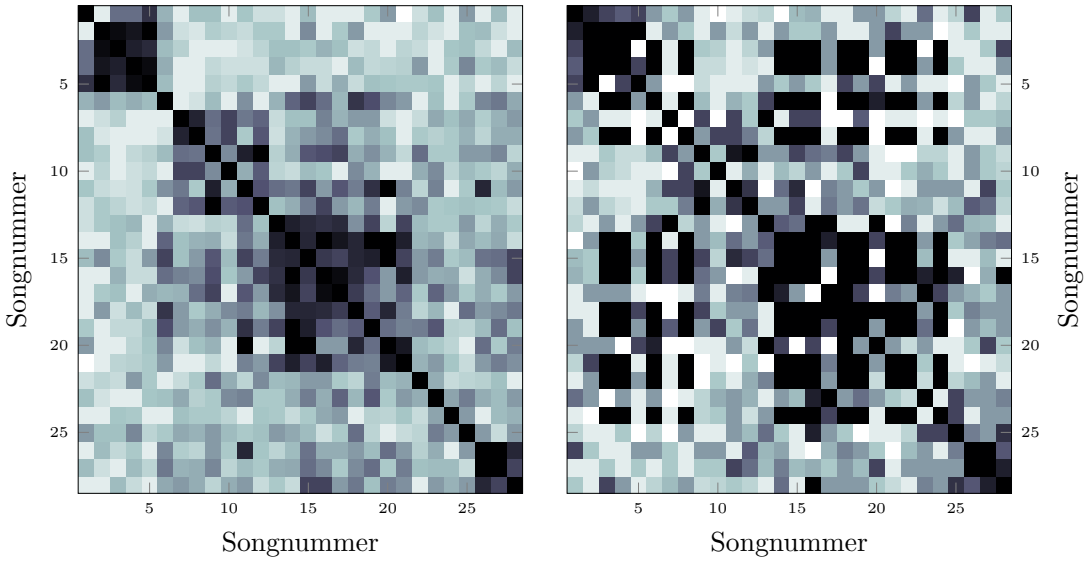

Abbildung 5.4.: Vergleich der Distanzmatrizen der Gruppen „spiele ein Instrument“ Ja (links) und Nein (rechts). Dunkle Felder stehen für rhythmisch ähnliche Paare, weiße Felder enthalten keine Bewertung.

auf den Bereich von 0 („gar nicht ähnlich“) bis 1 („sehr ähnlich/gleich“) normiert, die Zwischenschritte liegen bei 0.33 („etwas ähnlich“) und 0.66 („ähnlich“). Die mittlere Bewertung der Ähnlichkeit liegt mit 0.32 also leicht unterhalb unterhalb von „etwas ähnlich“.

Ein signifikanter Unterschied in der Varianz der Bewertung oder der mittleren Bewertung der rhythmischen Ähnlichkeit zwischen Versuchspersonen, die ein Instrument spielen und denen, die kein Instrument spielen, ist nicht zu erkennen. Ein Vergleich der Distanzmatrizen nach dieser Gruppierung in Abbildung 5.4 zeigt jedoch einen deutlichen Unterschied. Auf der linken Seite (Instrument: Ja) bilden sich vier deutliche Gruppen (Offbeat, Bossa, Rock2+ und Funk), auf der rechten Seite (Instrument: Nein) werden einige Songs der Gruppe Offbeat mit den Stücken der verschiedenen Rock-Gruppen als sehr ähnlich eingestuft, lediglich die Stücke des Funk heben sich etwas von den anderen Stücken ab. Zu beachten ist allerdings, dass einige Songpaare insbesondere bei der Gruppe Nein, gar keine Bewertung erhalten haben, diese Felder sind weiß. In Abbildung 5.7 (links) ist die relative Verteilung der gegebenen Bewertungen dieser beiden Gruppen im direkte Vergleich aufgeführt. Auffällig ist, dass die Versuchspersonen, die kein Instrument spielen, fast 25 % aller Paare als „sehr ähnlich/gleich“ bewerteten. Dieser Anteil hebt sich deutlich von den Anteilen aller anderen Gruppen ab.

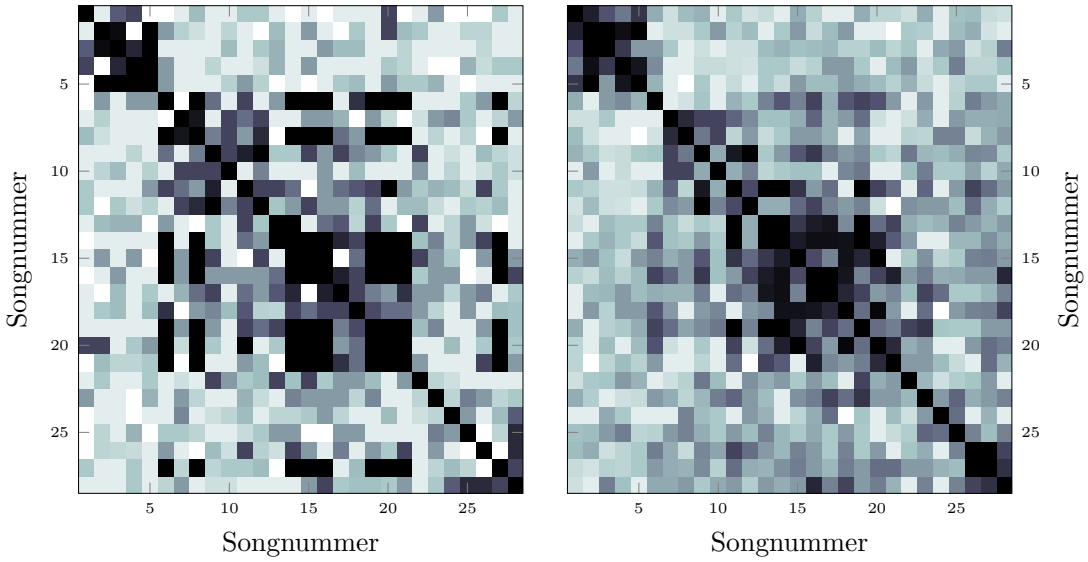

Abbildung 5.5.: Vergleich der Distanzmatrizen der Gruppen „spiele Schlagzeug“ Ja (links) und Nein (rechts). Dunkle Felder stehen für rhythmisch ähnliche Paare, weiße Felder enthalten keine Bewertung.

Versuchspersonen, die angaben, Schlagzeug zu spielen, bewerteten die Rhythmen im Mittel als weniger ähnlich (0.278 gegen 0.342), stimmten in ihren Bewertungen jedoch besser miteinander überein (Varianz: 0.05 gegenüber 0.073). Beide Unterschiede sind signifikant (Varianzanalyse mittels Kruskal-Wallis-Test, $p < 0.01$). Der Blick auf die Distanzmatrizen dieser Gruppierung in Abbildung 5.5 (links: Schlagzeug Ja, rechts: Schlagzeug Nein) zeigt ein gemischtes Bild. Zwischen den Stücken des Bossa, des Rock und des Funk finden sich Paare, die als rhythmisch sehr ähnlich bewertet werden, die Vorgruppierung in fünf Gruppen zeigt sich nicht. Auch in dieser Distanzmatrix finden sich allerdings viele fehlende Songpaare. Die Schlagzeug spielenden vergaben häufiger die Bewertungen „gar nicht ähnlich“ und „sehr ähnlich/gleich“, was auch im Histogram der mittleren Bewertungen zu erkennen ist (Abbildung 5.7 mitte). Gerade im Vergleich zu den Nicht-SchlagzeugerInnen zeigt sich eine andere Form der Verteilung mit deutlich weniger Anteilen in den mittleren Bewertungen.

Die Aufteilung nach der musikalischen Aktivität in Band oder Orchester führt zu etwa gleich großen Gruppen. Interessant ist, dass hierbei die Varianz der gegebenen Antworten der Gruppe musikalisch aktiver Versuchspersonen signifikant größer ist (Varianzanalyse mittels Kruska-Wallis-Test, $p < 0.01$). Die Distanzmatrizen sind in Abbildung 5.6 dargestellt, sie zeigen ähnliche Gruppierungen wie die Aufteilung nach Spielen eines Instrumentes. Allerdings korrelieren die Gruppenmitglieder sehr stark. So gehören alle

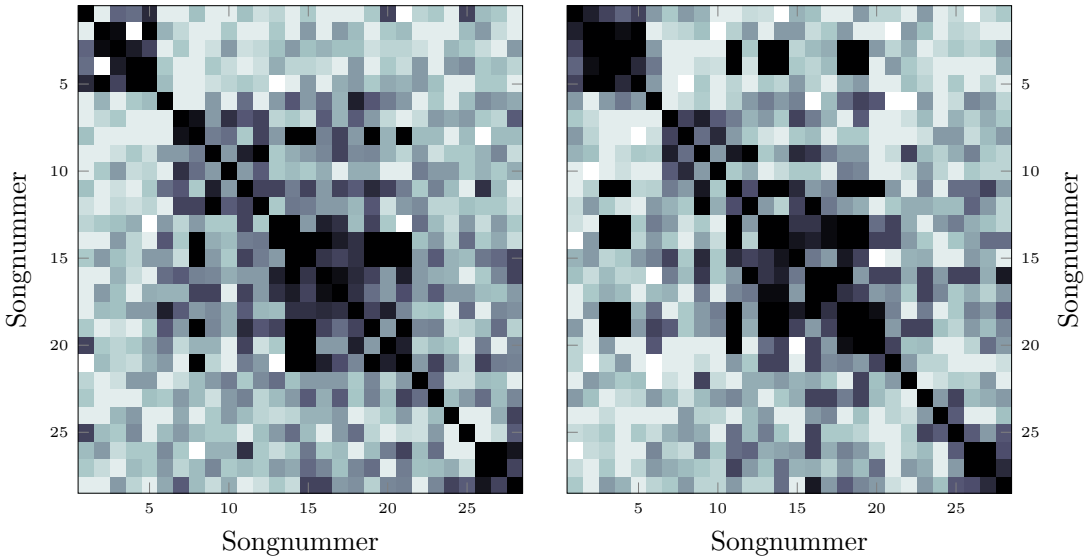

Abbildung 5.6.: Vergleich der Distanzmatrizen der Gruppen „spiele in Band/Orchester“ Ja (links) und Nein (rechts). Dunkle Felder stehen für rhythmisch ähnliche Paare, weiße Felder enthalten keine Bewertung.

Versuchspersonen, die in Band oder Orchester spielen zu denjenigen, die ein Instrument spielen. Versuchsteilnehmende, die kein Instrument spielen, fallen auch zwangsläufig in die Gruppe derjenigen, die nicht in Band oder Orchester spielen. Die Verteilung der mittleren Bewertung unterscheidet sich zwischen diesen beiden Gruppen kaum (Abbildung 5.7 rechts).

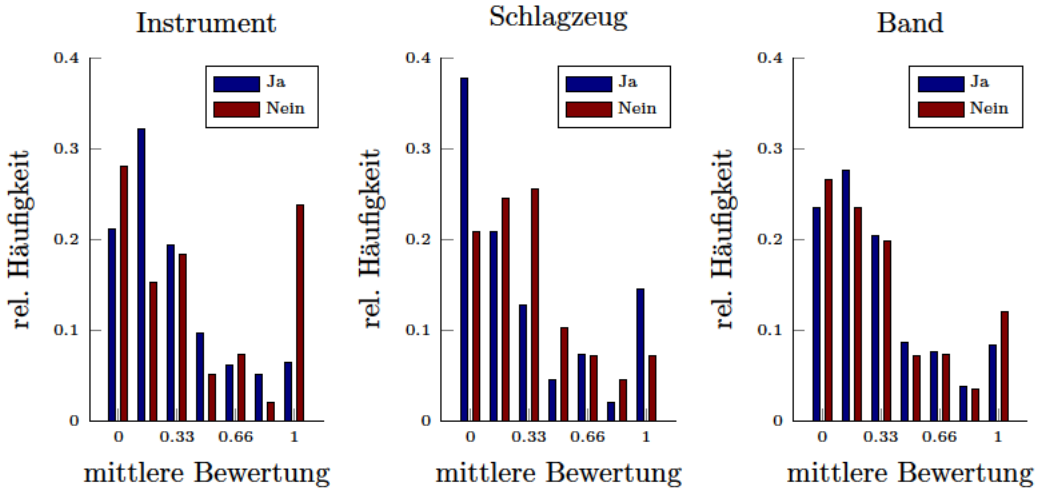

Abbildung 5.7.: Histogramme über die mittleren Ähnlichkeitsbewertungen pro Songpaar nach den drei Gruppierungen getrennt. Durch die Normierung entspricht der Wert 0 der Bewertung „gar nicht ähnlich“, der Wert 1 entspricht „sehr ähnlich“.

6. Evaluation

Die Ergebnisse der empirischen Untersuchung zeigen eine gute Abbildung der manuellen Gruppierungen der Rhythmen, die im Vorfeld vorgenommene grobe Einschätzung der rhythmischen Ähnlichkeit spiegelt sich in den Antworten der Versuchteilnehmer wieder. Wie gut die vorgestellten algorithmischen Verfahren zur Bewertung der Rhythmusähnlichkeit mit den empirischen Daten übereinstimmen, wird in diesem Kapitel untersucht. Neben dem Vergleich mit den Hörversuchsdaten wird zusätzlich der Ballroom Datensatz zur Auswertung der verschiedenen Verfahren herangezogen.

Jedes Maß für die Bewertung der rhythmischen Ähnlichkeit, ob basierend auf den Bar Patterns, den Rhythmusfamilien, oder bekannten Verfahren wie den Fluctuation Patterns, liefert, angewendet auf den jeweiligen Datensatz, eine zur Hauptdiagonalen symmetrische Distanzmatrix, die den paarweisen Vergleich aller beteiligten Musikausschnitte enthält. Da für die Daten des Hörversuchs eine Referenz-Distanzmatrix vorliegt, die Musikausschnitte des Ballroom Datensatzes jedoch nur in Klassen eingeteilt sind, ergeben sich für beide Datensätze verschiedene Evaluationsmethoden, die in den jeweiligen Unterkapiteln erläutert werden.

Folgende Abstandsmaße wurden für die Auswertung implementiert und miteinander verglichen:

Fluctuation Patterns nach Pampalk (2001), Vergleich mittels Euklidischer Distanz

Onset Patterns nach Pohle et al. (2009), Vergleich mittels Euklidischer Distanz

Bar Patterns, Vergleich mittels Kreuzkorrelation

Bar Patterns aligned*, korrekt (manuell) ausgerichtete Bar Patterns, Vergleich mittels Korrelationskoeffizient

Rhythmic Families 4* Rhythmusfamilien basierend auf manuell ausgerichteten Bar Patterns und Viertelnoten (nur Hörversuchsdatensatz)

Rhythmic Families 8* Rhythmusfamilien basierend auf manuell ausgerichteten Bar Patterns und Achtelnoten

Rhythmic Families 16* Rhythmusfamilien basierend auf manuell ausgerichteten Bar Patterns und Sechzehntelnoten

Annotated Famlies 4** Rhythmusfamilien basierend auf annotierten Rhythmen in Instrumenten und Viertelnoten (nur Hörversuchsdatensatz)

Annotated Famlies 8** Rhythmusfamilien basierend auf annotierten Rhythmen in Instrumenten und Achtelnoten (nur Hörversuchsdatensatz)

Annotated Families 16** Rhythmusfamilien basierend auf annotierten Rhythmen in Instrumenten und Sechzehntelnoten (nur Hörversuchsdatensatz)

Rhythmic Patterns* nach Dixon et al. (2004)

Rhythmic Patterns min Die Ähnlichkeit der Rhythmic Patterns ist die minimale Manhattandistanz bei beliebiger Verschiebung der Patterns zueinander (nur Hörversuchsdatensatz).

Eine Markierung mit * zeigt an, dass dieses Abstandsmaß eine (manuelle) Taktausrichtung erfordert, Abstandsmaße mit ** erfordern eine manuelle Annotation des gesamten Rhythmus.

6.1. Datensatz des Hörversuchs

Die Distanzmatrizen der Hörversuchsdaten haben eine Größe von 28x28 Musikstücken. In Abbildung 6.1 sind die Distanzmatrizen der Fluctuation Patterns, der Onset Patterns, der Bar Patterns und der Rhythmic Patterns abgebildet. Ein erster Blick zeigt, dass auch die implementierten Vergleichsverfahren die Musikbeispiele gruppieren. Die Offbeat-starken ersten Stücke (1-6) werden von den Bar Patterns sehr stark, von den Rhythmic Patterns und den Onset Patterns deutlich weniger stark und von den Fluctuation Patterns kaum zusammengefasst. Die Bar Patterns berücksichtigen die rhythmische Phasenlage innerhalb der Takte und können so einen starken Offbeat erkennen, die Phaseninformationen der Rhythmic Patterns hingegen unterscheiden nicht zwischen den Frequenzbändern und somit Instrumenten und Klängen. Auch die Stücke der verschiedenen Rockgruppen (13-23) werden von den Bar Patterns gruppiert, die Abgrenzung der Funk Rhythmen (24-28) ist hingegen nicht sehr stark. Bei den Onset Patterns heben sich insbesondere die Bossa Rhythmen von den übrigen Musikausschnitten ab, die Fluctuation Patterns bilden weniger deutliche Gruppen. Abgesehen von den Funk Rhythmen und einzelnen Stücken wird für den Großteil der Musikausschnitte eine vergleichbare Ähnlichkeit ermittelt.

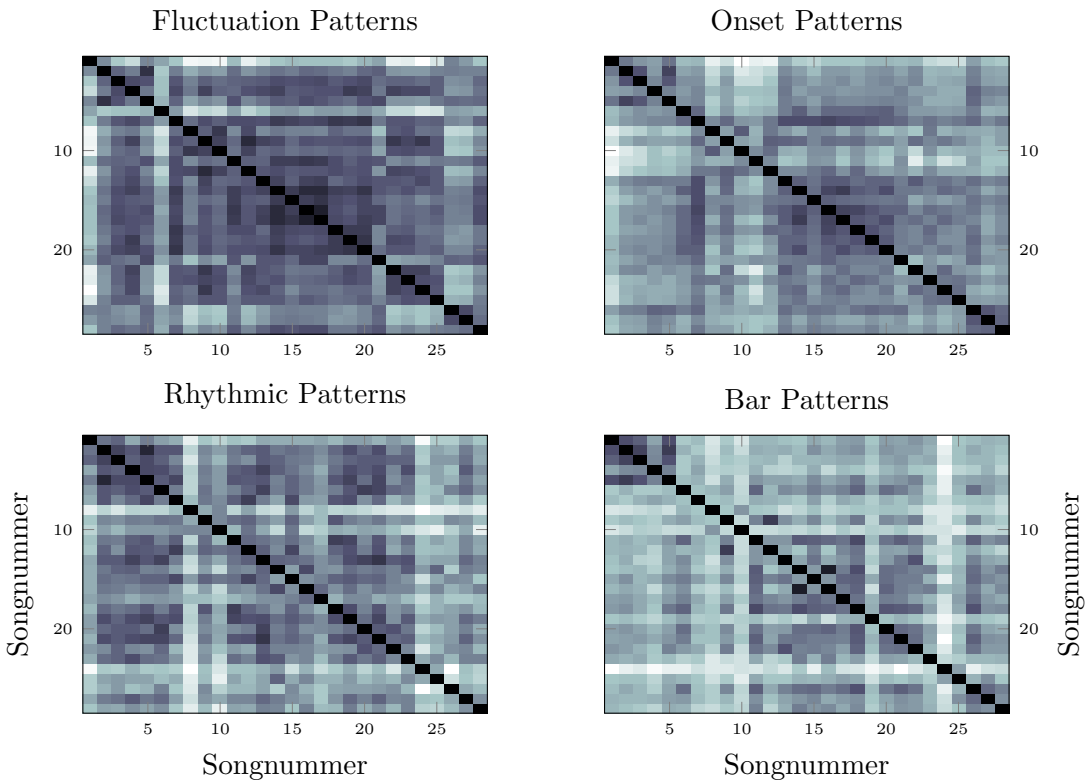

Abbildung 6.1.: Distanzmatrizen einiger Abstandsmaße. Dunkle Felder stehen für rhythmisch ähnliche Paare.

Für eine quantitative Evaluation der Distanzmaße wird die Korrelation der Distanzmatrizen mit der Distanzmatrix der empirischen Untersuchung bestimmt. Die obere Dreiecksmatrizen jedes Distanzmaßes werden dazu linearisiert, also jeweils in einen eindimensionalen Vektor mit 378 Elementen umgewandelt und der Korrelationskoeffizient mit der Referenzdistanz berechnet.

Abgesehen von den Rhythmic Patterns lagen für die anderen Abstandsmaße zusätzlich auch die Abstände in den einzelnen Frequenzbändern vor. So konnten für jedes Maß die Frequenzbänder verschieden gewichtet und an die empirischen Daten angepasst werden. Dies geschah über ein lineares Modell, bei dem lediglich positive Gewichte erlaubt waren (nonnegative-least-squares-Algorithmus nach Lawson und Hanson (1974)). Um ein Overfitting zu verhindern wurde das Modell über eine siebenfache Kreuzvalidation (7-fold-cross-validation) erstellt. Diese Anpassung ist jedoch mit Vorsicht zu betrachten, wie auch bei der Auswertung auf dem Ballroom Datensatz zu sehen sein wird (siehe Kapitel 6.2). Bei den Musikausschnitten des Hörversuch-Datensatzes wurde auf eine gewisse klangliche Ähnlichkeit geachtet, um den Fokus besser auf rhythmische Dif-

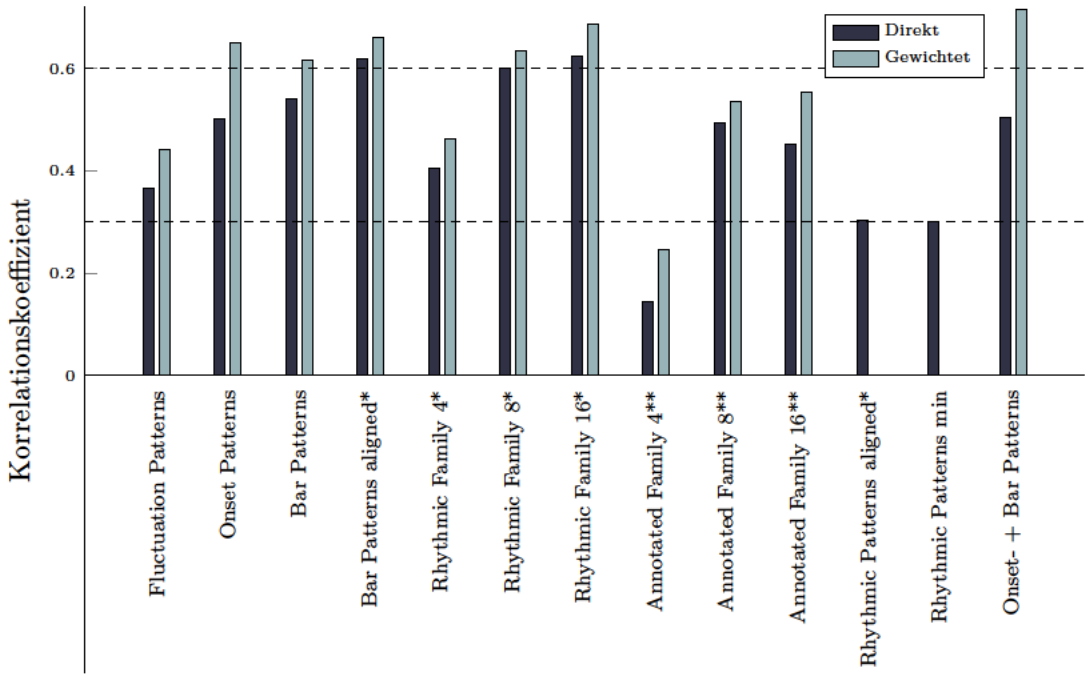

Abbildung 6.2.: Korrelation der algorithmisch bestimmten Rhythmusähnlichkeiten mit den empirischen Daten. Mit * versehene Methoden verlangen die korrekte Kenntnis der Taktanfänge, Methoden mit ** erfordern eine manuelle Annotation des gesamten Rhythmus.

ferenzen legen zu können. Die beteiligten Instrumente sind mit Schlagzeug, Bass und Gitarre in allen Tracks identisch und definieren bereits relevante Frequenzbereiche. Auf Musik mit anderen beteiligten (Rhythmus-)Instrumenten wird eine solche Anpassung nicht zwangsläufig zutreffen. Die Ergebnisse dieses Vergleichs sind in Abbildung 6.2 aufgeführt. Da die Rhythmic Patterns nicht in verschiedenen Frequenzbändern bestimmt werden, konnte dort eine Anpassung nicht vorgenommen werden.

Für alle Algorithmen gilt, dass eine angepasste Gewichtung der Frequenzbänder die Ergebnisse verbessert, insbesondere die Onset Patterns mit über 30 Frequenzbänder profitieren. Die Darstellung des Rhythmus als Rhythmusfamilie auf Basis von Viertelnoten (also eine Charakterisierung mit lediglich zwei Zahlen pro Frequenzband/Instrument) erzielt sowohl basierend auf Frequenzbändern als auch auf Instrumentenbasis verglichen mit den feiner auflösenden Rhythmusfamilien deutlich schlechtere Ergebnisse. Der Rhythmusvergleich mittels Rhythmusfamilien auf Basis von Sechzehnteln ist zumindest auf diesem Datensatz dem Vergleich der Rhythmusfamilie auf Basis von Achtelnoten überlegen, das gilt sowohl für die Rhythmusfamilien in Frequenzbändern, als auch in Instrumenten mit Gewichtung.

Die Rhythmic Patterns erzielen auf diesem Datensatz einen sehr geringen Korrelationkoeffizienten von 0.3, unabhängig von der Kenntnis der korrekten Ausrichtung. Dies steht entgegen den bisherigen Ergebnissen der Rhythmic Pattern. Allerdings ist die grundlegende Ähnlichkeit einiger Rhythmusparameter der Musikbeispiele (gleiche Taktart, lediglich binäre Unterteilungen) ein Unterschied zu anderen Datensätzen. Die verschiedenen Rhythmen dieses Datensatzes zeichnen sich primär durch unterschiedliche Instrumente auf den Zählzeiten aus, die reine Betrachtung der Energieverteilung innerhalb der Takte ist offensichtlich für den genauen Rhythmusvergleich unzureichend.

Die höchsten Korrelationskoeffizienten von über 0.65 erzielen die frequenzbandbasierten Rhythmusfailien mit Sechzehntelunterteilung und die korrekt ausgerichteten Bar Patterns. Beide Abstandsmaße basieren jedoch auf der Kenntnis der korrekten Taktanfänge.

Von den Ansätzen, die keine korrekte Taktausrichtung erfordern, zeigen die Onset Patterns und die Bar Patterns die deutlich besseren Ergebnisse. Die Vorteile der Onset Patterns gegenüber den Fluctuation Patterns können also auch in dieser Untersuchung bestätigt werden. Bei angepasster Gewichtung der Frequenzbänder erzielen die Bar Patterns und die Onset Patterns vergleichbare Ergebnisse. Der Unterschied zu den Rhythmusvergleichsverfahren, die eine korrekte Taktausrichtung verlangen, ist sehr klein. Valide Aussagen über Rhythmusähnlichkeit können demzufolge zumindest auf diesem Datensatz auch ohne explizite Downbeaterkennung gemacht werden.

Allgemein ist zu beobachten, dass eine höhere Dimensionalität der Novelty Function die Ergebnisse verbessert. Das zeigt sich bei den Vergleichen der frequenzbandbasierten Rhythmusfamilien gegenüber den instrumentenbasierten, der Onset Patterns mit den Fluctuation Patterns und der Bar Patterns mit den Rhythmic Patterns, in allen Fällen auch ohne eine explizite Anpassung der Gewichte.

Die Onset Patterns erfassen Fluktuationen in vielen Frequenzbereichen, ignorieren jedoch die Phaseninformationen, welche in den Bar Patterns (in weniger Frequenzbereichen) enthalten sind. Die enthaltenen Informationen sind in beiden Darstellungen also nicht identisch. Tatsächlich kann durch ihre Kombination ein Informationgewinn erzielt werden, die gewichtete Summe der Distanzmatrizen der angepassten Onset Patterns (d_{OP}) und der Bar Patterns (d_{BP})

$$d_{Komb} = d_{BP} + 2 \cdot \sqrt{d_{OP}}$$

zeigt einen Korrelationkoeffizienten mit der empirisch gewonnenen Distanzmatrix von 0.73.

6.2. Ballroom Datensatz

Um die Ergebnisse des relativ kleinen erhobenen Datensatzes zu verifizieren, wurde eine Genreklassifikation auf dem Ballroom Datensatz durchgeführt. Der Ballroom Datensatz besteht aus 698 Songausschnitten von etwa 30 Sekunden Länge, die acht verschiedenen Genres bzw. Standardtänzen zugeordnet sind. Diese Aufteilung entspricht primär rhythmischen Eigenschaften, eine Klassifizierung nach rhythmischer Ähnlichkeit sollte den Genres also gut entsprechen.

Der Ballroom Datensatz wurde für viele Untersuchungen der rhythmischen Ähnlichkeit herangezogen, so wurden für die Fluctuation Patterns korrekte Klassifikationsraten von 75 %, für die Onset Patterns Klassifikationsraten von 86,7 % berichtet (Pohle et al., 2009), durch Erweiterung des Modells um Timbreinformationen erreichten Pohle et al. (2009) sogar Werte um 90 %. Unter Berücksichtigung der korrekten (manuell annotierten) Taktanfänge konnten mit den Rhythmic Patterns, kombiniert mit einer ganzen Reihe rhythmischer Features (insgesamt 79), korrekte Klassifikationsraten von bis zu 96,0 % erreicht werden (Dixon et al., 2004).

Für den Ballroom Datensatz wurden die gleichen Ähnlichkeitsmaße verwendet, wie bei der Evaluation mit dem Hörversuchs-Datensatz. Auf die Rhythmusfamilien auf Basis von Viertelnoten wurde verzichtet, zudem liegen für den Ballroom Datensatz keine instrumentenweise Annotationen des Rhythmus vor, diese Maße fließen ebenso nicht in den Vergleich ein.

Da sich die Genres des Ballroom Datensatzes neben den Rhythmen insbesondere auch durch die Tempi unterscheiden, werden zusammen mit den auf Taktlängen normalisierten Patterns (Bar Patterns, Rhythmusfamilien und Rhythmic Patterns) auch die Tempoangaben (Taktlängen in Sekunden) extrahiert. Auf diese Weise wurden zwei Distanzmatrizen erzeugt, die paarweisen rhythmischen Distanzen $d_{Pattern}$, sowie die „Abstände“ (der Quotient kürzere Taktlänge durch längere Taktlänge) der Tempi d_{Tempo}. Für die Rhythmusfamilien, die korrekt ausgerichteten Bar Patterns und die Rhythmic Patterns wird auf die annotierten Taktlängen zurückgegriffen, die „normalen“ Bar Patterns arbeiten mit den automatisch extrahierten Taktlängen.

Für die Klassifikation wird das *k-nearest-neighbours*-Verfahren (KNN) mit $k = 7$ verwendet, die beste Gewichtung von Pattern- und Tempodistanzen wird mit einer *leave-one-out cross-validation* ermittelt. Die Ergebnisse der Klassifikationen sind in Tabelle 6.1 zusammengefasst.

Distanzmaß	**Klassifikationsrate**	**Modell**
Fluctuation Patterns	75,0 %[1]	
Onset Patterns	86,7 %[1]	
Bar Patterns	89,9 %	$d_{BP} + 0.3 \cdot d_T$
Bar Patterns + Onset Patterns	92,4 %	$d_{BP} + 0.01 * d_{OP} + 0.15 \cdot \sqrt{d_T}$
Bar Patterns Aligned*	96,7 %	$d_{BPA} + 0.5 \cdot d_T$
Rhythmic Families 8*	91,0 %	$d_{RF8} + 2 \cdot d_T$
Rhythmic Families 16*	91,7 %	$d_{RF16} + 1.5 \cdot d_T$
Rhythmic Patterns*	87,1 %[2]	
Rhythmic Patterns + Features*	96,0 %[2]	

Tabelle 6.1.: Ergebnisse der Klassifikation des Ballroom Datensatzes. Die Distanzmaße im unteren Teil mit * erfordern eine manuelle Annotation der Taktanfänge.

Die Verfahren, die auf die Annotation der Takte zurückgreifen erzielen erwartungsgemäß im Allgemeinen höhere Klassifikationsraten. Der Rhythmusvergleich mittels Rhythmusfamilien zeigt auch hier gute Klassifikationsraten, wieder erzielen die Rhythmusfamilien auf Basis von Sechzehntelnoten etwas bessere Ergebnisse.

Beide Distanzmaße der Rhythmusfamilien klassifizieren besser als die Rhythmic Patterns mit Tempoangaben kombiniert, eine Kombination der Rhythmic Patterns mit einem Satz rhythmussensitiver Features nach Dixon et al. (2004) setzt sich mit 96,0 % hingegen deutlich ab.

Die korrekt ausgerichteten Bar Patterns erreichen mit 96,7 % sogar eine noch etwas höhere Klassifikationsrate. Die Vorteile der getrennten Betrachtung verschiedener Frequenzbänder zeigt sich gegenüber den Rhythmic Patterns mit lediglich Tempoannotationen deutlich.

Auch bei den Verfahren, die ohne Annotationen auskommen, bestätigen sich die Ergebnisse des Hörversuchsdatensatzes teilweise. Die Onset Patterns klassifizieren deutlich besser als die Fluctuation Patterns, liegen hier jedoch unterhalb der Bar Patterns, welche mit fast 90 % eine sehr hohe Klassifikationsrate aufweisen. Der sehr hohe Aufwand zur Kompensation eines möglichen Fehlers der Taktlängenabschätzung und den Faktor Zwei (vgl. Kapitel 4.1.6) trägt hier lediglich zu einer Verbesserung von etwa einem

[1] aus Pohle et al. (2009)
[2] aus Dixon et al. (2004)

	Cha.	Jive	Quick	Rum	Sam	Tan	WW	LW
Cha-Cha-Cha	108	0	0	3	0	0	0	0
Jive	0	57	0	0	0	2	0	1
Quickstep	1	0	77	4	0	0	0	0
Rumba	5	3	0	76	1	5	0	8
Samba	2	0	1	9	72	1	0	1
Tango	4	0	0	1	0	80	0	1
Wiener Walzer	0	0	0	0	0	1	61	3
Langs. Walzer	0	0	0	0	0	1	11	94

Tabelle 6.2.: Klassifikation der Musikausschnitte des Ballroom-Datensatzes, nach Genres aufgeteilt für die Bar Patterns ohne manuelle Ausrichtung.

Prozentpunkt bei. Die Berechnungszeit steigt jedoch um mehr als das Vierfache. Dieser Aufwand ist also möglicherweise gar nicht erforderlich.

Da die Extraktion der Bar Patterns auf einer Erkennung der Taktlänge beruht, müssen auch die Fehlerraten dieser Abschätzung berücksichtigt werden. Insbesondere im Genre Walzer finden sich Stücke klassischer Instrumentierung, bei denen das hier verwendete Verfahren Schwächen aufweist. Dies ist auch in Tabelle 6.2 zu erkennen, welche die jeweilige Klassifikation der Musikstücke nach Genre sortiert zeigt. Ein nicht unerheblicher Teil der Fehler liegt in der Vertauschung von Wiener Walzer und Langsamem Walzer.

Die Distanzmatrix der Rhythmen aus dem Ballroom Datensatz wurden ohne Gewichtung der Frequenzbänder bestimmt. Eine Anwendung der Gewichte, die sich aus dem Hörversuchsdatensatz ergeben, verschlechterte das Ergebnis erheblich (Klassifikationsraten von ca. 75 %). In den Stücke des Ballroom Datensatzes spielen mehr Instrumente als nur Schlagzeug, Bass und Gitarre, gerade bei Tänzen wie dem Walzer finden sich rhythmisch relevante Stimmen auch gerne in mittleren Frequenzbändern. Dies war im anderen Datensatz kaum der Fall und zeigt, dass die gewonnenen Gewichtungen nicht allgemeingültig sind.

Auch der Ballroom Datensatz kann mit dem kombinierten Ähnlichkeitsmaß aus Bar Patterns und Onset Patterns klassifiziert werden. Dazu werden wieder die einzelnen Distanzmatrizen (Bar Patterns, Onset Patterns und extrahiertes Tempo) gewichtet addiert. Die Klassifikationsrate dieser Kombination liegt mit 92,4 % oberhalb der bisher beobachteten Werte für Verfahren, die ohne manuelle Annotationen auskommen. Die Klassifikationsmatrix (Tabelle 6.3) zeigt, dass durch die Kombination mit den Onset Patterns insbesondere bei den Genres Wiener Walzer, Langsamer Walzer, Tango

	Cha.	Jive	Quick	Rum	Sam	Tan	WW	LW
Cha-Cha-Cha	105	0	0	6	0	0	0	0
Jive	1	54	1	3	0	1	0	0
Quickstep	1	0	79	2	1	0	0	0
Rumba	1	0	0	82	0	6	0	9
Samba	2	1	2	6	74	1	0	0
Tango	0	0	0	1	0	84	0	1
Wiener Walzer	0	0	0	1	0	0	60	4
Langs. Walzer	0	0	0	3	0	0	2	105

Tabelle 6.3.: Klassifikation der Musikausschnitte des Ballroom-Datensatzes, nach Genres aufgeteilt für die Kombination Bar Patterns und Onset Patterns.

und Rumba bessere Klassifikationsergebnisse erzielt werden, als mit den Bar Patterns alleine.

7. Zusammenfassung und Ausblick

In dieser Arbeit wurde ein neuer Ansatz für den rhythmischen Vergleich zweier als Audiomaterial vorliegender Musikausschitte, die Bar Patterns, entwickelt, implementiert und getestet, der als Kombination von bekannten Ansätzen aufgefasst werden kann: Die Phaseninformationen der Rhythmic Patterns werden mit einer Frequenzbandaufteilung von Fluctuation- bzw. Onset Patterns kombiniert. Die Bar Patterns basieren auf einer mehrdimensionalen novelty function, die die Änderung des Energieverlaufs in vielen Frequenzbändern, sowie eine Chroma Differenz Funktion enthält. Diese Novelty Function wird in Takte eingeteilt und auf eine einheitliche Länge normiert, eine korrekte Ausrichtung auf musikalische ist nicht erforderlich. Der Median über die Takte rhythmisch ähnlicher Bereiche bildet das Bar Pattern. Der Vergleich von Bar Patterns wird mit einer zeilenweisen Kreuzkorrelation durchgeführt.

Zur Evaluation wurde neben der Verwendung des Ballroom Datensatzes in einem empirischen Hörversuch ein neuer Datensatz zur rhythmischen Ähnlichkeit erstellt. 28 Musikausschnitte wurden paarweise von 86 Personen in 2789 einzelnen Einstufungen auf rhythmische Ähnlichkeit bewertet. Die Ergebnisse des Hörversuchs lassen trotz des für empirische Untersuchungen üblichen Rauschens eine sehr einheitliche Bewertung vieler Songpaare erkennen, eine Vorgruppierung nach rhythmischen Kriterien konnte in den Ergebnissen gut wiedergefunden werden.

Neben den Bar Patterns wurde der Ansatz, Rhythmen in Rhythmusfamilien einzuteilen und zu vergleichen, von binären, eindimensionalen Rhythmen aus Klickgeräuschen auf Musik erweitert. Als Vergleichsverfahren wurden insbesondere die Rhythmic Patterns und die Onset Patterns betrachtet, die in bisherigen Arbeiten zu Rhythmusähnlichkeit und -klassifikation gute Ergebnisse erzielten.

Im Vergleich der Ähnlichkeitsprognosen zeigten die meisten der untersuchten Verfahren einen vergleichbar hohen Korrelationskoeffizienten mit den Daten des Hörversuchs um 0,6. Die Bar Paterns und die Onset Patterns erreichten erst mit einer angepassten Gewichtung der Frequenzbänder einen solchen Wert. Die Rhythmic Patterns erzielten trotz Kenntnis der Taktanfänge keine hohe Korrelation mit den Empirischen

Daten. Der Vergleich auf Basis der Rhythmusfamilien und der ausgerichteten Bar Patterns, beide mit sehr hoher Korrelation mit den empirischen Daten, basieren ebenfalls auf der Kenntnis der Taktanfänge, bei einer rein automatisierten Erfassung ist dieser Schritt noch häufig fehleranfällig. Die Onset Patterns kommen komplett ohne Takterkennung aus, die Bar Patterns erfordern eine Abschätzung der Taktlängen. Diese Abschätzung ist jedoch deutlich weniger fehleranfällig und erzielt insbesondere bei rhythmisch prägnanter Musik sehr gute Ergebnisse.

Eine Genreklassifikation auf dem Ballroom-Datensatz konnte die Ergebnisse der Auswertung der Rhythmusvergleichsverfahren auf dem Hörversuchsdatensatz bestätigen. Die Bar Patterns zeigten sehr gute Klassifikationsraten von 96,7 % bei Berücksichtigung annotierter Takte, auch der Vergleich auf Basis von Rhythmusfamilien verschiedener Unterteilung unterscheidet die Rhythmen verschiedener Genres mit Klassifikationsraten von 91,0 % bzw. 91,7 % zuverlässig.

Die Klassifikationsraten von Verfahren, die nicht auf annotierte Informationen zurückgreifen sind erwartungsgemäß etwas geringer, hier erreichen die Bar Patterns mit 89,9 % sogar noch etwas bessere Werte als die Onset Patterns (86,7 %).

Eine Kombination der Onset Patterns mit den Bar Patterns kann für beide Datensätze die Ergebnisse deutlich verbessern, so wird eine Korrelation von 0.73 mit den empirischen Daten und eine korrekte Klassifikationsrate von 92,4 % auf dem Ballroom Datensatz erreicht.

Dies zeigt, dass weder die Bar Patterns noch die Onset Patterns eine vollständige Darstellung aller relevanten Aspekte des Rhythmus bieten, eine einfache Kombination beider Rhythmusdarstellungen enthält jedoch viele redundante Informationen. Das Finden einer kompakten Rhythmusdarstellung, die die nötigen Informationen beider Rhythmusdarstellungen enthält, kann Gegenstand weiterer Untersuchungen sein.

Bei der Extraktion der Bar Patterns wird zunächst das gesamte Musiksignal in einzelne Takte zerlegt und für jeden Takt eine zu den Bar Patterns analoge Darstellung gewonnen. Die guten Ergebnisse des Vergleichs von Bar Patterns lassen darauf schließen, dass diese Darstellung von Takten auch für die Segmentierung von Musik auf rhythmischer Basis verwendet werden kann. So können durch einen Vergleich aller Takte innerhalb eines Musikstückes diejenigen Takte und Bereiche gefunden werden, deren Rhythmus vom Rhythmus der anderen Takte abweicht.

Zuletzt ist auch eine Verwendung der Bar Patterns zur Detektion der Taktanfänge denkbar. Aus der Kreuzkorrelation, die sich als gutes Maß für die Ähnlichkeit zweier Bar Patterns erwies, liefert ebenso die Verschiebung, die nötig ist, dass die Bar Patterns

gut übereinander liegen. Durch den Vergleich mit einer Reihe von bekannten Rhythmen könnte die Verschiebung, die eine maximale Korrelation ergibt, ein Indikator für die Position des Taktanfanges sein. Der Vorteil gegenüber herkömmlichen Verfahren zur Taktanfangserkennung liegt hierbei in der Berücksichtigung des gesamten Taktes, insbesondere der in vielen Musikrichtungen auf der zweiten und vierten Zählzeit erklingenden Snaredrum. Eine Erweiterung um ausführliche harmonische Betrachtung könnte diese Ergebnisse noch verbessern.

A. Anhang

Rhythmusannotation der Hörversuchsmusikausschnitte

```
1 TMEX2BT; (24.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |--x- --x- --x- --x-|
BD    |x--- x--- x--- x---|
Guit  |--x- --x- --x- --x-|
Bass  |x--- x--- x--- x---|

2 TRAINAS; (25.mp3):
HH    |---- ---- ---- ----|
Sn/T  |--xg --xg --xg ggxg|
BD    |x--- x--- x--- x---|
Guit  |--x- --x- --x- --x-|
Bass  |x--- x--- x--- x---|

3 TRAINBT; (26.mp3):
HH    |---- ---- ---- ----|
Sn/T  |--xg --xg --xg ggxg|
BD    |x--- x--- x--- x---|
Guit  |--x- -x-- x--x -gxg|
Bass  |x--- x--- x--- x---|

4 TRAINFP; (27.mp3):
HH    |---- ---- ---- ----|
Sn/T  |ggxg ggxg ggxg ggxg|
BD    |x--- x--- x--- x---|
Guit  |--x- -x-- x--- -gg-|
Bass  |x--- x--- x--- x---|

5 TRAINPF; (28.mp3):
HH    |---- ---- ---- ----|
Sn/T  |ggxg ggxg ggxg ggxg|
BD    |x--- x--- x--- x---|
Guit  |--x- --x- --x- --x-|
Bass  |x--- x--- x--- x---|

6 BZMRCK; (9.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- --x-|
Guit  |--x- --x- --x- --x-|
Bass  |x-x- x-x- x-x- x-x-|
```

```
7 BOLERO2; (1.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- --x- --x-|
BD    |x--- ---- x--- ----|
Guit  |x--- --x- --x- --x-|
Bass  |x--- --x- x--- --x-|

8 BNGGRSL; (5.mp3):
HH    |---- ---- ---- ----|
Sn/T  |--x- x--- --x- x---|
BD    |x--- ---- --x- ----|
Guit  |x--- x--- --x- --x-|
Bass  |x--- --x- x--- ----|

9 HAPN120; (14.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- ----|
Guit  |x--- --x- --x- ----|
Bass  |x--- --x- -xxx x---|

10 SLOBOTS; (19.mp3):
HH    |---- ---- ---- ----|
Sn/T  |g-x- x-g- g-g- x-g-|
BD    |---- ---- ---- ----|
Guit  |---- x--- ---- x---|
Bass  |x--- --x- x--- x-x-|

11 BOSSARK; (6.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- --x- x--- ----|
Guit  |x--- --x- ---- x---|
Bass  |x--- --x- x--- ----|

12 POPMHRN; (16.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- ----|
Guit  |x--- --x- ---- ----|
Bass  |x--- --x- ---- x---|

13 PUPLIFT; (18.mp3):
HH    |xxxx xxxx xxxx xxxx|
Sn/T  |---- ---- ---- x---|
BD    |x--- x--- x--- ----|
Guit  |x-g- g-x- g-g- x-g-|
Bass  |x-x- x--- x-x- x-x-|

14 PSHINIH; (17.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- ----|
Guit  |x-g- g-x- g-g- x-g-|
Bass  |x--- ---- ---- ----|

15 SOFTROK; (20.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- --x- x--- --x-|
Guit  |x-x- --x- ---- x---|
Bass  |x--- --x- x--- x---|

16 AMENROK; (4.mp3):
HH    |x--- x--- x--- x---|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- --x- ----|
Guit  |x-g- --x- g-g- g-g-|
Bass  |x--- --x- --x- x-x-|
```

```
17 BRITPOP; (7.mp3):
HH    |x-x- x-x- x-x- x-xx|
Sn/T  |---- x--- ---- x---|
BD    |x--- --x- ---- ----|
Guit  |x-g- g-x- g-g- x-g-|
Bass  |x--- --x- --x- x---|

18 BRSOLC3; (8.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- ----|
Guit  |x-x- ---- x-x- ----|
Bass  |x-x- --x- x-x- x---|

19 JAZROK1; (15.mp3):
HH    |b--- ---- b--- ----|
Sn/T  |---- x-g- ---- x-g-|
BD    |x--- ---- x-x- ----|
Guit  |x--- ---- ---- ----|
Bass  |x--- --x- x--- x---|

20 STZ_120; (21.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- --x- x--- ----|
Guit  |x-x- g-g- x-g- x-g-|
Bass  |x-x- x-x- x-x- x-x-|

21 TBLZSBB; (23.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- ----|
Guit  |x--- x--- x--- x-xx|
Bass  |x--- --x- x--- --x-|

22 80MPBBP; (2.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- ---- ---- x---|
BD    |x--- x--- x--- x---|
Guit  |x--- x-x- --x- ----|
Bass  |x--- x--- x--- x---|

23 ELPJHSS; (10.mp3):
HH    |x-x- x-x- x-x- x-x-|
Sn/T  |---- x--x ---- x---|
BD    |x--- ---- x-x- --x-|
Guit  |x--- ---- x--- ----|
Bass  |x--- ---g x-x- --x-|

24 SUGA_SX; (22.mp3):
HH    |x-xg x-xg x-o- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--x ---- --x- ----|
Guit  |x--- x--- --x- x-x-|
Bass  |x-x- x-x- x-xg x-x-|

25 ACFNK16; (3.mp3):
HH    |x-x- x-x- x-o- x-x-|
Sn/T  |---- x--- ---- x---|
BD    |x--x ---- --x- ----|
Guit  |xggg x-gx ---- xg--|
Bass  |x--- x-x- --x- x-x-|

26 FNKSYN2; (11.mp3):
HH    |x-xg x-xg x-xg x-xg|
Sn/T  |---- x--- ---- x---|
BD    |x--- ---- x--- --x-|
Guit  |x--- x--- x--- --x-|
Bass  |x--- ---x x-x- x-x-|
```

```
27 FNKSYN3; (12.mp3):
HH    |x-x- x-x- x-o- x-o-|
Sn/T  |---- x--x ---- x---|
BD    |x--- ---- --x- --x-|
Guit  |x--- x--- x--- x---|
Bass  |x--- x--- x--- x-x-|
```

```
28 FNKSYN4; (13.mp3):
HH    |x-g- x-g- x-o- x-g-|
Sn/T  |---- x--- -g-- x---|
BD    |x--- ---- --x- --x-|
Guit  |xg-- xg-- x--- gggg|
Bass  |x--- x--- xxx- x-x-|
```

Daten CD

```
./
├── fulltext.pdf ..... Arbeit im Volltext
├── README
├── scriptEvaluateSurvey.m
├── loadParameters.m
├── scriptEvaluateBallroom.m
├── scriptEvaluateTatumAndBarLength.m
├── getEmpiricalDistanceMatrix.m
├── datasets ..... verwendete Datensätze
│   ├── ballroom
│   │   ├── annotations
│   │   └── audio
│   └── gtzan
│       ├── annotations
│       └── audio
├── literatur ..... verwendete Literatur (Volltext)
│   └── bibliography.bib
├── survey ..... Online Survey Ergebnisse
├── rhythmicRepresentations ..... Rhythmusextraktion
│   ├── getFluctuationPatterns.m
│   ├── getRhythmicPatterns.m
│   ├── getBarPatterns.m
│   ├── getOnsetPatterns.m
│   ├── getRhythmFamily.m
│   ├── barPatternsFunctions
│   └── MA-toolboxElias-Pampalk
└── toolsAndFunctions
```

Literatur

Behrends, Ehrhard (2000). *Introduction to Markov chains*. Bd. 228. Wiesbaden: Springer.

Bregman, Albert S. (1994). *Auditory scene analysis: The perceptual organization of sound*. Cambridge: MIT press.

Bregman, Albert S. et al. (1985). „Spectral integration based on common amplitude modulation“. In: *Perception & Psychophysics* 37.5, S. 483–493.

Bruhn, Herbert (2000). „Rhythmus. Ein interdisziplinäres Handbuch.“ In: Bern: Huber. Kap. Zur Definition von Rhythmus, S. 41–56.

Cao, Erica, Max Lotstein und Philip N. Johnson-Laird (2014). „Similarity and Families of Musical Rhythms“. In: *Music Perception* 31, S. 444–469.

Carlyon, Robert P. und S. McAdams (1992). „The Psychophysics of Concurrent Sound Segregation“. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 336.1278, S. 347–355.

Cha, Sung-Hyuk (2007). „Comprehensive survey on distance/similarity measures between probability density functions“. In: *International Journal of Mathematical Models and Methods in Applied Sciences* 1.2, S. 1.

Clarke, Eric F. (1985). „Musical structure and cognition“. In: Hrsg. von Peter Howell, Ian Cross und Robert West. London: Academic Press. Kap. Structure and expression in rhythmic performance, S. 209–236.

Clarke, Eric F. (1999). „The psychology of music“. In: Hrsg. von Diana Deutsch. 2. Aufl. New York: Academic Press. Kap. Rhythm and timing in music, S. 473–500.

Cooper, Grosvenor und Leonard B. Meyer (1963). *The rhythmic structure of music*. Bd. 118. University of Chicago Press.

Darwin, C. J. (2005). „Pitch: Neural coding and perception“. In: *Pitch: Neural coding and perception*. Hrsg. von Christopher J. Plack et al. New York: Springer. Kap. Pitch and auditory grouping, S. 278–305.

Darwin, C. J. und Valter Ciocca (1992). „Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component“. In: *The Journal of the Acoustical Society of America* 91.6, S. 3381–3390.

Deutsch, Diana (1999). „The psychology of music“. In: Hrsg. von Diana Deutsch. 2. Aufl. San Diego: Academic Press. Kap. Grouping Mechanisms in Music, S. 299–348.

Dixon, Simon, Fabien Gouyon und Gerhard Widmer (2004). „Towards Characterisation of Music via Rhythmic Patterns“. In: International Society for Music Information Retrieval.

Esparza, Tlacael Miguel, Juan Pablo Bello und Eric J. Humphrey (2016). „From Genre Classification to Rhythm Similarity: Computational and Musicological Insights“. In: *Journal of New Music Research* 44, S. 39–57.

Fischinger, Timo (2009). *Zur Psychologie des Rhythmus: Präzision und Synchronisation bei Schlagzeugern.* Kassel university press GmbH.

Foote, Jonathan und Shingo Uchihashi (2001). „The Beat Spectrum: A new approach to rhythm analysis“. In: *IEEE International Conference on Multimedia and Expo.* Tokyo.

Forney, G. David (1973). „The Viterbi algorithm“. In: *Proceedings of the IEEE* 61.3, S. 268–278.

Fraisse, Paul (1982). „The Psychology of Music“. In: Hrsg. von Diana Deutsch. 1. Aufl. Bd. 1. New York: Academic Press. Kap. Rhythm and Tempo, S. 149–180.

Fujishima, Takuya (1999). „Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music.“ In: International Computer Music Conference, S. 464–467.

Goldstone, Robert L. (1994). „The role of similarity in categorization: Providing a groundwork“. In: *Cognition* 52.2, S. 125–157.

Gómez-Marin, Daniel, Sergi Jorda und Perfecto Herrera (2015). „Pad and sad: two awareness-weighted rhythmic similarity distances“. In: International Society for Music Information Retrieval Conference. Málaga, Spain.

Gómez-Marín, Daniel, Sergi Jorda und Perfecto Herrera (2015). „Strictly Rhythm: Exploring the effects of identical regions and meter induction in rhythmic similarity perception“. In: International Symposium on Computer Music Multidisciplinary Research, S. 449–463.

Gouyon, Fabien (2005). „A computational approach to rhythm description“. Diss. Universitat Pompeu Fabra.

Gouyon, Fabien et al. (2006). „An experimental comparison of audio tempo induction algorithms“. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.5, S. 1832–1844.

Gruhne, Matthias, Christian Dittmar und Daniel Gaertner (2009). „Improving rhythmic similarity computation by beat histogram transformations.“ In: International Society for Music Information Retrieval, S. 177–182.

Hampton, James A. (1998). „Similarity-based categorization and fuzziness of natural categories“. In: *Cognition* 65.2, S. 137–165.

Handel, Stephen (1993). *Listening: An introduction to the perception of auditory events.* Cambridge, MA, US: The MIT Press.

Holzapfel, André und Yannis Stylianou (2011). „Scale transform in rhythmic similarity of music“. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.1, S. 176–185.

Honing, Henkjan (2012). „Without it no music: beat induction as a fundamental musical trait“. In: *Annals of the New York Academy of Sciences* 1252.1, S. 85–91.

Houtsma, Adrianus J.M. (1995). „Pitch Perception“. In: *Hearing - Handbook of Perception and Cognition.* Hrsg. von B.J.C. Moore. 2nd Edition. San Diego, CA: Academic Press, S. 267–295.

Huron, David und Matthew Royal (1996). „What is melodic accent? Converging evidence from musical practice“. In: *Music Perception: An Interdisciplinary Journal* 13.4, S. 489–516.

Jensen, Jesper Højvang, Mads Grœsbøll Christensen und Søren Holdt Jensen (2009). „A tempo-insensitive representation of rhythmic patterns“. In: *Signal Processing Conference.* IEEE, S. 1509–1512.

Jensen, Kristoffer (2004). „A causal rhythm grouping“. In: *International Symposium on Computer Music Modeling and Retrieval.* Berlin: Springer, S. 83–95.

Krebs, Florian, Sebastian Böck und Gerhard Widmer (2013). „Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio“. In: *ISMIR*, S. 227–232.

Lawson, Charles L. und Richard J. Hanson (1974). *Solving least squares problems.* Englewood Cliffs: Prentice-Hall.

Leman, Marc (1995). *Music and schema theory: Cognitive foundations of systematic musicology.* Berlin: Springer.

Lerch, Alexander (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics.* Hoboken, New Jersey: John Wiley & Sons.

Lin, Dekang et al. (1998). „An information-theoretic definition of similarity“. In: *International Conference on Machine Learning.* 1998, S. 296–304.

Mandel, Michael I. und Dan Ellis (2005). „Song-Level Features and Support Vector Machines for Music Classification.“ In: *Proceedings of the International Society for Music Information Retrieval Conference*, S. 594–599.

Marchand, Ugo, Quentin Fresnel und Geoffroy Peeters (2015). „GTZAN-Rhythm: Extending the GTZAN Test-Set with Beat, Downbeat and Swing Annotations“. In: *Extended abstracts for the Late-Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference.* Malaga.

McKinney, Martin und Jeroen Breebaart (2003). „Features for audio and music classification“. In: *Proceedings of the International Symposium on Music Information Retrieval*. Baltimore, S. 151–158.

Mongeau, Marcel und David Sankoff (1990). „Comparison of musical sequences“. In: *Computers and the Humanities* 24, S. 161–175.

Moore, Brian CJ (2012). *An introduction to the psychology of hearing*. 6. Aufl. Leiden: Koninklijke Brill NV.

Moore, Brian CJ (1995). „Hearing“. In: Hrsg. von Brian CJ Moore. London: Academic Press. Kap. Frequency analysis and masking, S. 161–205.

Möser, Michael (2009). *Technische Akustik*. 9. Aufl. Springer.

Pampalk, Elias (2006). „Computational models of music similarity and their application in music information retrieval“. Diss. Technische Universität Wien.

Pampalk, Elias (2001). „Islands of music: Analysis, organization, and visualization of music archives“. Magisterarb. Wien: Technischen Universität Wien.

Patel, Aniruddh D. et al. (2009). „Experimental evidence for synchronization to a musical beat in a nonhuman animal“. In: *Current biology* 19.10, S. 827–830.

Plack, Christopher J. und Robert P. Carlyon (1995). „Hearing“. In: Hrsg. von Brian Moore. London: Academic Press. Kap. Loudness perception and intensity coding, S. 123–160.

Pohle, Tim et al. (2009). „On Rhythm and General Music Similarity“. In: *Proceedings of the International Society for Music Information Retrieval Conference*, S. 525–530.

Pöppel, Ernst (1997). *Grenzen des Bewußtseins: wie kommen wir zur Zeit, und wie entsteht Wirklichkeit?* Frankfurt a.M.: Insel-Verlag.

Povel, Dirk-Jan (1981). „Internal representation of simple temporal patterns“. In: *Journal of Experimental Psychology: Human Perception and Performance* 7.1, S. 3–18.

Rabiner, Lawrence und B. Juang (1986). „An introduction to Hidden Markov Models“. In: *IEEE Transactions on Acoustics, Speech and Signal Processing magazine* 3.1, S. 4–16.

Rasch, Rudolf und Reinier Plomp (1999). „The psychology of music“. In: Hrsg. von Diana Deutsch. 2. Aufl. San Diego, CA: Academic Press. Kap. The perception of musical tones, S. 89–112.

Risset, Jean-Claude und David L. Wessel (1999). „The psychology of music“. In: Hrsg. von Diana Deutsch. San Diego: Academic Press. Kap. Exploration of timbre by analysis and synthesis, S. 113–169.

Snyder, Bob (2000). *Music and memory: An introduction*. Massachusetts: MIT press.

Todd, Neil P. McAngus (1994). „The auditory “primal sketch”: A multiscale model of rhythmic grouping“. In: *Journal of new music Research* 23.1, S. 25–70.

Toussaint, Godfried T. (2004). „A Comparison of Rhythmic Similarity Measures“. In: *Proceedings of the International Societyon Music Information Retrieval Conference*. Barcelona, S. 242–245.

Tzanetakis, George und Perry Cook (2002). „Musical genre classification of audio signals“. In: *IEEE Transactions on Speech and Audio Processing* 10.5, S. 293–302.

Weinzierl, Stefan (2008). *Handbuch der Audiotechnik*. Berlin: Springer.

Zwicker, Eberhard und Hugo Fastl (2006). *Psychoacoustics: Facts and models*. Bd. 22. Berlin: Springer.