Aktuelle Aspekte des Music Information Retrieval

Meinard Müller, Frank Kurth, Michael Clausen
Institut für Informatik, Universität Bonn, Römerstr. 164, D-53117 Bonn, Email: {meinard,frank,clausen}@iai.uni-bonn.de

Music Information Retrieval (MIR) ist eine junge Disziplin, die es sich zur Aufgabe gemacht hat, Methoden zu erforschen und Systeme zu entwickeln, die Benutzern große, in digitaler Form vorliegende Musikkollektionen in vielfältiger Weise zugänglich machen. Während noch vor ca. zehn Jahren relativ einfache Retrievalszenarien wie die Suche nach kurzen Melodieverläufen in monophonen Musikstücken studiert wurden, liegt der Fokus aktueller Forschungsbestrebungen auf semantisch komplexen Problemstellungen wie dem Auffinden aller musikalisch ähnlichen Passagen zu einem gegebenen akustischem Musikausschnitt. Im Hinblick auf effiziente Navigation und inhaltsbasierte Suche in umfangreichen und heterogenen Musikdatenbeständen ist die automatisierte Analyse und inhaltliche Erschließung der Musikdokumente in Verbindung mit leistungsfähigen Indexierungs- und Retrievaltechniken von großer Bedeutung. In diesem Artikel werden aktuelle Fragestellungen der automatisierten Musikdatenerschließung vorgestellt und einige Grundideen bisheriger Lösungansätze skizziert.

1 Einleitung

Moderne digitale Musikbibliotheken enthalten multimediale Dokumente in zahlreichen Ausprägungen und Formaten, die ein Musikwerk auf verschiedenen Ebenen semantischer Ausdruckskraft beschreiben. Man denke hier beispielsweise an CD-Aufnahmen diverser Interpreten, Noten, MIDI-Daten oder Gesangstexte. Allgemein gesprochen ist das Hauptziel des *Music Information Retrieval* (MIR) die Nutzbarmachung solch inhomogener und komplexer Musikdatenbstände. Eine zentrale Aufgabe ist hierbei die Entwicklung effizienter Such- und Navigationssysteme, die es dem Benutzer erlauben, den Datenbestand bezüglich unterschiedlichster musikrelevanter Aspekte zu durchsuchen. Während die textbasierte Suche nach Musik anhand von Komponistennamen, Songtitel, Werkverzeichnisnummer oder dergleichen mit klassischen Datenbanktechniken möglich ist, stellt die inhaltsbasierte Suche in Musikdaten ohne das Zurückgreifen auf manuell erzeugte Annotationen ein schwieriges Problem dar. Was ist zu tun, wenn man nur ein Melodiefragment vorpfeifen kann oder nur einen kurzen akustischen Ausschnitt von einem Musikstück vorliegen hat? Wie geht man vor, wenn der Benutzer an allen CD-Aufnahmen (samt der genauen Zeitpositionen innerhalb der jeweiligen Aufnahmen) interessiert ist, die gewisse Notenkonstellationen, Harmonieverläufe, oder Rhythmen aufweisen? Wie können Partiturdaten oder Musikaufnahmen hinsichtlich wiederkehrender Muster durchsucht werden? Dies ist nur eine kleine Auswahl aktueller MIR-Fragestellungen, die eng mit der automatisierten Analyse von Musikdaten verknüpft sind.

Bei der Entwicklung inhaltsbasierter Such- und Navigationsmechanismen führt die oben angesprochene Inhomogenität und Komplexität existierender Musikdokumentensammlungen zu großen, weitgehend noch ungelösten Problemen. Eine entscheidende Rolle kommt hier der umfassenden Annotation und Verlinkung des Datenbestandes zu, was allerdings aufgrund der enormen Datenmassen manuell nicht bewerkstelligt werden kann. Genau diesem Punkt widmet sich die *automatisierte Musikdatenerschließung*, bei der es allgemein gesprochen um die automatische Generierung semantisch hochwertiger Annotationen geht, mittels derer dann inhaltsbasierte Anfragen an Musikdatenbanken effizient bearbeitet werden können.

In diesem Artikel werden drei aktuelle Themenkomplexe der automatisierten Musikdatenerschließung vorgestellt, die insbesondere im Hinblick auf effizientes und effektives Musikretrieval von fundamentaler Bedeutung sind. Es wird also nicht der Versuch eines umfassenden Überblicks über das komplexe und weitläufige Gebiet des MIR unternommen. Stattdessen werden exemplarisch konkrete Lösungsansätze skizziert und in einen allgemeineren methodischen Rahmen eingeordnet. Literaturangaben und eine Diskussion offener Probleme findet man in den jeweiligen Abschnitten. Konkret werden die folgenden Fragestellungen diskutiert. Bei der *Musiksynchronisation* (Abschnitt 2) geht es um die automatische Verlinkung von Musikdatenströmen unterschiedlicher Formate, die dasselbe Musikstück repräsentieren. In Abschnitt 3 werden zwei grundlegende Problemstellungen des *Audioretrieval* behandelt. Ziel der kurz skizzierten *Audioidentifikation* ist dabei die Erkennung einer in einer Datenbank enthaltenen Aufnahme anhand eines kurzen Audiofragments. Die Fragestellung des *Audiomatching* kann als Verallgemeinerung der Audioidentifikation angesehen werden. Hierbei besteht die Anfrage aus einem 10–30 sekündigen Audioausschnitt. Ziel ist dann die automatische Identifikation und Extraktion aller zu dieser Anfrage musikalisch ähnlichen Abschnitte, z. B. unabhänigig vom Interpreten oder von der Instrumentation, in der gegebenen Datenbank. Eine verwandte Fragestellung stellt die *Strukturanalyse* (Abschnitt 4) dar, bei der automatisch sich wiederholende Strukturen innerhalb eines Musikstücks (unter Zulassung gewisser musikalischer Variationen) erkannt werden sollen. Schließlich beschreiben wir in Abschnitt 5, wie sich diese unterschiedlichen Technologien verbinden und in ein Gesamtsystem zur effektiven und effizienten Musiksuche integrieren lassen. Hierzu wurde an der Universität Bonn ein erster Prototyp (SyncPlayer) entwickelt.

2 Musiksynchronisation

Ein großes Problem innerhalb multimedialer Datenbanken sind die verschiedenen Datenformate, die allein schon bei reinen Musikdaten anzutreffen sind. Gängige Datenformate sind wellenformbasierte Audiodatenformate wie WAV oder MP3, Partiturdatenformate wie Score oder Capella sowie partiturnahe Datenformate wie MIDI. (MIDI entspricht i.w. der Klavierwalzendarstellung.) So wird z. B. die Interpretation eines Musikstücks vollständig durch die Darstellung der *akustischen Wellenform* beschrieben und kann über geeignete Ausgabe-

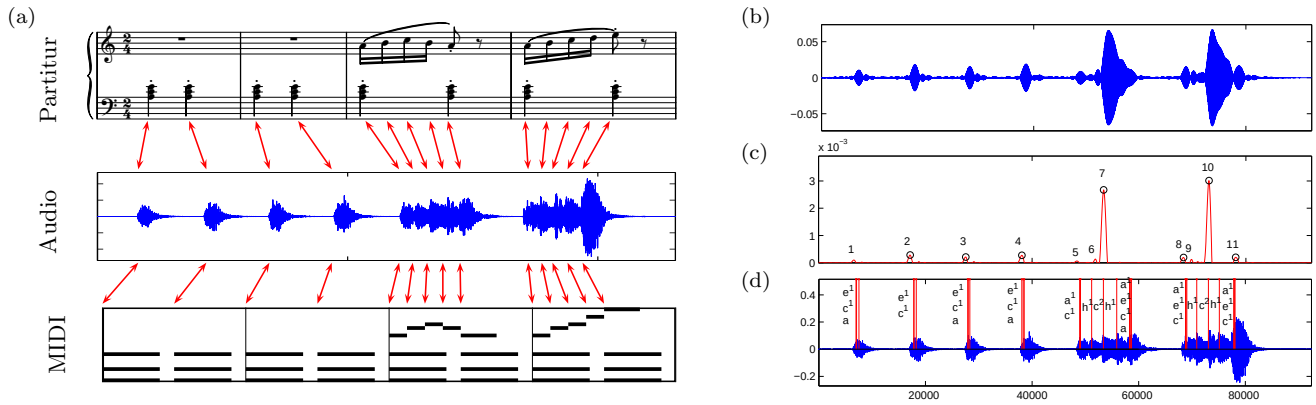

Abbildung 1: (a) Verlinkung von Musikdaten in unterschiedlichen Formaten (Partitur, Audio, MIDI), die dasselbe Musikstück (die ersten vier Takte der Etüde Nr. 2, op. 100, F. Burgmüller) repräsentieren. (b) Subbandsignal zur Tonhöhe C5 (MIDI pitch 72) für das Audiosignal in (a). (c) Resultierende Onset-Kurve und (durch Nummern gekennzeichnete) Kandidaten für Einsatzzeiten zur Tonhöhe C5. (d) Die extrahierten notenähnlichen Parameter aus dem Audiodatenstrom sind durch vertikale Linien gekennzeichnet (samt Bezeichnung für die jeweilige Tonhöhe).

geräte, z. B. mit einem CD-Gerät, wiedergegeben werden. Die Wellenform hat allerdings den Nachteil, dass *inhaltsbezogene Informationen* – wie z. B. die dem Musikstück zugrundeliegenden Noten – praktisch nicht ablesbar sind. Im Gegensatz hierzu wird ein Musikstück durch die *Partiturdarstellung* rein durch inhaltsbezogene Informationen wie Tonhöhen, Tondauern oder Einsatzzeiten beschrieben, welche aber großen interpretatorischen Spielraum hinsichtlich des Tempos, der Dynamik oder der Ausführung von Notengruppen zulassen. Im Hinblick auf eine effiziente Navigation in inhomogenen Musikdatenbeständen ist die Entwicklung von *Synchronisationstechniken*, die zur automatischen Verlinkung zweier Datenströme unterschiedlicher Formate eingesetzt werden können, von großer Bedeutung, siehe u. a. [ACKM03, DW05, HDT03, MKR04, Ra04, SRS03, TE03]. Hierbei verstehen wir unter *Synchronisation* (man sprich oft auch von *Alignment*) ein Verfahren, das zu einer bestimmten Position innerhalb einer Darstellung eines Musikstücks (z. B. in einer CD-Aufnahme) die entsprechende Stelle innerhalb einer anderen Darstellung (z. B. in einer Partitur) bestimmen kann, siehe Abb. 1 (a).

2.1 Audio-Partitur-Synchronisation

Abhängig von den zugrundliegenden Formaten lassen sich eine ganze Reihe unterschiedlicher Synchronisationsaufgaben formulieren [ACKM03]. Exemplarisch soll in diesem Abschnitt auf das Szenario eingegangen werden, in dem ein Musikstück sowohl als CD-Aufnahme (Audio) als auch in einem symbolischen Notenformat (Partitur) vorliegt. Unter einer *Audio-Partitur-Synchronisation* verstehen wir dann ein Verfahren, das zu einer bestimmten *physikalischen Einsatzzeit* im Audiodatenstrom die entsprechende *musikalische Einsatzzeit* in der Partitur bestimmen kann. In diesem Sinne kann eine Audio-Partitur-Synchronisation als automatisierte Annotation des Audiodatenstroms durch die Noten der Partitur oder auch als Extraktion bzw. Lokalisation von Noteninformation im Audiodatenstrom unter Ausnutzung des Vorwissens der Partiturdaten angesehen werden.

Da sich die wellenformbasierten Audiodaten grundsätzlich von den rein symbolischen Partiturdaten unterscheiden, gehen die meisten der gängigen Ansätze zur Audio-Partitur-Synchronisation in zwei Schritten vor. In einem ersten Schritt werden aus dem Audiodatenstrom geeignete Parameter extrahiert, die einen Vergleich mit den Partiturdaten erlauben. Im zweiten Schritt wird dann eine optimale Zuordnung mittels dynamischer Programmierung (DP) unter Verwendung geeigneter lokaler Ähnlichkeitsmaße berechnet. Für Details und weitere Literaturhinweise verweisen wir auf [ACKM03, MKR04, SRS03]. Der Arbeit [MKR04] folgend gehen wir nun auf einige Grundideen genauer ein.

2.2 Merkmalsextraktion

DP-basierte Algorithmen, wie sie im Verlinkungsschritt eingesetzt werden, weisen ein quadratisches Laufzeitverhalten in der Eingabegröße auf und stellen daher meist den Flaschenhals bei der Audio-Partitur-Synchronisation dar. Daher wird in dem folgenden Verfahren eine kleine Anzahl von semantisch ausdrucksstarken Merkmalen verwendet, die sowohl effizient aus dem Audiosignal extrahiert werden können als auch eine hohe Zeitauflösung aufweisen, wie sie im Hinblick auf eine präzise Synchronisation wichtig ist. Unter Verwendung fortgeschrittener Filtertechniken (Filterbank bestehend aus 88 elliptischen IIR-Filtern) wird das Audiosignal in 88 Bänder (gemäß den Klaviertönen) zerlegt. Zum Beispiel zeigt Abb. 1 (b) für ein Audiosignal das resultierende Subbandsignal zur Tonhöhe C5 (MIDI pitch 72). Mittels energiebasierter Verfahren und diskreter Ableitung wird für jedes Subband eine Onset-Kurve berechnet, aus der sich durch geeignetes Peak-picking Kandidaten für Einsatzzeiten der jeweiligen Tonhöhe ergeben, siehe Abb. 1 (c) und Abb. 1 (d).

Im Fall polyphoner Musik stellt die Extraktion von Notenparametern ein extrem schwieriges Problem dar. Für Klaviermusik bereiten z. B. Obertöne, Resonanz- und Schwebungseffekte, Vermischung von Klangspektren (verursacht durch das Haltepedal) oder auch das Vorliegen starker inharmonischer Komponenten (verursacht durch den Tastenanschlag) große Schwierigkeiten. Auch wenn die extrahierten Merkmale im Hinblick auf eine *Musiktranskription*, also der Übertragung einer Musikaufnahme in Notenschrift, unzureichend sein mögen, ermöglichen sie dennoch im Allgemeinen eine ausgezeichnete *Musiksynchronisation*.

2.3 Synchronisationsschritt

Nach einer geeigneten Aufarbeitung und Kodierung der Partiturdaten wird nun im zweiten Schritt mittels DP eine kostenoptimale zeitliche Verlinkung zwischen den Partitur- und Extraktionsparametern berechnet. Hierbei wird ein Verlinkungsmodell verwendet, welches sich an klassische auf „Dynamic Time Warping“ (DTW) basierende Methoden anlehnt, siehe z. B. [RJ93, SRS03]. Um eventuellen Unstimmigkeiten zwischen dem Partitur- und Audiodatenstrom, bedingt z. B. durch interpretatorische Abweichungen oder fehlerhafte Extraktion, Rechnung zu tragen, wird nicht die Zuordnung aller Partitur- bzw. Extraktionsparameter erzwungen, sondern es werden auf beiden Seiten auch unverlinkte Ereignisse erlaubt – ganz nach dem Motto: „Besser keine Zuordnung als eine schlechte Zuordnung.“ Darüber hinaus basiert die Definition des lokalen Ähnlichkeitsmaßes auf folgendem einfachen aber weitreichenden Prinzip: Die Partitur gibt vor, wonach im Audiodatenstrom zu suchen ist. Bei der Verlinkung werden also nur Extraktionsparameter berücksichtigt, die sich in der Partitur widerspiegeln.

2.4 Fazit und Ausblick

Das oben beschriebene Verfahren wurde in MATLAB implementiert und anhand zahlreicher Beispiele polyphoner Klaviermusik unterschiedlicher Komplexität getestet, einschließlich Etüden von Chopin und Klaviersonaten von Beethoven. Hierbei hat sich gezeigt, dass unser Verfahren für die eingeschränkte Musikklasse polyphoner Klaviermusik gute Synchronisationsergebnisse hoher Auflösung erzielt, die für Anwendungen wie die inhaltsbasierte Musiksuche oder zum Zwecke der zeitgleichen Notendarstellung beim Abspielen einer CD-Aufnahme mehr als ausreichend sind. Selbst plötzliche Tempoänderungen, Ritardandi, Accelerandi oder Fermaten konnten im Allgemeinen gut erfasst werden. (Eine geeignete Sonifikation von Synchronisationsergebnissen findet man unter `http://www-mmdb.iai.uni-bonn.de/projects/sync/`.)

Ähnliche DTW-basierte Algorithmen werden zur Audio-Audio-Synchronisation eingesetzt, bei der zwei unterschiedliche Interpretationen desselben Musikstücks zu verlinken sind. Hierbei liefern insbesondere chromabasierte Ansätze, wie sie in Abschnitt 3.1 diskutiert werden, vielversprechende Ergebnisse für Musik mit prägnanten harmonischen Komponenten [HDT03, MMK06c]. Für die Zukunft sind Synchronisationsalgorithmen zu entwickeln, die andere Musikaspekte wie Rhythmus oder Klangfarbe berücksichtigen. Noch weitgehend ungelöst ist die automatisierte Text-Audio-Synchronisation von vorliegenden Liedtexten mit entsprechenden CD-Aufnahmen.

3 Audioretrieval

Im Kontext der inhaltsbasierten Musiksuche hat das „Query-by-Example“ Paradigma viel Aufmerksamkeit auf sich gezogen. Hierbei ist eine Anfrage an eine Musikdatenbank in Form eines Musikausschnitts gegeben. Das Ziel besteht dann darin, alle in der Datenbank enthaltenen Abschnitte zu bestimmen, die der Anfrage in gewisser, semantisch sinnvoller Weise ähneln. Das Problem der ähnlichkeitsbasierten Suche stellt insbesondere für als Wellenform gegebene digitale Audiodaten ein schwieriges und noch in vielen Teilen ungelöstes Forschungsproblem dar. Unterschiedliche Ähnlichkeitsbegriffe führen hier zu unterschiedlich schwierigen Retrievalaufgaben.

Aufgabe der *Audioidentifikation* ist es, einen gegebenen Ausschnitt einer Audioaufnahme als Bestandteil genau dieser Aufnahme zu erkennen. Mit Hilfe der Audioidentifikation kann man also exakt ermitteln, in welchem Stück einer Audio-CD ein bestimmter Audioausschnitt enthalten ist. Zusätzlich fordert man oft, auch die exakte Position des gesuchten Ausschnitts in der Originalaufnahme lokalisieren zu können. Eine weitere wichtige Anforderung zahlreicher Anwendungsszenarien ist, dass auch stark qualitätsreduzierte (etwa verrauschte oder mit einem Mikrophon aufgenommene) Versionen des Originals identifiziert werden können. Der bei dieser Retrievalaufgabe zugrunde liegende Ähnlichkeitsbegriff ist also im wesentlichen die *Identität* von Anfrage und Treffer. Im Hinblick auf die enormen im Audiobereich anfallenden Datenmassen (aufgrund der Speicherintensität des Datenformats und aufgrund umfangreicher Kollektionen an CD-Aufnahmen) sind geeignete Mechanismen zur kompakten merkmalsbasierten Beschreibung von Audioaufnahmen, sowie die Verwendung leistungsfähiger Indexierungs- und Retrievaltechniken erforderlich. Hier wurden in den letzten sechs Jahren mit unterschiedlichen Ansätzen erfolgreiche Lösungsansätze vorgeschlagen [AHFC01, CBMN02, Wa03], so dass die Aufgabe der Audioidentifikation als im wesentlichen gelöst angesehen werden kann.

Die im folgenden ausführlich diskutierte Problemstellung des *Audiomatching* kann als Verallgemeinerung der Audioidentifikation aufgefasst werden. Ausgangspunkt ist hier eine große Musikdatenbank, die typischerweise mehrere verschiedene Aufnahmen desselben Musikstücks enthält, wobei diese Aufnahmen im Allgemeinen von unterschiedlichen Interpreten und in eventuell verschiedenen Instrumentierungen eingespielt wurden. Bei Anfrage eines kurzen Audioausschnitts sollen dann automatisch alle entsprechenden Abschnitte in allen in der Datenbank enthaltenen Interpretationen des zugrundeliegenden Musikstücks gefunden werden. Das Problem des Audiomatching ist noch weitgehend unerforscht. Ein erster harmoniebasierter Ansatz wird in [MKC05] vorgeschlagen, den wir im folgenden skizzieren. Grundlage sind eine Klasse statistischer Chromamerkmale, wobei sich die Chroma auf die zwölf Tonhöhenklassen der wohltemperierten Stimmung beziehen. Folgen solcher Merkmale korrelieren stark mit dem Harmonieverlauf des zugrundeliegenden Musikstücks und sind hochgradig invariant bezüglich Änderungen von Parametern wie Dynamik, Klangfarbe, Artikulation sowie gegenüber lokalen Tempodeformationen. Weiterhin beschreiben wir ein robustes Matchingverfahren, durch welches man lokale und globale Tempovariationen in den Griff bekommt.

3.1 CENS-Merkmalsfolgen

Im ersten Schritt der Merkmalsextraktion wird, wie in Abschnitt 2.2 beschrieben, das zu transformierende Audiosignal in 88 Tonhöhenbänder zerlegt. Für jedes Band wird durch Faltung mit einem 200-Millisekunden Rechteckfenster eine lokale Energiekurve berechnet, deren Datenrate auf 10 Hz reduziert wird. Anschließend werden alle zu gleichen Tonhöhenklassen korrespondierenden Energiewerte zu einem Chroma-Energiewert aufsummiert. (Z. B. werden die Energiewerte der Bänder zu den Tonhöhen A0,A1,. . .,A7 zu einem

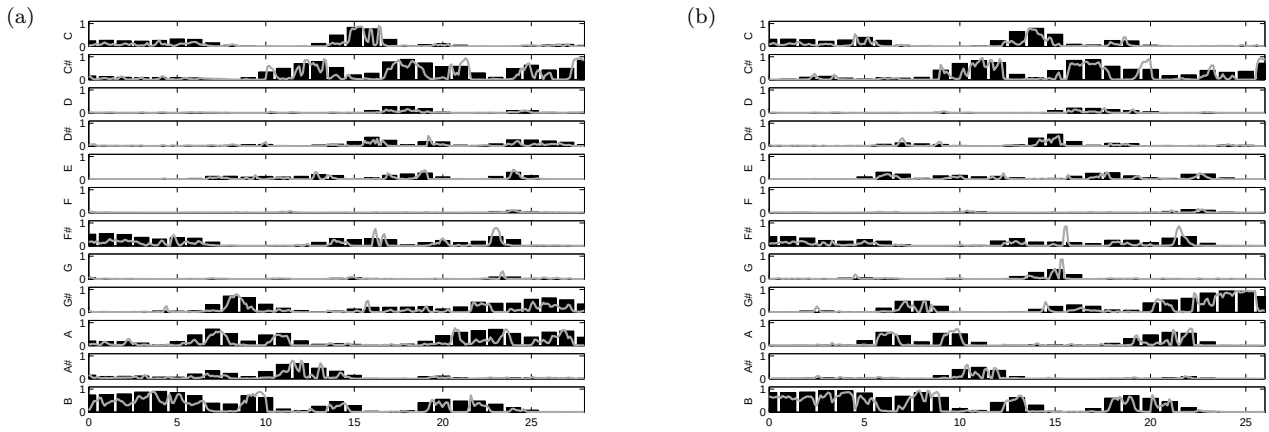

Abbildung 2: (a) Takte 44–55 (Sekunden 112–140) einer Zukerman-Interpretation von Vivaldis Frühling RV 269, Nr. 1. Die hellen Kurven stellen die lokalen Chroma-Energieverteilungen (10 Hz) dar, während die dunklen Balken die CENS-Merkmalsfolge (1 Hz) zeigen. (b) Entsprechende Takte (Sekunden 102–128) einer Perlman-Interpretation. Obwohl die beiden Interpretationen erhebliche Unterschiede aufweisen, sind die Verläufe der zugehörigen CENS-Merkmalsfolgen (bis auf eine globale zeitliche Skalierung) sehr ähnlich.

Energiewert zum Chroma A zusammengefasst.) Nach einem anschliessenden Normalisierungsschritt erhält man schließlich eine Folge von 12-dimensionalen Chromavektoren (10 Vektoren pro Sekunde), wobei jeder Vektor die lokale Energieverteilung der im Audiosignal vorkommenden Frequenzen auf die 12 Chromabänder widerspiegelt, siehe Abb. 2.

Die so erhaltenen Chromamerkmale sind durch die Identifikation modulo Oktaven gleicher Tonhöhenbänder robust unter Klangfarbenänderung und durch die Normalisierung zusätzlich invariant unter Dynamikveränderungen. Zur Erhöhung der Robustheit gegenüber lokalen zeitlichen Verzerrungen werden die Merkmale noch weiter vergröbert und lokale Statistiken über geeignet quantisierte Chroma-Energieverteilungen innerhalb eines 4100-Millisekunden Analysefensters berechnet. Anschliessend werden die 12-dimensionalen Statistikvektoren erneut normalisiert und die Datenrate der Vektorfolge wird durch Downsampling um den Faktor 10 auf 1 Hz reduziert, siehe Abb. 2. Die so resultierenden Merkmale werden abkürzend mit *CENS* (**C**hroma **E**nergy distribution **N**ormalized **S**tatistics) bezeichnet, vgl. [MKC05] für Details.

3.2 Matchingverfahren

Auf der Grundlage der CENS-Merkmale gehen wir nun auf die wichtigsten Ideen eines robusten Verfahrens zum Audiomatching ein. In einem Vorverarbeitungsschritt werden zunächst die CENS-Merkmalsfolgen aller Aufnahmen der Datenbank berechnet. Durch Konkatenation dieser Folgen (und unter Protokollierung der Dateigrenzen) kann die gesamte Datenbank durch eine einzige Folge $\mathcal{D} := (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^N)$ von CENS-Merkmalen repräsentiert werden.

Im folgenden Szenario besteht eine typische Anfrage aus einem Musikausschnitt von 10–30 Sekunden Dauer. Diese Anfrage wird zunächst in eine CENS-Merkmalsfolge $\mathcal{Q} := (\vec{w}^1, \vec{w}^2, \ldots, \vec{w}^M)$ transformiert und dann mit jeder Teilfolge $(\vec{v}^i, \vec{v}^{i+1} \ldots, \vec{v}^{i+M-1})$, $1 \leq i \leq N - M + 1$, bestehend aus M aufeinanderfolgenden Vektoren von $\mathcal{D}$ verglichen. Hierzu wird in Abhängigkeit von $\mathcal{D}$ und $\mathcal{Q}$ eine Abstandsfunktion $\Delta : [1 : N - M + 1] \rightarrow [0, 1]$ mit $\Delta(i) := 1 - \frac{1}{M} \sum_{m=1}^{M} \langle \vec{v}^{i+m-1}, \vec{w}^m \rangle$ definiert, siehe Abb. 3 (a). Aufgrund der Normierung der CENS-Vektoren entspricht dabei das Skalarprodukt $\langle \vec{v}^{i+m-1}, \vec{w}^m \rangle$ gerade dem Cosinus des Winkels zwischen den beiden Vektoren $\vec{v}^{i+m-1}$ und $\vec{w}^m$. $\Delta(i)$ beschreibt den Abstand zwischen $\mathcal{Q}$ und der ab Position i beginnenden Teilfolge von $\mathcal{D}$ der Länge M. Unterschiedliche Interpretationen weisen häufig verschiedene globale Tempi auf. Um diesen gerecht zu werden, erzeugen wir unterschiedliche Anfrageversionen, die zu verschiedenen Tempi korrespondieren. Die Tempoanpassungen werden durch Modifikation obiger Statistik-Analysefenster und der Downsampling-Faktoren simuliert. Zum Beispiel simuliert die Verwendung eines 5300-Millisekunden Analysefensters (anstelle 4100) und eines Downsampling-Faktors von 13 (anstelle 10) eine Tempoänderung um den Faktor $10/13 \approx 0.77$. In unseren Experimenten benutzen wir 8 verschiedene Anfrageversionen, welche globale Tempovariationen von -40 bis $+40$ Prozent abdecken. Für jede dieser Anfragen wird dann eine separate Abstandsfunktion Δ^j, $j \in [1 : 8]$, berechnet.

Seien nun $i_{\min} \in [1 : N - M + 1]$ und $j_{\min} \in [1 : 8]$ diejenigen Indizes, für die $\Delta^{j_{\min}}(i_{\min})$ minimal ist unter allen $\Delta^j(i)$. Die beste Übereinstimmung der Anfrage mit der Datenbank entspricht dann dem Musikausschnitt zur Teilfolge $(\vec{v}^{j_{\min}}_{i_{\min}}, \vec{v}^{j_{\min}}_{i_{\min}+1} \ldots, \vec{v}^{j_{\min}}_{i_{\min}+M-1})$. Zur Bestimmung des zweitbesten Treffers werden zur Vermeidung von Überschneidungen mit dem besten Treffer die in einer Umgebung von $i_{\min}$ liegenden Indizes für die weiteren Betrachtungen ausgeschlossen. Danach wird in analoger Weise fortgefahren, bis eine gewisse Anzahl an Treffern erreicht oder eine vorab festgelegte Abstandsschranke überschritten wird.

3.3 Experimente

Das Matchingverfahren wurde auf einem Datenbestand von 1167 Stücken (112 Stunden) klassischer Musik diverser Komponisten wie Bach, Beethoven, Chopin, Mozart, Ravel, oder Vivaldi getestet. Dabei liegen für die meisten Stücke mehrere Interpretation vor. Exemplarisch gehen wir auf das Vivaldi-Beispiel aus Abb. 2 ein, für das unsere Datenbank sieben verschiedenen Interpretationen (Abbado, Carmirelli, Lizzio, Mae, Nishizaki, Perlman, Zukerman) enthält. (Weiteres Material/Audiobeispiele zu den Experimenten findet man unter `http://www-mmdb.iai.uni-bonn.de/projects/audiomatching/`.) Für die besten sieben Treffer des Matchingverfahrens erhält man genau die Ab-

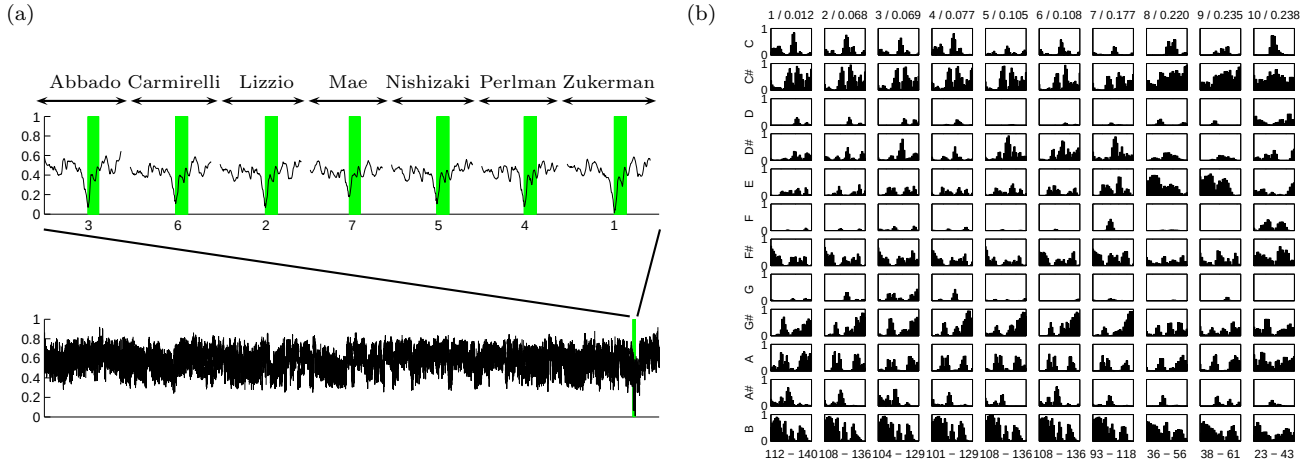

Abbildung 3: (a) Werte der Abstandsfunktion Δ bezüglich der in Abb. 2 (a) dargestellten Zukerman-Audioanfrage für die gesamte Testdatenbank (untere Abbildung) und für den die Treffer enthaltenden Teil der Testdatenbank (obere Abbildung). Die Treffer unterhalb des Δ-Schwellwerts von 0.2 sind durch die transparenten Balken gekennzeichnet. Diese sind gerade die relevanten Stellen der sieben in der Datenbank enthaltenen Interpretationen. (b) CENS-Merkmalsfolgen und Δ-Abstände für die ersten 10 Treffer bezüglich der Zukerman-Audioanfrage. Die ersten sieben Spalten korrespondieren zu den sieben Treffern in (a).

schnitte in den sieben Interpretationen, die den Takten 44–55 der Anfrage entsprechen. Abb. 3 (b) zeigt die CENS-Merkmalsvektoren der ersten 10 Treffer mit einer von links nach rechts aufsteigenden Distanz. Der beste Treffer auf Rang 1 stimmt (bis auf eine kleine durch die Auflösung der Merkmale bedingte Verschiebung) mit der Anfrage überein und weist einen Δ-Abstand von 0.012 (vgl. 1. Zeile von Abb. 3 (b)) auf. Die Position des Abschnitts innerhalb der Interpretation (Sekunden 112–140) findet man in der untersten Zeile. Die entsprechenden Parameter sind für die übrigen neun Treffer in analoger Weise angegeben. So hat der zweitbeste Treffer einen Δ-Abstand von 0.068 und entspricht den Sekunden 108–136 der Lizzio-Interpretation. Die Mae-Interpretation unterscheidet sich beträchtlich von der Anfrage hinsichtlich Artikulation, Tempo, und Notenrealisation (Mae spielt zusätzliche Verzierungen). Dennoch wird der entsprechende Abschnitt als siebter und letzter „korrekter“ Treffer mit einem Δ-Abstand von 0.177 identifiziert. Der achte und erste „falsche“ Treffer weist schon einen Δ-Abstand von 0.220 auf und korrespondiert zu den Sekunden 36–56 der Zukerman-Interpretation des dritten Satzes desselben Werks. Der zehnte Treffer korrespondiert gar zu einem Abschnitt von Bachs Sinfonia Nr. 12, BWV798 für Klavier. Auch wenn die „falschen“ Treffer oft keinen unmittelbaren Bezug zur Anfrage aufzuweisen scheinen, gibt es bezüglich des groben harmonischen Verlaufs einen großes Maß an Übereinstimmung.

3.4 Fazit und Ausblick

Die Grundidee des Audiomatchings basiert darauf, die erwünschten Invarianzen in die Merkmale selbst zu integrieren, um auf diese Weise robuste und effiziente Matchingverfahren anwenden zu können. Die hier vorgestellten CENS-Merkmale sind sehr grob, wodurch sich im Allgemeinen unter den besten Treffern auch eine Reihe von „falschen“ Treffern einschleichen. Um diese zu eliminieren müssen in einem Nachverarbeitungsschritt feinere Kriterien herangezogen werden, deren Berechnung allerdings nur noch auf der stark reduzierten Treffermenge durchzuführen sind. Aktuelle Forschungsergebnisse (Veröffentlichung in Vorbereitung) zeigen weiterhin, dass die Effizienz des Audiomatchings durch geeignete Indexierung der CENS-Merkmale deutlich gesteigert werden kann.

4 Strukturanalyse

Während die Musiksynchronisation und das Audiomatching dazu eingesetzt werden können, um zwischen verschiedenen Versionen eines Musikstücks hin- und herzuspringen, diskutieren wir nun die Fragestellung der *Strukturanalyse*, auf deren Basis die Navigation innerhalb eines Musikstücks ermöglicht wird. Ein Hauptziel der *Strukturanalyse von Musikstücken* ist die automatische Erkennung sich wiederholender Strukturen beziehungsweise die Bestimmung der musikalischen Form. Als Beispiel sei Brahms' Ungarischer Tanz Nr. 5 angeführt. Dieses Stück hat die musikalische Form $A_1A_2B_1B_2CA_3B_3B_4D$, bestehend 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 und einem kurzen Schlussteil D, siehe Abb. 4. Eine solche musikalische Struktur kann dann dem Benutzer durch ein geeignetes Navigationssystem zugänglich gemacht werden, das ihm im Fall des Brahmsbeispiels beliebig zwischen den A-Teilen hin- und herspringen oder direkt den Mittelteil C ansteuern lässt.

Die automatisierte Strukturanalyse im Audiobereich stellt ein reges Forschungsgebiet dar, siehe unter anderem [BW05, CF02, DH02, Go02, LWH04, MCKS04, MK06b, PBR02]. Eine Hauptschwierigkeit bei dieser Aufgabe besteht darin, dass musikalisch ähnliche Abschnitte erhebliche Variationen hinsichtlich Parametern wie Dynamik, Klangfarbe, der Spielweise von Notengruppen (z. B. Triller, Verzierungsnoten, Arpeggien), Tonhöhe (z. B. Modulationen), oder Tempo (z. B. Artikulation, Ritardandi, Accelerandi) aufweisen können. Zum Beispiel dirigiert Ormandy in seiner Interpretation des Ungarischen Tanzes den B_2-Teil wesentlich schneller als den B_1-Teil. Im folgenden skizzieren wir eine harmoniebasierte Strategie zur Strukturanalyse, siehe [MK06b].

4.1 Kostenmatrizen

Die meisten Ansätze zur automatisierten Strukturanalyse basieren auf so-

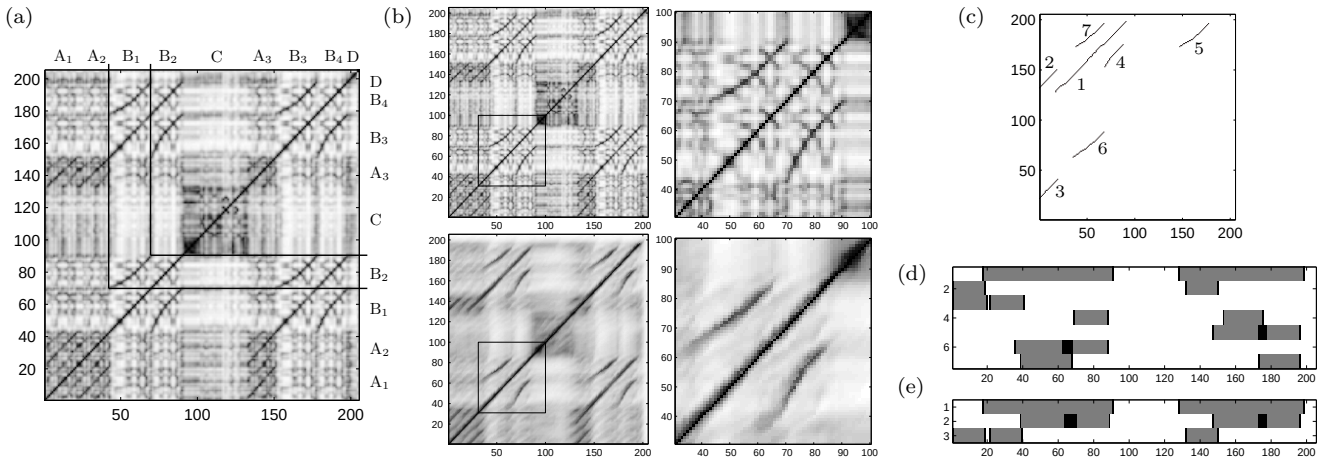

Abbildung 4: (a) Kostenmatrix einer CD-Aufnahme von Brahms' Ungarischem Tanz Nr. 5 dirigiert von Ormandy. (Einheiten in Sekunden). Das Stück hat die musikalische Form $A_1A_2B_1B_2CA_3B_3B_4D$. (b) Kostenmatrix mit vergrössertem Auschnitt vor (obere Reihe) und nach (untere Reihe) der Verbesserung der Pfadstruktur. (c) Extrahierte Pfade. (d) Zu den Pfaden korrespondierende Paare ähnlicher Abschnitte repräsentiert durch die horizontalen Balken. Dabei entsprechen die Reihen den Pfadnummern aus (c). Überlappungen der Abschnitte sind schwarz gekennzeichnet. (e) Abgeleitete globale Wiederholungsstruktur.

genannten *Kostenmatrizen* (häufig findet man in der Literatur auch den Begriff der *Ähnlichkeitsmatrix*). Das zu analysierende Audiosignal wird in einem ersten Schritt in eine Folge $\vec{V} := (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^N)$ von Merkmalsvektoren $\vec{v}^n \in \mathcal{F}$, $1 \leq n \leq N$, transformiert, wobei $\mathcal{F}$ einen geeigneten Merkmalsraum bezeichnet. Weiterhin sei $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ ein Kostenmaß auf $\mathcal{F}$. Die quadratische *Kostenmatrix* $\mathcal{S}$ ist dann durch $\mathcal{S}(n,m) := d(\vec{v}^n, \vec{v}^m)$ für $1 \leq n, m \leq N$ definiert. Mit anderen Worten werden also alle Merkmale $\vec{v}^n$ und $\vec{v}^m$ paarweise miteinander verglichen.

Im folgenden werden wir den Raum $\mathcal{F}$ der in Abschnitt 3.1 beschriebenen CENS-Merkmalsvektoren betrachten, die sich aufgrund ihrer angesprochenen Invarianzen besonders gut für die Strukturanalyse harmoniebasierter Musik eignen. Es sei daran erinnert, dass jeder CENS-Vektor einer Sekunde im Audiosignal entspricht. Das Kostenmaß d sei definiert als die Differenz von Eins und dem Cosinus des Winkels der zu vergleichenden CENS-Vektoren. Damit ist d nahe Null (niedrige Kosten) für ähnliche und nahe Eins (hohe Kosten) für sich unterscheidende (orthogonale) CENS-Vektoren.

Abb. 4 (a) zeigt die resultierende Kostenmatrix für eine CD-Aufnahme von Brahms' Ungarischem Tanz Nr. 5 dirigiert von Ormandy. Offensichtlich sind alle Kosten entlang der Hauptdiagonale identisch Null, da hier jeder Merkmalsvektor mit sich selbst verglichen wird. Die grundlegende Beobachtung ist nun, dass Paare von ähnlichen Teilfolgen in $\vec{V}$ zu Pfaden niedriger Kosten entlang von Nebendiagonalen korrespondieren. Zum Beispiel besagt der Pfad von Position $(1, 22)$ bis $(22, 42)$ (in Sekunden) in Abb. 4 (a), dass der Abschnitt $[1, 22]$ im Audiosignal ähnlich dem Abschnitt $[22, 42]$ ist. Eine manuelle Prüfung ergibt, dass der Abschnitt $[1, 22]$ gerade dem A_1-Teil und der Abschnitt $[22, 42]$ gerade dem A_2-Teil des Ungarischen Tanzes entspricht. Analog offenbart der nach oben gebogene Pfad von $(42, 69)$ bis $(69, 89)$ die Änlichkeit der Abschnitte $[42, 69]$ (Teil B_1) und $[69, 89]$ (Teil B_2). Hierbei ist in der Ormandy-Interpretation der B_2-Teil (20 Sekunden) wesentlich schneller als im B_1-Teil (27 Sekunden). Diese Tatsache drückt sich im Gradienten des Pfads aus, welcher den lokalen Tempounterschied zwischen den Abschnitten kodiert.

4.2 Strukturextraktion

Aus der vorherigen Diskussion geht hervor, dass ein schräg nach oben verlaufender Pfad niedriger Kosten einem Paar von ähnlichen Audiosegmenten entspricht. Der nächste Schritt der Strukturanalyse besteht in der Extraktion dieser Pfadstruktur aus der Kostenmatrix der zugrundeliegenden Audioaufnahme. Allerdings ist dieser Schritt im Allgemeinen problematisch, da die Pfade Verästelungen (z. B. durch mehrfaches Auftreten kurzer Wiederholungsmuster) und Unstetigkeiten (z. B. durch kleine lokale Unterschiede in sich entsprechenden Abschnitten) aufweisen oder durch Plateaus gleichbleibend niedriger Kosten (z. B. bei Segmenten gleichbleibender Harmonien) verlaufen können. Im Hinblick auf eine robuste Pfadextraktion wird daher zunächst die diagonale Pfadstruktur der Kostenmatrix herausgearbeitet. Die Strukturverbesserung kann dadurch erreicht werden, dass bei der Berechnung eines Matrixeintrags der lokale Kontext des Audiosignals mitberücksichtigt wird, siehe [MK06a]. Grob gesprochen entspricht dies einer Tiefpass-Filterung der Kostenmatrix entlang der Nebendiagonalen. Der Glättungseffekt wird durch Abb. 4 (b) illustriert.

Aus der so verbesserten Kostenmatrix können nun die Pfade mit einer einfachen, aber effizienten Greedy-Strategie unter Verwendung geeigneter Schwellwerte extrahiert werden. Es sei bemerkt, dass aufgrund der Symmetrie der Kostenmatrix nur der Teil oberhalb der Hautpdiagonalen betrachtet werden muss. Als Ausgangspunkt für die Konstruktion eines Pfads dient ein Matrixeintrag minimaler Kosten. Dieser wird sukzessive mittels geeignet vorgegebener Schrittweiten nach links unten und rechts oben verlängert, solange die Kosten der Verlängerung unter einem gewissen Schwellwert liegen. Kann der Pfad nicht mehr verlängert werden, wird eine gewisse Umgebung des Pfades für das weitere Vorgehen ausgeschlossen. Die Extraktion weiterer Pfade erfolgt dann in analoger Weise bis alle Matrixeinträge niedriger Kosten abgearbeitet sind. In einem Nachverarbeitungsschritt werden die so extrahierten Pfade verbessert und kurze Pfadfragmente verworfen. Abb. 4 (c) zeigt die extrahierten Pfade für die Ormandy-Interpretation. Jeder der sieben Pfade kodiert die Ähnlichkeit jeweils zweier Abschnit-

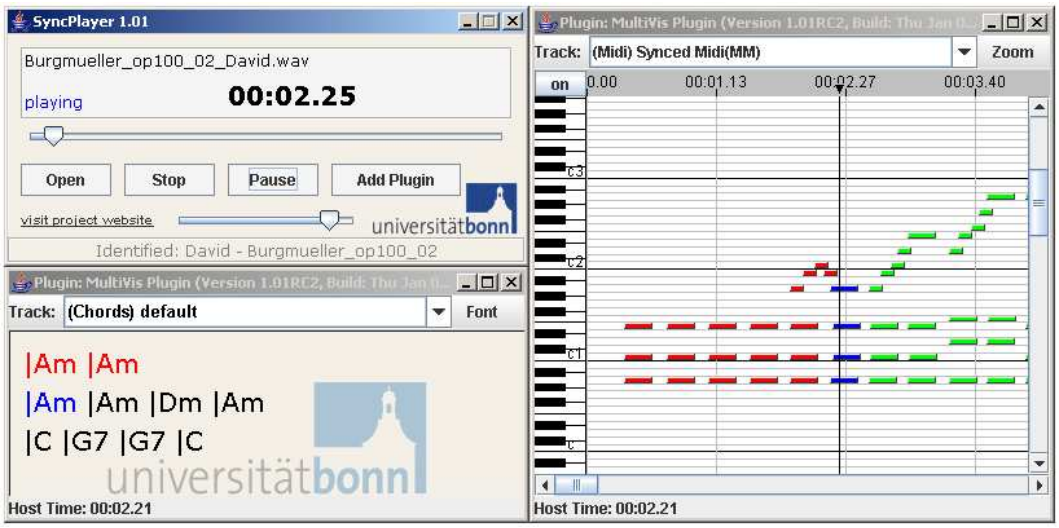

Abbildung 5: Benutzerseitige Komponente des SyncPlayer-Systems bei der Wiedergabe einer Audioaufnahme der Burgmüller-Etüde von Abb. 1. Links oben ist der eigentliche Audioplayer zu sehen. Weiterhin sind zwei der verfügbaren Visualisierungs Plug-Ins dargestellt. Rechts ist die Darstellung der an der Wiedergabeposition vorkommenden Noten in Klavierwalzendarstellung zu sehen, während links unten korrespondierende musikalische Akkorde dargestellt sind. Die aktuelle Wiedergabeposition ist jeweils farblich bzw. durch einen zusätzlichen Marker (Klavierwalzendarstellung) hervorgehoben.

te, siehe Abb. 4 (d). Zum Beispiel offenbart Pfad 1 die Ähnlichkeit zwischen Abschnitt $[20, 89]$ (entspricht $A_2B_1B_2$) und Abschnitt $[127, 195]$ (entspricht $A_3B_3B_4$).

In einem letzten Schritt wird nun die globale Wiederholungsstruktur der Audioaufnahme aus den paarweise gegebenen Ähnlichkeitsrelationen abgeleitet. Auch dieser Schritt ist nicht einfach, da die extrahierte Pfadstruktur im Allgemeinen Fehler aufweist wie z. B. Ungenauigkeiten in den Pfadlängen, lückenhafte Pfade, oder bedeutungslose, nicht verworfene Pfadfragmente. Zum Ausgleich dieser Fehler werden Clusteringmethoden in Verbindung mit einem Transitivitätsschritt eingesetzt, siehe [MK06b]. Die sich ergebende globale Struktur ist für das Brahmsbeispiel in Abb. 4 (e) dargestellt. Sowohl die vier Wiederholungen der B-Teile (zweite Zeile) als auch drei Wiederholungen der A-Teile (dritte Zeile) wurden trotz erheblicher Variationen richtig erkannt. Die erste Zeile entspricht der Wiederholung $A_2B_1B_2$ und $A_3B_3B_4$.

Um auch transponierte Wiederholungen (z. B. die Wiederholung eines Themas in der Dominante) erfassen zu können, wurde obige Strukturanalyse grob gesprochen mit zwölf verschiedenen Versionen der CENS-Merkmale durchgeführt. Die Hauptidee beruht dabei auf der Beobachtung, dass die zwölf zyklischen Shifts der 12-dimensionalen CENS-Merkmale genau den zwölf möglichen Transponierungen (modulo Oktave) entsprechen, siehe [Go02, MK06b].

4.3 Fazit und Ausblick

Das vorgestellte Verfahren wurde in MATLAB implementiert und für etwa 100 CD-Aufnahmen (inbesondere klassische Musik) getestet. (Eine repräsentative Auswahl an Ergebnissen ist unter `http://www-mmdb.iai.uni-bonn.de/projects/audiostructure/` verfügbar.) Für die meisten Stücke entspricht die automatisch extrahierte Wiederholungsstruktur der musikalischen Form des zugrundeliegenden Musikstücks. Hierbei erweist sich das Verfahren als sehr robust gegenüber relativen Änderungen in der Klangfarbe, der Lautstärke, oder des Tempos. Unterscheiden sich allerdings musikalisch ähnliche Abschnitte in ihrem harmonischen Verlauf (z. B. Wiederholung eines Dur-Themas in moll oder Hinzufügen eines ganzen Takts in der Mitte der Wiederholung), funktioniert das beschriebene Analyseverfahren nicht mehr. Um den vielfältigen Aspekten von Musik Rechnung zu tragen, sollten daher harmoniebasierte Merkmale mit anderen Merkmalstypen (Dynamik, Rhythmus, Timbre) kombiniert werden. Eine weitere anspruchsvolle Aufgabe besteht in der Entwicklung von Verfahren zur hierarchischen Strukturanalyse, bei der Wiederholungen auf zeitlich unterschiedlichen Auflösungstufen (z. B. Wiederholungen ganzer Passagen, einzelner Themen, Phrasen, oder Motive) erkannt und entsprechend angeordnet werden können.

5 SyncPlayer

Die vorgestellten Techniken der automatisierten Musikdatenerschließung erlauben die Entwicklung von leistungsfähigen Systemen zum inhaltsbasierten Retrieval in umfangreichen Musikdatenbeständen, zur Navigation innerhalb und zwischen Musikdokumenten, sowie zur synchronen Darstellung musikrelevanter Informationen. Das SyncPlayer-System [KM05] stellt einen ersten Prototypen für ein solches System dar, dessen Funktionsweise im folgenden zusammengefasst wird.

Der Benutzer des Client-Server-basierten SyncPlayer-Systems kann mittels eines Audio-Players lokal auf seinem Rechner abgelegte Audioaufnahmen wiedergeben (Abb. 5, links oben). Während der Audiowiedergabe extrahiert das System lokal eine Merkmalsmenge aus dem Audiodokument und überträgt diese zur Server-Anwendung. Dort wird, basierend auf einer vorab durchgeführten Indexierung einer großen Audiokollektion, eine Audioidentifikation anhand der übermittelten Merkmalsmenge durchgeführt. Eine wichtige Eigenschaft des hierzu von uns verwendeten Identifikationsalgorithmus [CK03a] ist, dass nicht nur die Audioaufnahme an

sich, sondern auch die exakte Position des aktuell im SyncPlayer wiedergegebenen Audioausschnitts innerhalb der Audioaufnahme ermittelt werden kann.

Neben dem zur Audioidentifikation verwendeten Index sind auf dem Server zu jeder Aufnahme der Audiokollektion inhaltbasierte Daten abgelegt. Zu diesen Daten gehören unter anderem die von Synchronisationsalgorithmen automatisch erzeugten Verlinkungen zwischen Musikdatenströmen (z. B. Ergebnisse einer Audio-Partitur- oder Audio-Audio-Synchronisation) und die automatisch extrahierten Wiederholungsstrukturen. Wurde die vom Benutzer mit dem SyncPlayer wiedergegebene Audioaufnahme erfolgreich vom Server identifiziert, so erhält der SyncPlayer die Möglichkeit, die verfügbaren inhaltsbasierten Daten sukzessive abzufragen. Hierzu überträgt der Server eine Liste von für die identifizierte Audioaufnahme verfügbaren, in einzelnen *Tracks* organisierten Daten an den Client. Der Benutzer erhält auf Grundlage dieser Trackliste die Möglichkeit, verschiedene Plug-Ins zur Visualisierung auszuwählen. Derzeit existieren Plug-Ins für die Darstellung von Partiturdaten im MIDI-Format, von Gesangstexten, von Akkorden, von Audiostrukturdaten, sowie Plug-Ins zur Wellenform- und Spektraldarstellung. Abb. 5 zeigt zwei verschiedene Plug-Ins für die Audioaufnahme zu dem Burgmüller-Beispiel von Abb. 1. Im rechten Bildteil sind zur aktuellen Position in der Audioaufnahme zugehörige Partiturdaten im Klavierwalzenformat dargestellt, während im linken unteren Bildteil entsprechende Akkordinformationen zu sehen sind.

Während der akustischen Wiedergabe der Audioaufnahme werden nun, unter Ausnutzung der vorliegenden Verlinkung zwischen den Positionen der Audioaufnahme und den übermittelten inhaltsbasierten Daten, immer genau diejenigen Informationen angezeigt, die zu der aktuell wiedergegebenen Stelle der Aufnahme korrespondieren. In der Klavierwalzendarstellung werden hierzu z. B. die Noten in einer gewissen Umgebung der Wiedergabeposition dargestellt, wobei ein senkrechter Balken die exakte aktuelle Notenposition markiert. In der Akkorddarstellung werden die aktuellen Akkorde farblich markiert. Ein weiteres Plug-In ermöglicht die Navigation mittels der zuvor extrahierten Audiostruktur, wobei die Visualisierung der Struktur analog zu Abb. 4 (e) erfolgt. Aktuell in Entwicklung befindet sich ein Plug-In-Modul zur synchronen Darstellung gescannter (gedruckter oder handschriftlicher) Partituren. Zur Navigation zwischen verschiedenen CD-Aufnahmen zu einem musikalischen Werk entwickeln wir ein Plug-In, das dem Benutzer bei der Audiowiedergabe erlaubt, zwischen unterschiedlichen Interpretationen des Musikwerks unter Beibehaltung der aktuellen musikalischen Zeitposition hin- und herzuspringen.

Das SyncPlayer-System wird durch eine weitere Komponente zur Volltextsuche in der Gesangsstimme von Musikstücken ergänzt. Hierbei werden unter Verwendung geeigneter Synchronisationsdaten von der Suchmaschine alle Musikstücke samt der exakten Positionen identifiziert, an denen die angefragten Stichworte oder Textpassagen vorkommen. Als besondere Eigenschaft des Systems ist hier hervorzuheben, dass mittels einer Textanfrage direkt an eine entsprechende Stelle einer CD-Aufnahme gesprungen werden kann. Eine entsprechende Funktionalität wird derzeit für die inhaltsbasierte Notensuche in das System integriert. Hierbei werden nach einer rein symbolischen Suche in polyphonen Partiturdaten mittels zuvor berechneter Partitur-Audio-Synchronisationsdaten die entsprechenden Treffer in den CD-Aufnahmen identifiziert.

Danksagung

Wir danken der Deutschen Forschungsgemeinschaft für die finanzielle Unterstützung bei der Entwicklung des SycPlayer-Systems im Rahmen des DFG-Leistungszentrums Probado DFG-GZ: 554 975 (1).

Literatur

[AHFC01] E. Allamanche, J. Herre, B. Fröba, and M. Cremer: AudioID: Towards Content-Based Identification of Audio Material. Proc. *110th AES Convention*, Amsterdam, NL, 2001.

[ACKM03] Vlora Arifi, Michael Clausen, Frank Kurth, Meinard Müller: Synchronization of Music Data in Score-, MIDI- and PCM-Format. Computing in Musicology 13, MIT Press, 9–33, 2004.

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

[CBMN02] Pedro Cano, Eloi Battle, Ton Kalker, and Jaap Haitsma: A Review of Audio Fingerprinting. *Proc. 5. IEEE Workshop on MMSP, St. Thomas, Virgin Islands, USA*, 2002.

[CK03a] Michael Clausen, Frank Kurth: A Unified Approach to Content-Based and Fault Tolerant Music Recognition. *IEEE Trans. on Multimedia*, 6(5), 717–731, 2004.

[CF02] Matthew Cooper, Jonathan Foote: Automatic Music Summarization via Similarity Analysis. Proc. *3th ISMIR*, Paris, France, 2002.

[DH02] Roger Dannenberg, Ning Hu: Pattern Discovery Techniques for Music Audio. Proc. *3th ISMIR*, Paris, France, 2002.

[DW05] Simon Dixon, Gerhard Widmer: Match: A music alignment tool chest. Proc. *6th ISMIR*, London, GB, 2005.

[Go02] Masataka Goto: A Chorus-Section Detecting Method for Musical Audio Signals. Proc. IEEE ICASSP, 437–440, 2003.

[HDT03] Ning Hu, Roger Dannenberg, George Tzanetakis: Polyphonic audio matching and alignment for music retrieval. Proc. *IEEE WASPAA*, New Paltz, NY, 2003.

[KM05] Frank Kurth, Meinard Müller, David Damm, Christian Fremerey, Andreas Ribbrock, Michael Clausen: SyncPlayer–An Advanced System for Multimodal Music Access Proc. *6th ISMIR*, London, GB, 2005.

[LWH04] Lie Lu, Muyuan Wang, Hong-Jiang Zhang: Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data. Proc. ACM Multimedia, NY, USA, 275–282, 2004.

[MCKS04] Namunu C. Maddage, Changsheng Xu, Mohan S. Kankanhalli, Xi Shao: Content-based music structure analysis with applications to music semantics understanding. Proc. ACM Multimedia, NY, USA, 112–119, 2004.

[MKR04] Meinard Müller, Frank Kurth, Tido Röder: Towards an Efficient Algorithm for Automatic Score-to-Audio Synchronization. Proc. *5th ISMIR*, Barcelona, Spain, 2004.

[MKC05] Meinard Müller, Frank Kurth, Michael Clausen: Audio Matching via Chroma-bases Statistical Features. Proc. *6th ISMIR*, London, GB, 2005.

[MK06a] Meinard Müller, Frank Kurth: Enhancing Similarity Matrices for Music Audio Analysis. Proc. IEEE ICASSP, 2006.

[MK06b] Meinard Müller, Frank Kurth: Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations. to appear in: Proc. *7th ISMIR*, Victoria, Canada, 2006.

[MMK06c] Meinard Müller, Henning Mattes, Frank Kurth: An Efficient Multiscale Approach to Audio Synchronization. to appear in: Proc. *7th ISMIR*, Victoria, Canada, 2006.

[PBR02] Geoffroy Peeters, Amaury La Burthe, Xavier Rodet: Toward Automatic Music Audio Summary Generation from Signal Analysis. Proc. *3th ISMIR*, Paris, France, 2002.

[RJ93] Rabiner, L. R., Juang, B. H.: Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series, 1993.

[Ra04] Christopher Raphael: A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores. Proc. *5th ISMIR*, Barcelona, Spain, 2004.

[SRS03] Ferréol Soulez, Xavier Rodet, Diemo Schwarz: Improving polyphonic and poly-instrumental music to score alignment. Proc. *4th ISMIR*, Baltimore, Maryland, 2003.

[TE03] Robert Turetsky, Dan Ellis: Ground-truth transcriptions of real music from force-aligned midi syntheses. Proc. *4th ISMIR*, Baltimore, Maryland, 2003.

[Wa03] A. Wang: An Industrial Strength Audio Search Algorithm. Proc. *4th ISMIR*, Baltimore, USA, 2003.