Strukturanalyse für Musiksignale

Meinard Müller[1], Nanzhu Jiang[1], Harald Grohganz[2], Michael Clausen[2]

[1]International Audio Laboratories Erlangen*

meinard.mueller@audiolabs-erlangen.de, nanzhu.jiang@audiolabs-erlangen.de

[2]Institut für Informatik, Universität Bonn

grohganz@cs.uni-bonn.de, clausen@cs.uni-bonn.de

Abstract: Bei der automatisierten Verarbeitung von Musiksignalen steht man aufgrund der Vielfältigkeit von Musik in Form und Inhalt vor großen Herausforderungen. Dieser Artikel gibt einen Überblick über unterschiedliche Aspekte der Segmentierung und Strukturierung von Musiksignalen. Hierbei gehen wir zum einen auf unterschiedliche musikalische Dimensionen wie Zeit, Rhythmus, Dynamik, Harmonie und Klangfarbe und zum anderen auf unterschiedliche Segmentierungsprinzipien wie Wiederholung, Homogenität und Novelty ein. Neben diesen Aspekten kann bei der Analyse auch ausgenutzt werden, dass ein Musikstück oft in unterschiedlichen Darstellungsformen und Versionen vorliegt, deren simultane Betrachtung zu einer Stabilisierung der automatisch berechneten Segmentierungsergebnisse führen kann. Zur Illustration der verschiedenen Strukturierungsaspekte stellen wir im zweiten Teil dieses Artikels einige konkrete Verfahren zur robusten und adaptiven Segmentierung vor und diskutieren zukünftige Herausforderungen.

1 Einleitung

Segmentierung und Strukturierung sind für die automatisierte Verarbeitung von Musiksignalen von grundlegender Bedeutung. Bei der *Segmentierung* geht es grob gesprochen um die Zerlegung eines Audiodatenstroms in inhaltlich sinnvolle Abschnitte und elementare Einheiten. Hierauf aufbauend werden bei der *Strukturierung* diese Abschnitte nach bestimmten Kriterien bezüglich ihrer Bedeutung oder Funktion semantischen Kategorien zugeordnet. Eine solche Strukturierung kann sich zum Beispiel auf die musikalische Form eines Musikstücks beziehen. Im Fall von Popmusik ist hierbei eine Audioaufnahme in Blöcke entsprechend der Intro (Einleitung), den Strophen, den Refrains, und der Outro zu segmentieren. Oder im Fall einer klassischen Sonate beziehen sich die Blöcke auf Exposition, Durchführung, Reprise und Coda. Musikalische Formen werden häufig durch Abfolgen von indizierten Buchstaben wie zum Beispiel $A_1 A_2 B_1 B_2 C A_3 B_3 B_4$ beschrieben, siehe auch Abbildung 2e. Hierbei beziehen sich gleiche Buchstaben auf sich wiederholende Blöcke und Indizes auf die jeweiligen Positionen der Wiederholungen. Obige

*Die International Audio Laboratories Erlangen sind eine gemeinsamen Einrichtung der Friedrich-Alexander-Universität Erlangen-Nürnberg und des Fraunhofer-Instituts für Integrierte Schaltungen IIS.

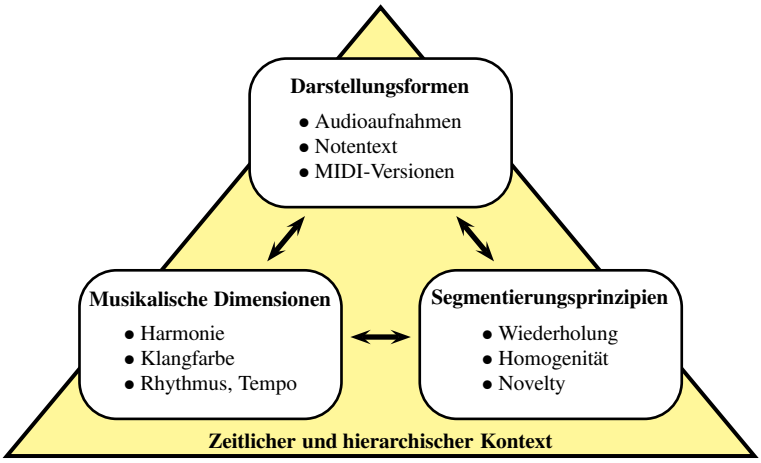

Abbildung 1: Schematische Darstellungen unterschiedlicher Aspekte und Prinzipien, die für die Segmentierung und Strukturierung von Musiksignalen grundlegend sind.

Abfolge besagt also, dass das zugrundeliegende Musikstück (hier: Ungarischer Tanz Nr. 5 von Johannes Brahms) aus drei sich wiederholenden A-Teilen A_1, A_2 und A_3, aus vier sich wiederholenden B-Teilen B_1, B_2, B_3 und B_4, sowie einem Mittelteil C besteht. Im Allgemeinen muss man bei Strukturierung von Musik ganz unterschiedliche zeitliche Stufen berücksichtigen, die oft hierarchisch angeordnet werden können. So können die Teile einer musikalischen Form häufig weiter untergliedert werden, indem man prägnante, sich wiederholende Ton- oder Akkordfolgen berücksichtigt. Diese können zum Beispiel ein Riff in Popmusik oder musikalische Themen und Motive im Fall klassischer Musik sein. Auf einer noch feineren zeitlichen Stufe können dann einzelne Akkorde, Töne, oder Noteneinsatzzeiten betrachtet werden.

Die Segmentierung und Strukturierung stellen oft den ersten Schritt für eine anschließende Weiterverarbeitung der Musiksignale dar, wie beispielsweise eine Klassifizierung, Annotation oder Indexierung. Diese Aufgaben sind zentrale Fragestellungen des *Music Information Retrieval* (MIR), eines noch relativ jungen Forschungsgebiets. Allgemeine Ziele dieses Gebiets liegen in der Entwicklung von Methoden und Systemen, die Benutzern große, in digitaler Form vorliegende Musikkollektionen in vielfältiger Weise zugänglich machen. MIR stellt ein interdisziplinäres Forschungsgebiet dar, das eine Vielzahl von Disziplinen wie z. B. die Informatik, Signalverarbeitung, Musikwissenschaft oder Bibliothekswissenschaft einschließt. Für einen allgemeinen Überblick über die beteiligten Disziplinen, über zentrale MIR-Fragestellungen und über bestehende MIR-Systeme verweisen wir auf die folgenden Überblicksartikel und Bücher [CVG+08, Dow03, KD06, Mül07, Ori06, TWV05]. Eine umfangreiche Sammlung an Forschungsartikeln, die den aktuellen Stand der MIR-Forschung repräsentieren, stellen die Sammelbände[1] der jährlich stattfindende Konferenz der *International Society on Music Information Retrieval* (ISMIR) dar.

[1] Die ISMIR-Sammelbände sind online auf der Webseite `http://www.ismir.net/` frei erhältlich.

In diesem Artikel wollen wir allgemeine musikalische Aspekte und Prinzipien diskutieren, die bei der Segmentierung und Strukturierung von Musiksignalen von Bedeutung sind, siehe auch Abbildung 1. Hierbei geht es uns weniger darum, einen umfassenden Überblick über relevante Arbeiten in diesem Bereich zu geben – hierzu sei zum Beispiel auf die Artikel [DG08, PMK10] verwiesen; vielmehr wollen wir anhand ausgewählter Arbeiten aktuelle Methoden der automatisierten Strukturanalyse skizzieren und hierbei auf einige, sich aus der Komplexität von Musiksignalen ergebenden Herausforderungen eingehen.

Die folgenden Abschnitte dieses Artikels sind wie folgt gegliedert: Zunächst diskutieren wir grundlegende Prinzipien zur Segmentierung und Strukturierung von Musiksignalen und erläutern deren musikalische Relevanz (Abschnitt 2). Dann gehen wir darauf ein, wie sich unterschiedliche musikalische Aspekte durch geeignete Merkmalsdarstellungen erfassen lassen (Abschnitt 3) und wie sich diese strukturell in Selbstähnlichkeitsmatrizen widerspiegeln (Abschnitt 4). Um das Zusammenspiel der unterschiedlichen Aspekte zu beleuchten, gehen wir dann in den folgenden vier Abschnitten auf unterschiedliche Strukturierungsalgorithmen ein. Der Artikel schließt mit einer Zusammenfassung und einem Ausblick (Abschnitt 8).

2 Prinzipien der Segmentierung und Strukturierung

In der Literatur findet man zahlreiche Verfahren zur automatisierten Strukturanalyse, die auf ganz unterschiedlichen Annahmen basieren [DG08, PMK10, Pee04]. Im Folgenden wollen wir die wesentlichen zugrundeliegenden Prinzipien und deren musikalischen Relevanz diskutieren. Hierbei folgen wir der in [PMK10] vorgeschlagenen Terminologie.

Bei *Novelty-basierten Verfahren* zur Segmentierung geht es um die Erkennung von Zeitpunkten oder Übergangsbereichen, in denen neuartige Signaleigenschaften auftreten [BDA+05, FC03]. Dies können zum einen plötzliche energie- und spektralbasierte Signaländerungen sein, wie sie zum Beispiel in der Attack-Phase beim Anspielen einer Note auftreten [BDA+05]. Zum anderen können dies auch eher glattere Signalübergänge sein, die zum Beispiel in Form von Klangfarbenwechseln durch Änderungen in der Instrumentierung auftreten. Novelty-basierte Verfahren können als Spezialfall von *ereignisbasierten Verfahren* angesehen werden, bei denen es um die unüberwachte bzw. überwachte Detektion statistisch hervorstechender bzw. als Vorwissen spezifizierter Ereignisse im Audiodatenstrom geht. Eine weitere wichtige Klasse von Segmentierungsverfahren stellen *homogenitätsbasierte Verfahren* dar, bei denen es um eine Unterteilung eines Audiodatenstroms in Abschnitte geht, die jeweils in sich bezüglich eines ausgezeichneten Aspekts homogen sind [LS08]. Die Homogenität kann sich zum Beispiel auf die Klangfarbe, die Dynamik oder Harmonik beziehen. Schließlich ist die Erkennung wiederkehrender Muster das Ziel von *wiederholungsbasierten Verfahren* [Cha06, Got06, Ong07, MXKS04, Pee07], die eine zentrale Rolle bei der Segmentierung von Musiksignalen spielen.

Alle genannten Segmentierungsparadigmen spielen im Musikbereich eine zentrale Rolle und spiegeln sich unmittelbar in musikalischen Gestaltungsprinzipien wider. Um dies zu erklären, wollen wir im Folgenden von der relativ stark vereinfachten Sichtweise aus-

gehen, dass sich die musikalische Form einer Musikaufnahme auf den groben zeitlichen Ablauf musikalisch sinnvoller Blöcke bezieht. Die musikalische Form ist sowohl für das Verständnis als auch die Erschließung von Musik von großer Bedeutung und steht mit der Gattung und Funktion eines Musikstücks in enger Beziehung [Lei87]. Die in der Abfolge auftretenden Blöcke stehen zueinander oft in gewissen Beziehungen, die einem festen Schema folgen und sich damit ins Gedächtnis einprägen können [NG98]. Im wesentlichen folgen diese Beziehungen drei Gestaltungsprinzipien. Bei der *Wiederholung* werden Gedanken und Teile mehr oder weniger unverändert aufgegriffen. Die entsprechenden Blöcke gleichen sich dann einander bezüglich bestimmter Aspekte wie der Melodik, Harmonik oder Rhythmik. Über solche wiederkehrenden Muster wird ein zeitlicher Bezug innerhalb des Stücks hergestellt, der vom Hörer nachvollzogen und das Gefühl der Vertrautheit und des musikalischen Verstehens hervorrufen kann. Als zweites Gestaltungsprinzip dient der *Kontrast*, bei dem zwei Blöcke unterschiedlichen Charakters aufeinandertreffen. Zum Beispiel folgt ein lauter Abschnitt einem leiseren, ein langsamer einem schnellen oder ein kammermusikalischer einem orchestralen [Mic]. Durch die bewusste Konstruktion eines Bruches wird eine kontrastierende Wirkung erzielt, die der Hörer quasi als Gegenstück zur Wiederholung als überraschendes Element erlebt und die dem Stück Farbigkeit verleiht. Das dritte Gestaltungsprinzip ist das der *Variation*. Hierbei werden Gedanken und Teile in abgewandelter Form aufgegriffen, wobei das Original immer noch durchscheint und wiedererkennbar bleibt. Die Abwandlungen können dabei unterschiedliche musikalische Aspekte wie Klangfarbe, Dynamik, Harmonik oder Polyphonie betreffen.

Das zentrale Ziel der Strukturanalyse von Musiksignalen besteht darin, mit automatisierten Methoden die grobe musikalische Form direkt aus den Audioaufnahmen eines Musikstücks zu bestimmen. Um den unterschiedlichen Gestaltungsprinzipien gerecht zu werden, sind alle Segmentierungsmethoden von zentraler Rolle. *Wiederholungsbasierte Verfahren* werden benötigt, um wiederkehrende Muster zu identifizieren. Durch *Noveltybasierte Verfahren* sind Übergänge zwischen konstrastierenden Blöcken aufzuspüren. *Homogenitätsbasierte Verfahren* dienen dazu, rhythmisch, harmonisch oder klanglich konsistente Bereiche zu erfassen, über die häufig Varianten identifiziert werden können (z. B. melodische Variationen über ähnlichen Harmonieabfolgen).

3 Musikalische Aspekte

Wie oben dargestellt können sich die Segmentierungsprinzipien auf ganz unterschiedliche musikalische Aspekte beziehen. So kann ein Segment homogen bezüglich der Klangfarbe, der Dynamik oder der Harmonik sein. Auf der anderen Seite können sich wiederholende Segmente durch ähnliche Harmonieverläufe auszeichnen, aber sich hinsichtlich der Instrumentierung oder Dynamik erheblich unterscheiden. Weiterhin sind Wiederholungsstrukturen auf einer zeitlich feingranularen Stufe oft mit rhythmischen Aspekten korreliert. Die unterschiedlichen musikalischen Aspekten spiegeln sich in den akustischen Eigenschaften eines als Wellenform gegebenen Musiksignals wider. Auch wenn der Mensch beim Hören einer Audioaufnahme oft sofort die relevanten Aspekte wahrnehmen kann, ist die automatische Extraktion musikalisch sinnvoller Merkmale aus Wellenformdaten eine im allge-

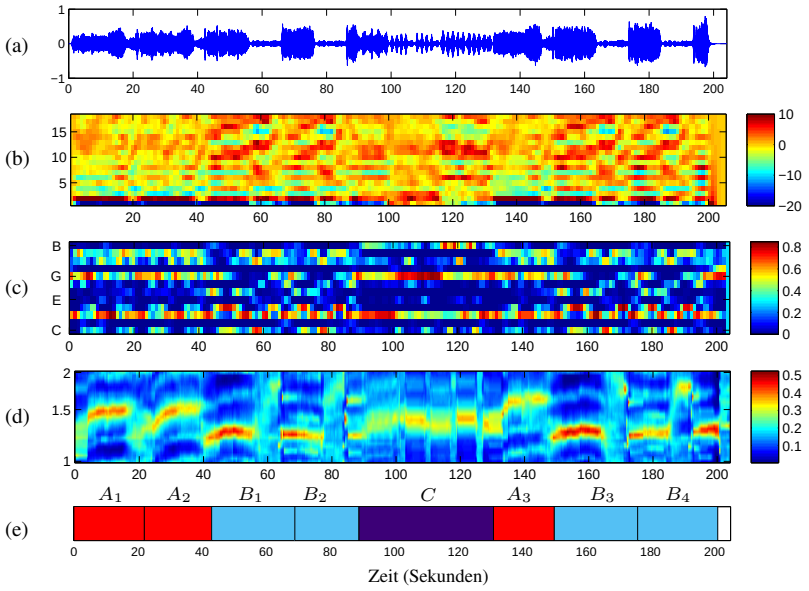

Abbildung 2: Merkmalsdarstellungen für eine Aufnahme des Ungarischen Tanzes Nr. 5 von Johannes Brahms. (a) Wellenform. (b) MFCC-basierte Merkmale. (c) Chroma-basierte Merkmale. (d) Tempo-basierte Merkmale.

meinen sehr schwierige Aufgabenstellung. Ein zentrales Ziel der Musiksignalverarbeitung besteht darin, ein gegebenes Musiksignal in geeignete Merkmalsdarstellungen zu transformieren, die zu unterschiedlichen musikalischen Aspekten korrelieren, siehe [MEKR11]. Im Folgenden wollen wir uns exemplarisch drei solchen Merkmalsdarstellungen zuwenden, siehe auch Abbildung 2.

Zum einen spielt bei der Strukturierung von Musiksignalen die *Instrumentierung* und *Klangfarbe* eine wichtige Rolle [BMK06]. Hierbei steht die Klangfarbe häufig mit der Energieverteilung und deren zeitlichen Entwicklung in sogenannten kritischen Spektralbändern in Beziehung. Bei der Analyse von Musiksignalen wird daher häufig auf als MFCCs (Mel Frequency Cepstral Coefficients) bekannte Merkmale zurückgegriffen, die ursprünglich für die Spracherkennung entwickelt wurden [DM90]. Nach Transformation des Musiksignals in eine Spektraldarstellung werden MFCC-basierte Merkmale durch Zusammenfassen geeigneter Frequenzbänder in perzeptuell motivierte Mel-Bänder und Anwendung einer dekorrelierenden Kosinustransformation gewonnen. Insbesondere die unteren MFCCs beschreiben die grobe Form der spektralen Einhüllenden, die wiederum zur Klangfarbe korreliert [TSB05]. Abbildung 2b zeigt für unser Brahms-Beispiel die MFCC-basierte Merkmalsfolge für die unteren 18 Koeffizienten. Die Merkmalsdarstellung spiegelt zum Beispiel wider, dass sich die A-Teile deutlich von den B-Teilen hinsichtlich der Klangfarbe unterscheiden.

Während sich MFCC-basierte Merkmale typischerweise für homogenitätsbasierte Segmentierungsaufgaben eignen, sind sie für wiederholungsbasierte Verfahren meist unge-

eignet. Ein Grund hierfür ist die Tatsache, dass melodisch sich wiederholende Passagen erhebliche Unterschiede in Bezug auf die Instrumentierung oder Klangfarbe aufweisen können. Man denke zum Beispiel an eine Strophe mit Gesang, die später nochmals als reine Instrumentalversion erscheint. Grundlage der wiederholungssbasierten Strukturanalyse sind daher oft *Chroma-basierte Merkmalsdarstellungen*, die stark mit dem *Harmonieverlauf* des zugrundeliegenden Musikstücks korrelieren und ein hohes Maß an Invarianz bezüglich Änderungen in der Klangfarbe aufweisen [BW05, G06, Mül07]. Ähnlich wie die MFCCs können Chromamerkmale aus einer Spektraldarstellung des Musiksignals abgeleitet werden. Hierbei werden allerdings die Frequenzbänder in musikalisch motivierte Tonhöhenbänder (gemäß der temperierten Stimmung entsprechend einer Klaviertastatur) zerlegt. Jede Tonhöhe kann eindeutig durch einen der zwölf Chromawerte $\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{H}$ und seine Oktavlage beschrieben werden. Im nächsten Schritt werden alle zum gleichen Chroma korrespondierenden Tonhöhenbänder zu einem Chromaband aufsummiert. Zum Beispiel werden die Energiewerte der Bänder zu den Tonhöhen $\mathrm{A0}, \mathrm{A1}, \ldots, \mathrm{A7}$ zu einem Energiewert zum Chroma A zusammengefasst. Nach einem anschließenden Normalisierungsschritt erhält man schließlich eine Folge von 12-dimensionalen Chromavektoren, wobei jeder Vektor die lokale Energieverteilungen der im Audiosignal vorkommenden Frequenzen auf die 12 Chromabänder widerspiegelt. So zeigt zum Beispiel in Abbildung 2c die *Chroma-basierte Merkmalsfolge* (oder auch *Chromagramm*[2]), dass der C-Teil harmonisch gesehen relativ homogen ist.

Neben der Klangfarbe und Harmonik stellen *Tempo* (Schläge pro Minute) und *Rhythmus* weitere wichtige Aspekte der Musik dar. Zur Erfassung solcher Eigenschaften gehen die meisten Verfahren in zwei Schritten vor. Im ersten Schritt werden Noteneinsatzkandidaten bestimmt. Hierbei nutzt man die Eigenschaft aus, dass das Anspielen von Noten meist mit einer meßbaren Änderung in der Signalenergie und der Frequenzzusammensetzung einhergeht [BDA$^+$05]. Im zweiten Schritt werden die Noteneinsatzkandidaten dann hinsichtlich periodischer Muster untersucht, aus denen sich dann eine *Tempo-basierte Merkmalsdarstellung* (oder auch *Tempogramm*[3]) ableiten lässt [Pee05, GM11]. Ähnlich wie bei den zyklischen Chromagrammen, bei denen sich um Oktaven unterscheidende Tonhöhen identifiziert werden, können auch zyklische Tempogramme betrachtet werden, bei denen sich um eine Zweier-Potenz unterscheidende Tempi identifiziert werden [GMK10]. Dies führt auf robuste Merkmalsdarstellungen, die lokale Tempounterschiede erfassen können. Das in Abbildung 2d dargestellte zyklische Tempogramm legt zum Beispiel die erheblichen Tempounterschiede zwischen den A-Teilen und B-Teilen offen.

4 Selbstähnlichkeitsmatrizen

Die Umwandlung eines Musiksignals in eine geeignete Merkmalsdarstellung stellt den ersten Schritt quasi jedes Strukturanalyseverfahrens dar. Hierbei hat der Typ der verwendeten

[2] Unterschiedliche Varianten Chroma-basierter Merkmale sind Teil der unter `www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/` frei erhältlichen Chroma Toolbox , siehe auch [ME11].

[3] Tempo-basierter Merkmale sind Teil der unter `http://www.mpi-inf.mpg.de/resources/MIR/tempogramtoolbox/` frei erhältlichen Tempogram Toolbox.

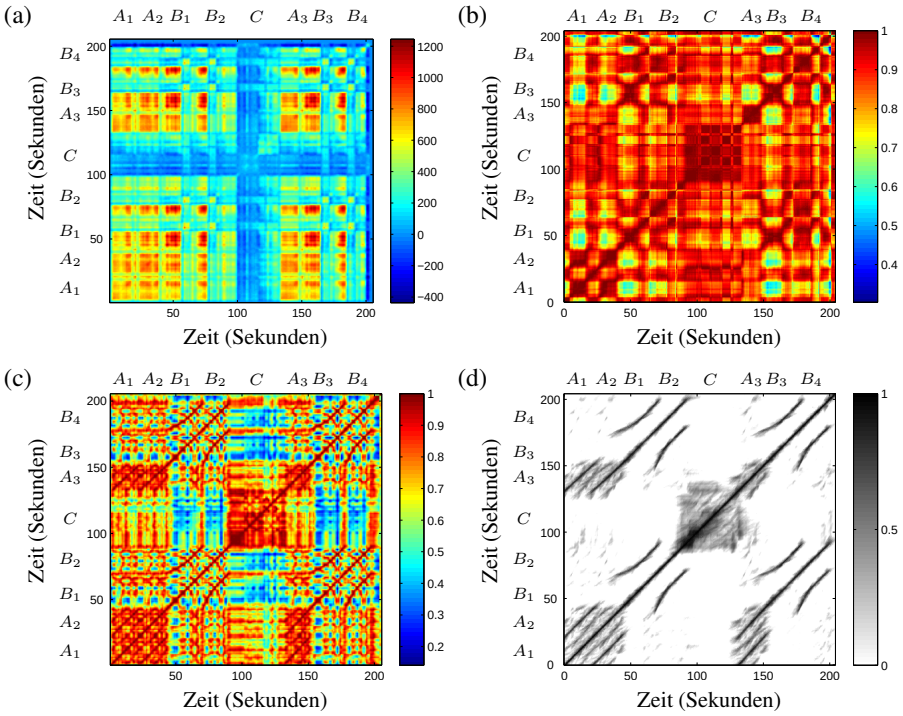

Abbildung 3: Selbstähnlichkeitsmatrizen für die in Abbildung 2 dargestellten Merkmale: **(a)** MFCC-basierte Merkmale, **(b)** Tempo-basierte Merkmale, **(c)** Chroma-basierte Merkmale. **(d)** zeigt eine strukturell verbesserte Version der Selbstähnlichkeitsmatrix aus (c).

Merkmalsdarstellung einen erheblichen Einfluss auf die erkennbaren Strukturen. Diese Tatsache spiegelt sich auch in sogenannten *Selbstähnlichkeitsmatrizen*[4] wider, auf die wir im Folgenden näher eingehen wollen. Wie bereits erwähnt, wird in den meisten Ansätzen zur automatisierten Strukturanalyse das Musiksignal in eine Folge von Merkmalsvektoren transformiert. Unter Verwendung eines geeigneten Ähnlichkeitsmaßes werden diese Vektoren dann paarweise verglichen. Die resultierenden Ähnlichkeitswerte können in einer quadratischen Selbstähnlichkeitsmatrix erfasst und durch eine geeignete Farbkodierung (z. B. dunkle Farbe für große und helle Farbe für kleine Werte) bildlich dargestellt werden. Abbildung 3 zeigt einige Selbstähnlichkeitsmatrizen, die auf den in Abbildung 2 dargestellten Merkmalsdarstellungen basieren.

Die strukturellen Bezüge innerhalb einer Merkmalsfolge werden in der resultierenden Selbstähnlichkeitsmatrix sichtbar. Grob kann man dabei zwischen zwei unterschiedlichen Mustern unterscheiden [DG08, PMK10]. Wenn sich zum einen die Elemente der Merkmalsfolge über einen gewissen Zeitraum nur wenig unterscheiden, so entspricht dies ei-

[4] Häufig findet man in der Literatur auch den Begriff des *Rekurrenzplots*.

nem homogenen Segment. Ein paarweiser Vergleich dieser Elemente führt durchgängig zu großen Ähnlichkeitswerten, so dass in der resultierenden Selbstähnlichkeitsmatrix ein quadratischer *Block* sichtbar wird. Wenn zum anderen die Merkmalsfolge zwei sich wiederholende Teilfolgen enthält, dann sind die sich entsprechenden Elemente der beiden Teilfolgen ähnlich. Im allgemeinen sind aber die jeweiligen Teilfolgen in sich nicht notwendigerweise homogen. Anstelle eines Blocks wird damit in der Selbstähnlichkeitsmatrix ein *Pfad* sichtbar, dessen Projektionen auf die beiden Achsen den beiden Teilfolgen entsprechen.

Als Illustration betrachten wir wieder die Selbstähnlichkeitsmatrizen in Abbildung 3. Deutliche Blockstrukturen sind zum Beispiel im Fall der Tempo-Selbstähnlichkeitsmatrix (Abbildung 3b) in denjenigen Passagen erkennbar, wo das Tempo mehr oder weniger konstant bleibt. Bei der Chroma-Selbstähnlichkeitsmatrix (Abbildung 3c) ist im C-Teil ein Block erkennbar, was an der gleichbleibenden Harmonik in diesem Teil liegt. Auf der anderen Seite zeigt diese Matrix auch zahlreiche Pfade, die die Wiederholungen der drei A-Teile und der vier B-Teile widerspiegeln.

Viele der in der Literatur beschriebenen Strukturanalyseverfahren basieren auf der Extraktion und Interpretation der Block- und Pfadmuster in geeignet definierten Selbstähnlichkeitsmatrizen. Allerdings sind diese Muster oft stark verrauscht und fragmentiert. Weiterhin hängt die Deutlichkeit der Blöcke und Pfade nicht nur vom Typ der Merkmalsdarstellung ab, sondern vom verwendeten Ähnlichkeitsmaß und vor allem auch von der Fenstergröße und zeitlichen Auflösung bei der Merkmalsberechnung. Oft ist bei der Strukturanalyse ein Vergröberungs- oder Glättungsschritt nicht nur hinsichtlich des Rechenaufwands sondern auch aus strukturellen Gründen von Vorteil. Abbildung 3d zeigt, wie man durch geeignete Glättungs- und Schwellwertverfahren die Pfadstruktur einer Selbstähnlichkeitsmatrix erheblich verbessern kann [MK06, Pee04, SGHS08].

5 Audio Thumbnailing

In diesem Abschnitt beschäftigen wir uns mit einem Spezialfall der wiederholungsbasierten Strukturierung. Beim sogenannten *Audio Thumbnailing* geht es darum, ein Segment (auch *Thumbnail* genannt) zu bestimmen, welches eine Audioaufnahme möglichst gut repräsentiert [BW05]. Hierbei macht man häufig die Modellannahme, dass ein solches Segment viele (mehr oder weniger ähnliche) Wiederholungen besitzt, welche wiederum möglichst große Teile der Audioaufnahme abdecken. Im Falle des in Abbildung 4 verwendeten Beatles-Songs “You Can’t Do That” von der musikalischen Form $IV_1V_2B_1V_3V_4B_2V_5O$ (I=Intro, V=Verse, B=Bridge, O=Outro) wäre zum Beispiel der Thumbnail eines der fünf V-Segmente.

Die meisten der in der Literatur beschriebenen Verfahren versuchen Thumbnails aus der Pfadstruktur einer auf Chromamerkmalen basierenden Selbstähnlichkeitsmatrix abzuleiten. Die musikalischen Unterschiede zwischen den (sogenannten) Wiederholungen verursachen allerdings oft starke Störungen in den Pfadstrukturen. In [MJG13] wird ein Verfahren vorgestellt, das auf einem Fitnessmaß basiert, welches jedem Audiosegment ein Fitnesswert zuordnet. Dabei drückt ein Fitnesswert gleichzeitig zwei Eigenschaften des Seg-

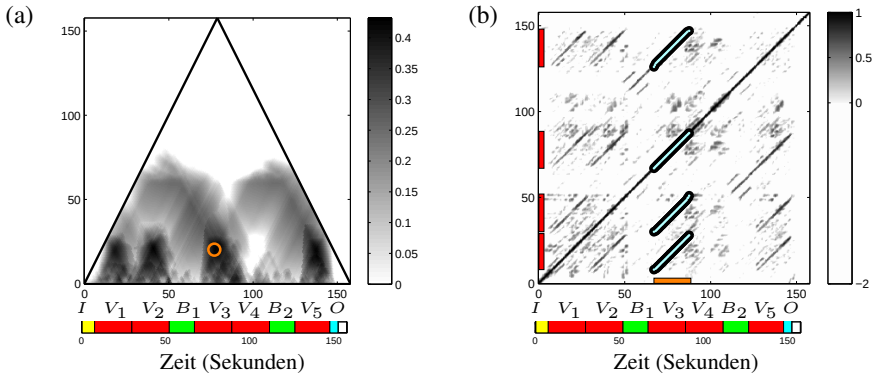

Abbildung 4: (a) Scape-Plot zum Beatles-Song "You Can't Do That." (b) Selbstähnlichkeitsmatrix mit Thumbnail (in orange, horizontale Achse) und allen zum Thumbnail in Beziehung stehenden Segmenten (in rot, vertikale Ache), die in Form einer sogenannten Pfadfamilie (in cyan) erfasst werden.

ments aus. Zum einen wird erfasst, *wie gut* das gegebene Segment andere in Beziehung stehende Segmente erklärt. Zum anderen wird ausgedrückt, *wie viel* von der Audioaufnahme von den in Beziehung stehenden Segmenten abgedeckt wird. Ähnlich wie [Pee07, CF02] ist ein Thumbnail dann als Fitness-maximierendes Segment definiert. Die wesentliche Idee in [MJG13] ist, bei der Berechnung des Fitnessmaßes die Pfadextraktion und Gruppierung der durch die Pfade ausgedrückten Relationen – zwei fehleranfällige Schritte, die in vorherigen Verfahren hintereinander ausgeführt wurden – in einem Optimierungsschritt durchzuführen. Als Ergebnis erhält man ein robustes Verfahren, das auch in Gegenwart von starken musikalischen Variabilitäten gute Thumbnails liefert.

Offensichtlich wird jedes Segment einer Audioaufnahme eindeutig durch Angabe seines Beginns und Endes festgelegt. Eine weitere eindeutige Beschreibung ist die Angabe des Segmentmittelpunkts und der Segmentlänge. Über die zweite Paramaterisierung können alle Segmente durch Punkte in einem Dreieck, einem sogenannten *Scape-Plot*, repräsentiert werden, bei dem die horizontale Achse die Segmentmittelpunkte und die vertikale Achse die Segmentlängen angeben. Solche Scape-Plots wurden in [Sap01] im Musikkontext für die hierarchische Darstellung harmonischer Eigenschaften verwendet, siehe auch Abbildung 7. Im Kontext der Strukturanalyse können Scape-Plots verwendet werden, um den Fitnesswert der jeweiligen Segmente in Form einer geeigneten Farbkodierung (z. B. dunkle Farbe für große und helle Farbe für kleine Fitnesswerte) anzugeben [MJG13]. Abbildung 4a zeigt einen solchen Scape-Plot für unser Beatles-Beispiel, aus dem man die hierarchische Wiederholungsstruktur ablesen kann. Das durch einen Kreis gekennzeichnete Fitness-maximierende Segment definiert den (zu V_3 korrespondierenden) Thumbnail. Abbildung 4b zeigt alle zu diesem Segment in Beziehung stehenden Segmente (d. h. Wiederholungen), die in Form einer sogenannten Pfadfamilie erfasst werden. Hierbei sei bemerkt, dass in diesem Beispiel das V_4-Segment nicht detektiert wurde. Dies liegt daran, dass V_4 zu einer Instrumentalversion mit zusätzlichen akustischen Elementen korrespondiert, die sich von den anderen V-Segmenten erheblich unterscheidet. Für weitere Details zu diesem Verfahren verweisen wir auf [MJG13].

6 Novelty-basierte Segmentierung mittels Strukturmerkmalen

Das Ziel der Novelty-basierten Segmentierung besteht darin, die Grenzen zwischen zwei aufeinanderfolgenden Strukturblöcken zu erkennen. In gängigen Verfahren werden dabei diejenigen Zeitpunkte bestimmt, in denen wesentliche Veränderungen einer geeignete Merkmalsdarstellung des Musiksignals feststellbar sind [Foo00]. Diese Zeitpunkte können z. B. durch Veränderungen in der Instrumentierung (bei Verwendung von MFCCs), in der Harmonik (bei Verwendung von Chromagrammen) oder im Tempo (bei Verwendung von Tempogrammen) hervorgerufen werden. Allerdings ist die Bestimmung solcher lokaler Veränderungen problematisch, da die Merkmalsdarstellungen stark verrauscht sein können und oft auch nur lokale Eigenschaften des Musiksignals erfassen. Die berechneten Segmentgrenzen sind daher oft nicht sehr zuverlässig und fehlerbehaftet.

In [SMGA12] wurde ein Verfahren zur Novelty-basierten Segmentierung vorgestellt, das *lokale* und *globale* Kriterien zur Identifikation von Segmentgrenzen heranzieht und die drei Segmentierungsprinzipien der Wiederholung, Homogenität und Novelty verbindet. Die Hauptidee besteht in der Verwendung sogenannter *Strukturmerkmale*, die globale, strukturelle Eigenschaften des Musiksignals erfassen. Hierzu wird zunächst eine geeignete Selbstähnlichkeitsmatrix (Abbildung 5a) berechnet, die dann in eine Zeit-Lag Matrix (Abbildung 5b) transformiert wird. Während eine der Zeitachsen unverändert bleibt, wird die andere Zeitachse durch Betrachtung von Zeitdifferenzen (auch *Lags* genannt) in eine Lag-Achse transformiert. Wie durch Abbildung 5 illustriert gehen dabei diagonale Strukturen in horizontale Strukturen über. Die Strukturmerkmale sind (unter Anwendung zusätzlicher Glättungsschritte) im Wesentlich definiert als die Spalten der Zeit-Lag Matrix. Aus dieser strukturell-basierten Merkmalsdarstellung wird dann quasi durch Ableitung (Differenzenbildung) eine Novelty-Kurve berechnet, deren lokale Maxima die Segmentgrenzen definieren (Abbildung 5c).

Durch die Verbindung globaler Eigenschaften (erfasst durch Strukturmerkmale) und lokaler Eigenschaften (erfasst durch die Novelty-Berechnung) erhält man vergleichsweise robuste Segmentgrenzen. Dies wurde auch durch eine vergleichende Auswertung im Rahmen der MIREX-Initiative[5] im Jahr 2012 bestätigt, bei dem das in [SMGA12] vorgestellte Verfahren bei den meisten Datensätzen und Auswertungskriterien die besten Resultate erzielte[6].

7 Cross-Version Strategie für Tempo-basierte Segmentierung

In diesem Abschnitt wollen wir zeigen, wie sich das Vorliegen unterschiedlicher Versionen eines Musikstücks ausnutzen lässt, um Segmentierungsresultate zu stabilisieren. Exemplarisch wollen wir hierzu auf die Studie [MPD12] eingehen, bei der eine solche *Cross-*

[5] MIREX steht für Music Information Retrieval Evaluation eXchange, siehe auch `http://www.music-ir.org/mirex/wiki/MIREX_HOME`.

[6] Siehe "Music Structure Segmentation Results" auf `http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results`

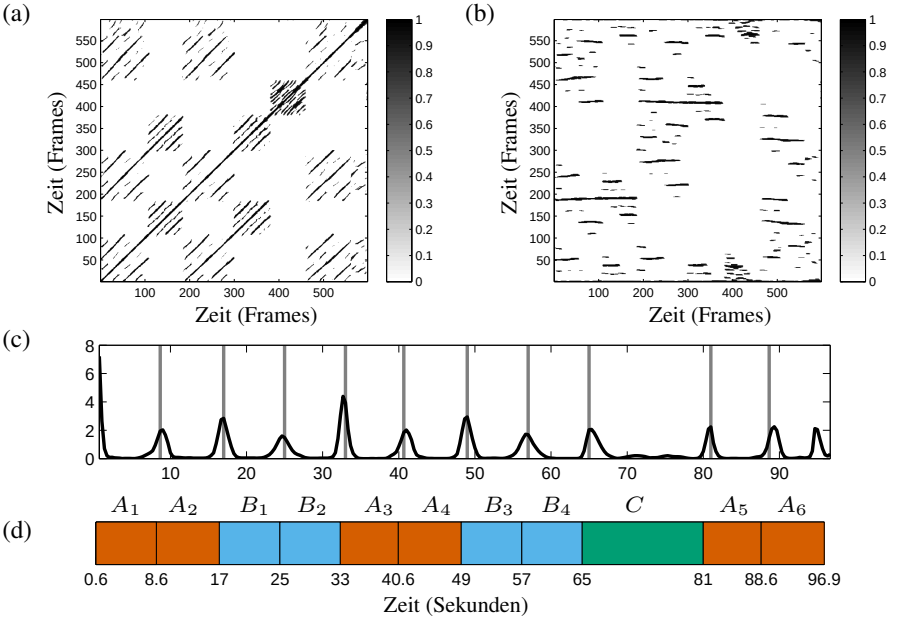

Abbildung 5: Novelty-basierte Segmentierung auf Strukturmerkmalen für eine Aufnahme der Mazurka Op. 24, No. 1 von Frédéric Chopin. **(a)** Selbstähnlichkeitsmatrix basierend auf Chromamerkmalen. **(b)** Zeit-Lag Matrix. **(c)** Novelty-Kurve. **(d)** Manuelle Annotation der musikalischen Form der Aufnahme.

Version Strategie im Kontext einer Tempo-basierten Segmentierung angewendet wurde. Ausgangspunkt der Studie ist ein Datensatz von 49 Mazurken von Frédéric Chopin[7], für die im Schnitt jeweils 57 unterschiedliche Interpretation vorliegen. So sind in diesem Datensatz zum Beispiel 51 Aufnahmen für die Mazurka Op. 68, No. 3 enthalten. Da es sich bei den Mazurken um romantische Klaviermusik handelt, nehmen sich die Pianisten oft erhebliche Freiheiten in der Gestaltung des Tempos. Allerdings gibt es auch interpretationsübergreifende Prinzipien, wann Pianisten das Tempo anziehen und wann sie das Tempo zurücknehmen. Solche Tempoänderungen gehen oft (aber bei weitem nicht immer) mit Strukturgrenzen des zugrundeliegenden Musikstücks einher.

In [MPD12] wurde ein auf Tempogrammen basierendes Novelty-Verfahren verwendet, um für jede Aufnahme eine Novelty-Kurve zu berechnen. Wie schon in dem im Abschnitt 6 beschriebenen Verfahren gesehen entsprechen die lokalen Maxima einer solchen Kurve den wesentlichen Änderungen (in diesem Fall Tempoänderungen). In Abbildung 6a sind die Novelty-Kurven der 51 Aufnahmen in farbkodierter Form zu sehen. Hierbei wurden unter Zuhilfenahme einer mit Taktangaben versehenen MIDI-Referenz die versionsabhängigen physikalischen Zeitachsen (gegeben in Sekunden) der unterschiedlichen Aufnahmen in versionsunabhängige musikalische Zeitachsen (gegeben in Takten)

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

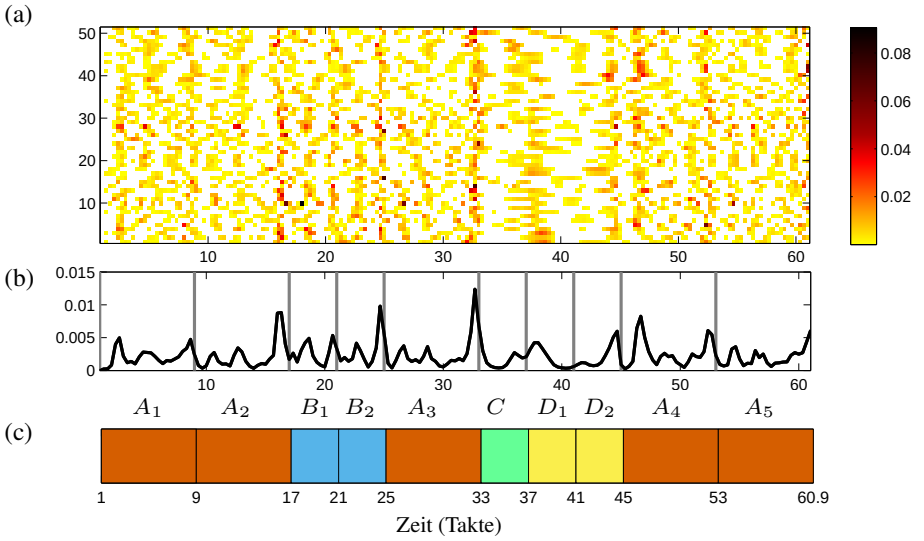

Abbildung 6: Cross-Version Strategie für Tempo-basierte Segmentierung am Beispiel von der Mazurka Op. 68, No. 3 von Chopin (aus [MPD12]). **(a)** Farbkodierte Novelty-Kurven für 51 Aufnahmen (mit transformierten Zeitachsen). **(b)** Versionsübergreifende Novelty-Kurve. **(c)** Manuelle Annotation der musikalische Form.

transformiert[8]. Auf Basis der gemeinsamen Zeitachsen kann nun über die unterschiedlichen Novelty-Kurven eine Mittelung vorgenommen werden. Man erhält als Resultat eine versionsübergreifende Novelty-Kurve (Abbildung 6b). Wie die Abbildung zeigt, werden hierdurch versionsabhängige und rauschartige Phänomene abgeschwächt und die lokalen Maxima der resultierenden Kurve korrelieren oft zu musikalisch sinnvollen Segmentgrenzen.

8 Fazit

In diesem Artikel haben wir das Gebiet der automatisierten Strukturanalyse von Musiksignalen aus Sicht der Informatik beleuchtet. Die Komplexität und Reichhaltigkeit der Aufgabenstellung begründet sich aus den unterschiedlichen Segmentierungsprinzipien, den zahlreichen musikalischen Aspekten, dem zeitlich-hierarchischen Kontext, und den unterschiedlichen Darstellungsformen. Nur in wenigen bisher in der Literatur beschriebenen Verfahren werden gleichzeitig mehrere Strukturierungsprinzipien betrachtet. Zum Beispiel wird in [PK09] eine Optimierungsstrategie beschrieben, die gleichzeitig Block- und Pfadstrukturen berücksichtigt. Ein weiteres Beispiel ist das in diesem Artikel betrachtete Verfahren, bei dem auf Selbstähnlichkeitsmatrizen basierende Merkmale (globale Eigenschaften) mit Novelty-basierten Methoden (lokale Eigenschaften) kombiniert wurden. Weiterhin haben wir gesehen, wie der hierarchische Kontext von Wiederholungsstrukturen

[8] Dies kann man mit wie in [EMG09] beschriebenen Synchronisationstechniken bewerkstelligen.

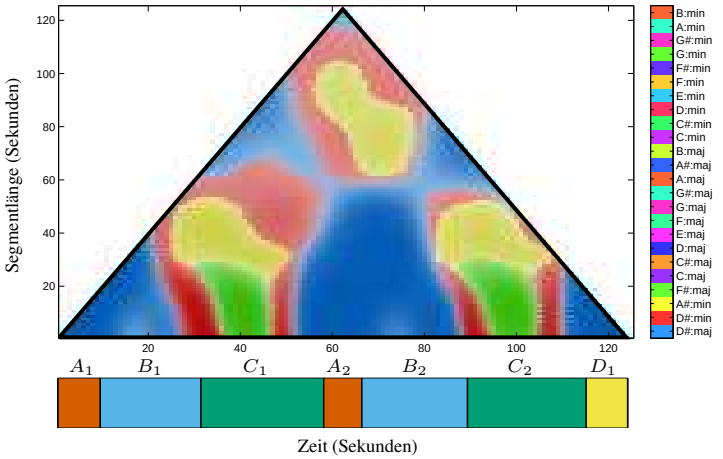

Abbildung 7: Scape-Plot-Darstellung der lokalen Harmonik für das Lied “Die Post” aus der Winterreise D911 von Franz Schubert, siehe [Sap01] für Details. Im unteren Teil ist die manuell erstellte Annotation der musikalischen Form zu sehen.

durch einen Scape-Plot dargestellt werden kann. Ähnlich zeigt Abbildung 7 einen solchen Scape-Plot zur Darstellung der lokalen Harmonik einer Audioaufnahme [Sap01]. Schließlich haben wir dargestellt, wie mehrere Versionen eines Musikstücks zur Stabilisierung der Analyseergebnisse verwendet werden können. Eine ähnliche Strategie wurde in [KMK13] angewendet, um harmonische Analysen zu stabilisieren.

Um der Vielfältigkeit musikalischer Strukturen besser gerecht zu werden, müssen verstärkt Verfahren entwickelt werden, die eine mehrschichtige Analyse und Strukturierung unter simultaner Berücksichtigung unterschiedlicher musikalischer Aspekte und Segmentierungsparadigmen erlauben. Weiterhin ist die Strukturanalyse oft auch schon aus rein musikalischer Sicht ein schlecht gestelltes Problem [SBF+11]. Daher sind gängige Evaluationsmaße durch Hinzuziehung musikalischer Experten kritisch zu hinterfragen. Auch sollte verstärkt über konkrete Anwendungsszenarien nachgedacht werden, welche an die zu extrahierenden Strukturen konkrete Anforderungen stellen. Solche Anwendungen ergeben sich zum Beispiel im Kontext inhaltsbasierter Suche und Navigation in Musikdatenbeständen [DFT+12, Got06].

Danksagung: Dieser Artikel sowie viele der beschriebenen Verfahren sind im Umfeld des von der DFG geförderten METRUM-Projekts (DFG CL 64/8-1, DFG MU 2682/5-1) entstanden.

Literatur

[BDA+05] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies und Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

[BMK06] Michael J. Bruderer, Martin McKinney und Armin Kohlrausch. Structural boundary perception in popular music. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Seiten 198–201, Victoria, Canada, 2006.

[BW05] Mark A. Bartsch und Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

[CF02] Matthew Cooper und Jonathan Foote. Automatic Music summarization via similarity analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Seiten 81–85, Paris, France, 2002.

[Cha06] Wei Chai. Semantic segmentation and summarization of music: methods based on tonality and recurrent structure. *IEEE Signal Processing Magazine*, 23(2):124–132, 2006.

[CVG+08] Michael A. Casey, Remco Veltkap, Masataka Goto, Marc Leman, Christophe Rhodes und Malcolm Slaney. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[DFT+12] David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth und Meinard Müller. A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction. *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, 12(2-3):53–71, 2012.

[DG08] Roger B. Dannenberg und Masataka Goto. Music structure analysis from acoustic signals. In David Havelock, Sonoko Kuwano und Michael Vorländer, Hrsg., *Handbook of Signal Processing in Acoustics*, Jgg. 1, Seiten 305–331. Springer, New York, NY, USA, 2008.

[DM90] Steven B. Davis und Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in Speech Recognition*, Seiten 65–74, 1990.

[Dow03] J. Stephen Downie. Music information retrieval. *Annual Review of Information Science and Technology (Chapter 7)*, 37:295–340, 2003.

[EMG09] Sebastian Ewert, Meinard Müller und Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 1869–1872, Taipei, Taiwan, 2009.

[FC03] Jonathan T. Foote und Matthew L. Cooper. Media segmentation using self-similarity decomposition. *Storage and Retrieval for Media Databases*, 5021(1):167–175, 2003.

[Foo00] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Seiten 452–455, New York, NY, USA, 2000.

[G06] Emilia Gómez. *Tonal description of music audio signals*. Dissertation, UPF Barcelona, 2006.

[GM11] Peter Grosche und Meinard Müller. Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1688–1701, 2011.

[GMK10] Peter Grosche, Meinard Müller und Frank Kurth. Cyclic tempogram – a mid-level tempo representation for music signals. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 5522 – 5525, Dallas, Texas, USA, 2010.

[Got06] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1783–1794, 2006.

[KD06] Anssi P. Klapuri und Manuel Davy, Hrsg. *Signal processing methods for music transcription*. Springer, New York, 2006.

[KMK13] Verena Konz, Meinard Müller und Rainer Kleinertz. A cross-version chord labelling approach for exploring harmonic structures – a case study on Beethoven's Appassionata. *Journal of New Music Research*, Seiten 1–17, 2013.

[Lei87] Hugo Leichtentritt. *Musikalische Formenlehre*. Breitkopf und Härtel, 12. Auflage, Wiesbaden, Germany, 1987.

[LS08] Mark Levy und Mark Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):318–326, 2008.

[MEKR11] Meinard Müller, Daniel P. W. Ellis, Anssi Klapuri und Gaël Richard. Signal processing for music analysis. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

[Mic] Michael Michaelis. Musik & Form. `http://www.michael-michaelis.de/htdocs/musikalischeformen`, Retrieved 18.06.2009.

[ME11] Meinard Müller und Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Seiten 215–220, Miami, FL, USA, 2011.

[MJG13] Meinard Müller, Nanzhu Jiang und Peter Grosche. A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *IEEE Transactions on Audio, Speech & Language Processing*, 21(3):531–543, 2013.

[MK06] Meinard Müller und Frank Kurth. Enhancing similarity matrices for music audio analysis. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seiten 437–440, Toulouse, France, 2006.

[MPD12] Meinard Müller, Thomas Prätzlich und Jonathan Driedger. A cross-version approach for stabilizing tempo-based novelty detection. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, Seiten 427–432, Porto, Portugal, 2012.

[Mül07] Meinard Müller. *Information retrieval for music and motion*. Springer Verlag, 2007.

[MXKS04] Namunu C. Maddage, Changsheng Xu, Mohan S. Kankanhalli und Xi Shao. Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the ACM International Conference on Multimedia*, Seiten 112–119, New York, NY, USA, 2004.

[NG98] Klaus Wolfgang Niemöller und Bram Gätjen, Hrsg. *Perspektiven und Methoden einer Systemischen Musikwissenschaft*. Peter Lang, 1998.

[Ong07] Bee Suan Ong. *Structural analysis and segmentation of music signals*. Dissertation, University Pompeu Fabra, Barcelona, Spain, 2007.

[Ori06] Nicloa Orio. Music retrieval: a tutorial and review. *Foundation and Trends in Information Retrieval*, 1(1):1–90, 2006.

[Pee04] Geoffroy Peeters. Deriving musical structure from signal analysis for music audio summary generation: "sequence" and "state" approach. In *Computer Music Modeling and Retrieval*, Jgg. 2771 of *Lecture Notes in Computer Science*, Seiten 143–166. Springer Berlin / Heidelberg, 2004.

[Pee05] Geoffroy Peeters. Time variable tempo detection and beat marking. In *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, 2005.

[Pee07] Geoffrey Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Seiten 35–40, Vienna, Austria, 2007.

[PK09] Jouni Paulus und Anssi P. Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170, 2009.

[PMK10] Jouni Paulus, Meinard Müller und Anssi P. Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, Seiten 625–636, Utrecht, The Netherlands, 2010.

[Sap01] Craig Stuart Sapp. Harmonic visualizations of tonal music. In *Proceedings of the International Computer Music Conference (ICMC)*, Seiten 423–430, La Habana, Cuba, 2001.

[SBF+11] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure und J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Seiten 555–560, Miami, FL, USA, 2011.

[SGHS08] Joan Serrà, Emilia Gómez, Perfecto Herrera und Xavier Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, 2008.

[SMGA12] Joan Serrà, Meinard Müller, Peter Grosche und Josep Lluis Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, Toronto, Ontario, Canada, 2012.

[TSB05] H. Terasawa, M. Slaney und J. Berger. The thirteen colors of timbre. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Seiten 323–326, 2005.

[TWV05] Rainer Typke, Frans Wiering und Remco C. Veltkamp. A survey of music information retrieval systems. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Seiten 153–160, London, GB, 2005.