Neue Entwicklungen im Bereich des Music Information Retrieval

Meinard Müller

Universität des Saarlandes und Max-Planck Institut für Informatik, Campus E1.4, 66123 Saarbrücken, Germany
E-Mail: meinard@mpi-inf.mpg.de
Web: http://www.mpi-inf.mpg.de/~mmueller/

Zusammenfassung

Bei der automatisierten Verarbeitung von Musikdaten steht man aufgrund der Vielfältigkeit von Musik in Form und Inhalt vor großen Herausforderungen. Bei dem noch jungen Forschungsgebiet des *Music Information Retrieval* (MIR) geht es um die Entwicklung von Methoden und Systemen, die Benutzern große, in digitaler Form vorliegende Musikkollektionen in vielfältiger Weise zugänglich machen. In diesem Beitrag geben wir einen Überblick über neue Entwicklungen im MIR-Bereich und diskutieren dabei Aspekte der inhaltsbasierte Musikanalyse, der Aufbereitung und Suche von Musikdaten, sowie der Entwicklung von neuartigen Benutzerschnittstellen.

1 Inhaltsbasierte Musikanalyse

Durch zunehmende Digitalisierung von Musikdaten aller Art sind in den letzten Jahren umfangreiche, oft unstrukturierte Musikdatenbeständen enstanden. In realen Anwendungsszenarien sind diese Bestände im allgemeinen heterogen und enthalten Bild-, Ton- und Textinhalte unterschiedlicher Formate. Man denke hier beispielsweise an CD-Aufnahmen diverser Interpreten, Noten, MIDI-Daten, Musikvideos oder Gesangstexte. Allgemein gesprochen ist das Hauptziel des *Music Information Retrieval* (MIR) die Nutzbarmachung solch multimodaler 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.

Wer Musikdatenbanken durchstöbert, will zum Beispiel wissen, wie der Komponist eines bestimmten Musikstücks heißt, welche Songtitel vorhanden sind oder zu welchem Gesamtwerk ein einzelnes Stück gehört. Solche Suchanfragen kann man schon heute mit klassischen Datenbanktechniken beantworten. Komplizierter wird es, wenn man anhand musikalischer Inhalte in Musikkollektionen suchen und navigieren möchte. Wer etwa nach einem Songtitel sucht, könnte dem Computer ein Melodiefragment vorpfeifen und danach suchen lassen. Oder er könnte unter Angabe eines akustischen Musikausschnitts danach fragen, in welchen Musikaufnahmen ähnliche Passagen vorkommen. Ein Musikwissenschaftler könnte sich außerdem dafür interessieren, wo bestimmte Notenkonstellationen, Harmonieverläufe oder Rhythmen zu finden sind, einschließlich der genauen Zeitpositionen innerhalb der jeweiligen Aufnahmen. Zur Bearbeitung solcher Suchanfragen benötigt man Methoden der *inhaltsbasierten* Musikanalyse, bei der nur auf die Rohdaten selbst und nicht etwa auf manuell erstellte und den Rohdaten beigefügte Annotationen zurückgegriffen wird. Im folgenden werden wir aktuelle MIR-Fragestellungen diskutieren, die eng mit der inhaltsbasierten Analyse von Musikdaten verknüpft sind.

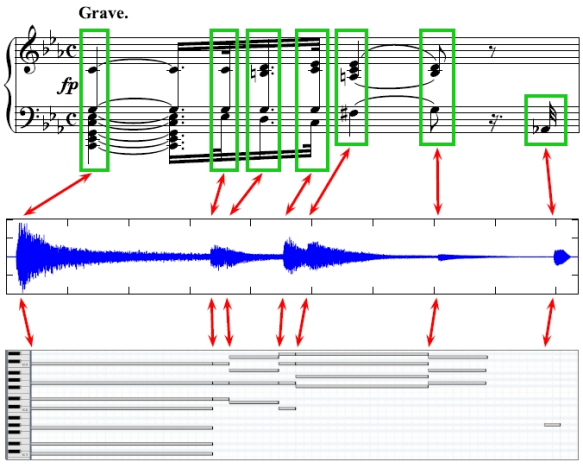

Abbildung 1: Verlinkung von verschiedenen Darstellungen unterschiedlicher Modalitäten (Notentext, Audio, MIDI) zu einem Musikstück. Dargestellt sind die ersten Takte der Klaviersonate Op. 13 (Pathétique) von Beethoven.

2 Multimodale Verlinkung

Bei der Entwicklung inhaltsbasierter Such- und Navigationsmechanismen führt die oben angesprochene Multimodalität und Komplexität existierender Musikdokumentensammlungen zu großen, weitgehend noch ungelösten Problemen. Eine entscheidende Rolle kommt hier der umfassenden Annotation, Verlinkung und Strukturierung des Datenbestandes zu, was allerdings aufgrund der enormen Datenmassen manuell nicht bewerkstelligt werden kann. Genau an diesem Punkt setzt die Aufgabenstellung der *Musiksynchronisation* an, bei der es um die automatische Verlinkung zweier Datenströme unterschiedlicher Formate geht [10]. Anschaulich können solche Verfahren 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 einem Notentext) bestimmen, siehe Abbildung 1.

Solche Verlinkungsstrukturen können unter anderem zur multimodalen Musiknavigation und zum Vergleich unterschiedlicher Interpretation eingesetzt werden. Exemplarisch zeigt Abbildung 2 eine Benutzerschnittstelle zur multimodalen Wiedergabe von Musik [4]. Basierend auf zuvor berechneten Verlinkungsstrukturen werden hierbei synchron zur akustischen Wiedergabe einer Audioaufnahme die musikalisch korrespondierenden Takte in einem Notentext visuell hervorgehoben. Weiterhin eröffnen Synchronisationstechniken neuartige Möglichkeiten der Navigation zwischen verschiedenen Aufnahmen eines Musikstücks. Hierbei kann der Benutzer mittels geeigneter Slider beim Abspielen von Aufnahmen nahtlos zwischen den unterschiedlichen Interpretationen hin- und herzuwechseln. Diese Funktionalität kann mit dem Anzeigen von Suchergebnissen (Abbildung 2) oder zuvor extrahier-

ter struktureller Informationen (Abbildung 3) kombiniert werden. Ein solcher *Interpretationswechsler* wurde unter anderem für das Digitale Beethoven-Haus[1] in Bonn in Form eines digitalen Exponats entwickelt. Dort stehen dem Musikliebhaber momentan 27 verschiedene Aufnahmen der berühmten "Appassionata"-Klaviersonate von Ludwig van Beethoven zur Verfügung, die man mit Hilfe des Interpretationswechsler direkt miteinander vergleichen kann.

Die meisten Verfahren zur *Musiksynchronisation* gehen in zwei Schritten vor [10]. Im ersten Schritt werden die zu verlinkenden Datenströme in geeignete Merkmalsdarstellungen umgewandelt, um hierdurch zum einen eine Datenreduktion und zum anderen Robustheit gegenüber nicht zu berücksichtigenden Variabilitäten zu erzielen. Im Musikkontext werden insbesondere *Chromamerkmale* mit großem Erfolg für unterschiedliche Retrieval- und Analyseaufgaben eingesetzt [2, 10]. Diese Merkmale korrelieren stark mit dem Harmonieverlauf des zugrundeliegenden Musikstücks und weisen einen hohen Grad an Robustheit gegenüber Änderungen in Instrumentierung, Dynamik, Klangfarbe und Artikulation auf. Insbesondere eignen sich chromabasierte Merkmale als gemeinsame Mid-Level Darstellung für sowohl akustische als auch symbolische Musikrepräsentationsformen und erlauben damit eine Verlinkung multimodal vorliegender Versionen. Im zweiten Schritten werden dann die beiden extrahierten Merkmalsfolgen mittels Alignment-Verfahren wie *Dynamic Time Warping* (DTW) oder *Hidden Markov Modelle* (HMM) – beides Techniken, die im Bereich der Sprachsignalverarbeitung entwickelt wurden [15] – synchronisiert.

3 Audio Retrieval

Insbesondere die Analyse von auf Wellenformen basierenden Audiodaten ist im Hinblick auf effizientes und effektives Musikretrieval von fundamentaler Bedeutung. Exemplarisch wollen wir auf drei aktuelle inhaltsbasierte Suchaufgaben eingehen.

Aufgabe der *Audioidentifikation* (auch als *Audio-Fingerprinting* bezeichnet) ist es, einen gegebenen Ausschnitt einer Audioaufnahme als Bestandteil genau dieser Aufnahme zu erkennen [1, 9, 17]. Mit Hilfe der Audioidentifikation kann man also exakt ermitteln, in welchem Stück und an welcher Position einer Audio-CD ein bestimmter Audioausschnitt enthalten ist. Die effiziente Audioidentifikation (unter Zulassung gewisser Signalverzerrungen und Kompressionsartefakte) kann als im wesentlichen gelöst angesehen werden und findet Anwendung in kommerziellen Produkten wie Shazam [2].

Die Fragestellung des *Audiomatching* kann als Verallgemeinerung der Audioidentifikation aufgefasst werden. Ausgangspunkt ist hier eine große Musikdatenbank, die typischer Weise mehrere verschiedene Aufnahmen desselben Musikstücks enthält. Diese Aufnahmen wurden dabei im Allgemeinen von unterschiedlichen Interpreten und in eventuell verschiedenen Instrumentierungen eingespielt. Bei Anfrage eines 10-30 sekündigen Audioausschnitts sollen dann automatisch alle musikalisch entsprechenden Passagen in allen in der Datenbank verfügbaren Interpretationen gefunden werden. Ein erster chromabasierter Ansatz zum Audiomatching wurde in [13] vorgeschlagen, der mittels indexbasierter Methoden wesentlich beschleunigt wurde [8]. Mit ähnlichen Techniken können auch multimodale Suchszenarien realisiert werden. Zum Beispiel zeigt die Abbildung 2 ein Szenario, bei dem die Anfrage aus in einem Notentext markierten Takten besteht; die Suche erfolgt dann aber in Audiodatenbeständen.

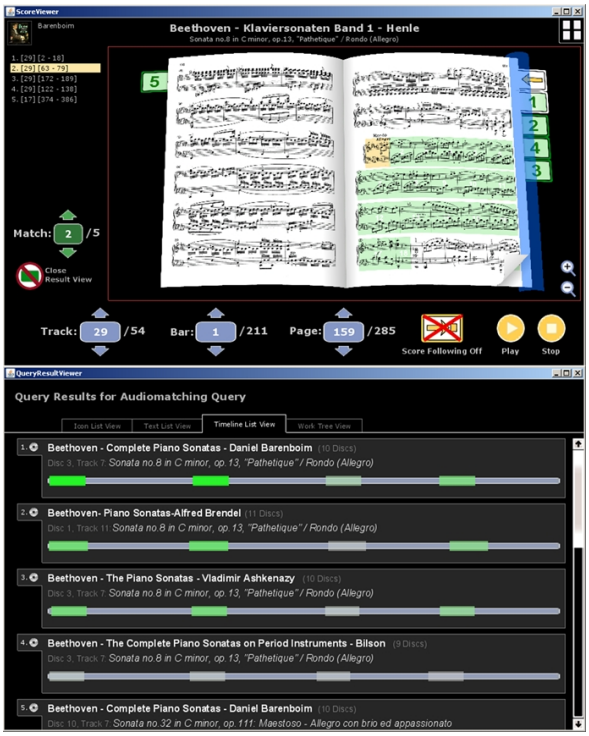

Abbildung 2: Der Score-Audio Viewer erlaubt eine simultane Präsentation eines Notentextes und die Wiedergabe von Audioaufnahmen. Dargestellt sind die ersten Takte des dritten Satzes (Rondo) von Beethovens Klaviersonate Op. 13 (Pathétique). Weiterhin kann multimodal gesucht werden. Bei einer visuellen Anfrage (grün markierte Takte; Thema des Rondos) werden alle Audioaufnahmen, die Treffer zu dieser Anfrage enthalten, aufgelistet. Hierbei kann eine Audioaufnahme mehrere Treffer (grüne Rechtecke; Thema erscheint vier mal im Rondo) enthalten.

Während es beim Audiomatching um das Auffinden von musikalisch in Beziehung stehenden Audioausschnitten geht, gibt es eine Reihe von verwandten Fragestellungen, bei denen die Ähnlichkeit ganzer Musikstücke zu bewerten ist. Besondere Beachtung findet hierbei die Fragestellung der *Cover Song Identifikation* mit dem Ziel zu einem gegebenen Song alle Varianten (u. a. Cover-, Remake-, Remix-Versionen) aufzufinden [3, 5, 16]. Auch hier wurden sehr erfolgreich chromabasierte Merkmale in Verbindung mit lokalen Alignmentverfahren (z. B. Smith-Waterman-Algorithmus) eingesetzt [16]. In [3] wird gezeigt, wie durch Shingling-Techniken und eine effiziente Nachbarschaftssuche unter Einsatz von *Locality Sensitive Hashing* (LSH) die Dokumentensuche wesentlich beschleunigt werden kann.

Ingesamt läßt sich feststellen, dass man es bei den diskutierten Aufgabenstellungen mit sich widersprechenden Prinzipien zu tun hat. Je spezifischer die Suchaufgabe ist, desto effizienter lässt sie sich unter Einsatz von Indexierungsmethoden lösen. Liegen große spektrale und insbesondere zeitliche Variabilitäten vor, müssen kostenintensive Alignmentverfahren eingesetzt werden, die sich nur schwer beschleunigen und damit auf große Datenkollektio-

[1] http://www.beethoven-haus-bonn.de/
[2] http://www.shazam.com/

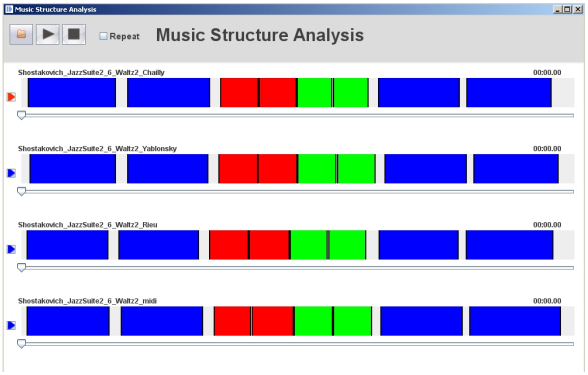

Abbildung 3: Benutzerschnittstelle zur Navigation innerhalb einer Audioaufnahme (mittels der Strukturblöcke) und zwischen unterschiedlichen Aufnahmen (Interpretationswechsler) eines Musikstücks. Die vier Slider entsprechen vier unterschiedlichen Interpretationen eines Walzers von Shostakovich. Über jedem Slider wird die jeweilige Wiederholungsstruktur mittels farblich kodierter Strukturblöcke angezeigt.

nen anwenden lassen. In diesen Fällen stellt die Skalierungung der Methoden auf Datenbestände mit Millionen von Dokumenten eine große Herausforderung mit noch vielen offenen Fragestellungen dar.

4 Strukturanalyse

Während die Musiksynchronisation dazu eingesetzt werden kann, 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 der zweite Walzer aus der "Suite for Variety Orchestra Nr. 1" von Dimitri Shostakovich angeführt, siehe Abbildung 3. Dieses Stück hat die musikalische Form $A_1A_2B_1B_2C_1C_2A_3A_4$, bestehend aus vier sich wiederholenden A-Teilen (blaue Blöcke) und aus jeweils zwei sich wiederholenden B-Teilen (rote Blöcke) und C-Teilen (grüne Blöcke). Eine solche musikalische Struktur kann dann dem Benutzer durch ein geeignetes Navigationssystem zugänglich gemacht werden, das ihm z. B. im Fall des Walzers beliebig zwischen den vier A-Teilen hin- und herspringen oder direkt den Mittelteil B_1 ansteuern lässt, siehe Abbildung 3.

Die automatisierte Strukturanalyse von Audioaufnahmen stellt ein reges Forschungsgebiet dar [2, 7, 12, 14]. Eine Hauptschwierigkeit bei dieser Aufgabenstellung besteht darin, dass musikalisch ähnliche Abschnitte erhebliche Variationen hinsichtlich Eigenschaften wie Dynamik, Klangfarbe, Spielweise bestimmter Notengruppen (z. B. Triller, Verzierungsnoten, Arpeggien), Tonhöhe (z. B. Modulationen) oder Tempo (z. B. Artikulation, Ritardandi, Accelerandi) aufweisen können. Zum Beispiel wird das Thema in den vier A-Teilen des obigen Walzers durch unterschiedliche Instrumente wiedergeben (Klarinette, Streicher, Posaune, Tutti). Darüber hinaus gibt es erhebliche Unterschiede zwischen den vier Teilen hinsichtlich der Begleitung und dem Vorliegen zusätzlicher Nebenstimmen. Um den unterschiedlichen Variabilitäten Rechnung tragen zu können,

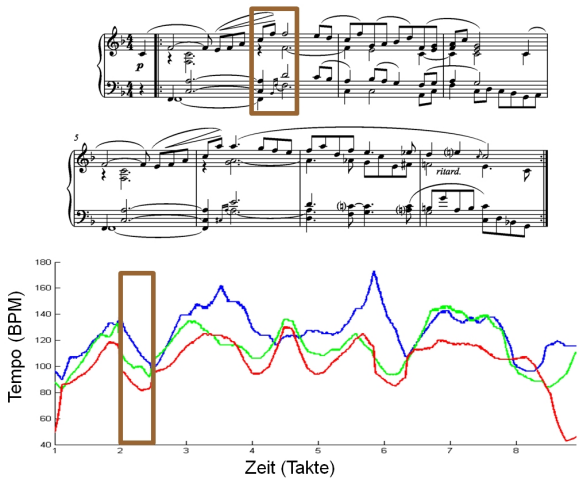

Abbildung 4: Tempoverläufe von drei unterschiedlichen Interpretationen zu Schumann's "Träumerei."

werden in einigen Arbeiten unterschiedliche Merkmale, die die Klangfarbe, die Harmonie und das Tempo betreffen, kombiniert [14]. In [7] wird vorgeschlagen, wie man die Strukturanalyse auch bei Vorliegen möglicher Transpositionen bewerkstelligen kann. Die meisten der bisherigen Verfahren setzen ein konstantes Tempo der Audioaufnahme voraus. In [12] wird ein erster Algorithmus beschrieben, der auch bei Temposchwankungen innerhalb einer Aufnahme gute Resultate bei der Strukturanalyse erzielt. Ein vielversprechender Ansatz zur kombinierten Strukturanalyse wird in [14] beschrieben, wobei sowohl homogenitätsbasierte Maße als auch wiederholungsbasierte Maße in einer Fitness-Funktion verquickt werden.

Beim Vorliegen signifikanter Variabilitäten stoßen allerdings bisherige Verfahren der Strukturfindung an ihre Grenzen. Die Entwicklung von kombinierten Methoden, mit denen man solche musikalisch begründeten Variabilitäten unter Berücksichtigung unterschiedlicher zeitlicher und polyphoner Hierarchiestufen besser in den Griff bekommen kann, ist Gegenstand aktueller und zukünftige Forschung.

5 Interpretationsanalyse

In den vorherigen Retrieval- und Strukturierungsaufgaben ging es darum, ähnliche musikalische Passagen trotz erheblicher Unterschiede in der jeweiligen Ausführung aufzufinden. Die automatisierte *Interpretationsanalyse* kann als eine quasi komplementäre Aufgabestellung angesehen werden, bei der es darum geht die Unterschiede und Gemeinsamkeiten in verschiedenen Interpretationen eines Musikstücks zu erfassen [18]. Hierbei sei bemerkt, dass der Notentext ein Musikstück nicht völlig festlegt, sondern dem Musiker künstlerische Freiheiten hinsichtlich der konkreten Ausführung lässt. So können unterschiedliche Aufnahmen zu einem Musikstück erheblich in Tempo, Phrasierung, Betonung, Agogik und Lautstärke variieren. Auf der anderen Seite gibt es häufig Gemeinsamkeiten über unterschiedliche Interpretationen hinweg, die auf allgemeine Interpretationsregeln schließen lassen.

Gängige Verfahren der automatisierten Interpretationsanalyse gehen in zwei Schritten vor. Im ersten Schritten werden aus den Audiodaten musikalische Parameter wie Noteneinsatzzeiten und Laustärke extrahiert. Dieser

Schritt erfolgt oft semi-automatisch oder gar manuell, da automatisierte Verfahren oft die nötige zeitliche Exaktheit vermissen lassen. Im zweiten Schritt werden dann auf Basis dieser Parameter die Gemeinsamkeiten und Unterschiede in den Tempo- und Dynamikverläufen untersucht. Zur Illustration zeigt Abbildung 4 die Tempoverläufe von drei Interpretationen der "Träumerei" von Robert Schumann. Trotz erheblicher Unterschiede bei der Tempowahl können auch Gemeinsamkeiten im relativen Verlauf der Kurven festgestellt werden. Zum Beispiel werden zu Beginn von Takt 2 (braun markierter Bereich) alle Interpreten langsamer, was auf die aufsteigende Melodielinie mit einem lokalen Höhepunkt auf der Subdominante (B-Dur) zurückzuführen ist. Nach diesem Höhepunkt kann man bei allen Interpretationen eine erhebliche Beschleunigung festellten.

In [11] wird ein auf Musiksynchronisation basierendes Verfahren beschrieben, das den Tempoverlauf einer Interpretation vollautomatsich bestimmt. Hierbei ist die Idee eine uninterpretierte MIDI-Datei mit der jeweiligen Audioaufnahme zu verlinken; der Tempoverlauf kann dann aus dem Gradienten des Alignmentpfads abgeleitet werden. Allerdings benötigt man für dieses Verfahren zeitlich hochauflösende und zugleich robuste Synchronisationsverfahren – ein für sich schwieriges Forschungsproblem [6].

6 Ausblick

Das interdisziplinärene Gebiet des *Music Information Reterival* hat sich in den letzten zehn Jahren zu einem eigenständigen und regem Forschungsbereich entwickelt, der sich neben klassischen Suchaufgaben mit ganz unterschiedlichen Aspekten der inhaltsbasierten Analyse von Musikdaten beschäftigt. Aufgrund der Komplexität und Vielschichtigkeit von Musik benötigt man intelligente Methoden und Werkzeuge, die Beziehungen zwischen den verschiedensten Modalitäten, Versionen und Interpretationen eines Musikstücks trotz erheblicher akustischer und musikalischer Variabiliäten erkennen und herstellen können. Eine vielverspechende Forschungsrichtung besteht daher in der Entwicklung *mehrschichtiger* Analyseverfahren, die *simultan* unterschiedliche zeitliche Auflösungsstufen, verschiedene musikalische Aspekte (z. B. Zeit, Rhythmus, Dynamik, Harmonie, Klangfarbe), und mannigfaltig vorliegende Versionen eines Musikstücks berücksichtigen.

Danksagung. Meinard Müller wird durch das Exzellenzcluster *Multimodal Computing and Interaction* (MMCI) der Universität des Saarlandes unterstützt. Ein spezieller Dank geht an die Arbeitsgruppe von Prof. Michael Clausen, Universität Bonn, für die langjährige Kooperation im Bereich der Musikinformatik.

Literatur

[1] Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, and Markus Cremer. AudioID: Towards content-based identification of audio material. In *Proc. 110th AES Convention*, Amsterdam, NL, 2001.

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

[3] Michael Casey, Christophe Rhodes, and Malcolm Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech & Language Processing*, 16(5), 2008.

[4] David Damm, Christian Fremerey, Frank Kurth, Meinard Müller, and Michael Clausen. Multimodal presentation and browsing of music. In *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI)*, pages 205–208, Chania, Crete, Greece, October 2008.

[5] Daniel P. W. Ellis and Graham. E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, Honolulu, Hawaii, USA, April 2007.

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

[7] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 437–440, Hong Kong, China, 2003.

[8] Frank Kurth and Meinard Müller. Efficient Index-Based Audio Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, February 2008.

[9] Frank Kurth, Andreas Ribbrock, and Michael Clausen. Identification of highly distorted audio material for querying large scale data bases. In *Proceedings of the 112th AES Convention*, 2002.

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

[11] Meinard Müller, Verena Konz, Andi Scharfstein, Sebastian Ewert, and Michael Clausen. Towards automated extraction of tempo parameters from expressive music recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 69–74, Kobe, Japan, October 2009.

[12] Meinard Müller and Frank Kurth. Towards structural analysis of audio recordings in the presence of musical variations. *EURASIP Journal on Advances in Signal Processing*, 2007(1):163–163, 2007.

[13] Meinard Müller, Frank Kurth, and Michael Clausen. Audio matching via chroma-based statistical features. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 288–295, 2005.

[14] Jouni Paulus and Anssi 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.

[15] Lawrence Rabiner and Bing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.

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

[17] Avery Wang. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Baltimore, USA, 2003.

[18] Gerhard Widmer, Simon Dixon, Werner Goebl, Elias Pampalk, and Asmir Tobudic. In search of the Horowitz factor. *AI Magazine*, 24(3):111–130, 2003.