Automatisierte Identifikation von Audioaufnahmen anhand Symbolisch Codierter Musikalischer Themen

Stefan Balke, Lukas Lamprecht, Vlora Arifi-Müller, Thomas Prätzlich, Meinard Müller
International Audio Laboratories Erlangen, D-91058 Erlangen, E-Mail: {stefan.balke, meinard.mueller}@audiolabs-erlangen.de

Zusammenfassung

Im Jahre 1948 veröffentlichten Barlow und Morgenstern eine Sammlung von ca. 10 000 musikalischen Themen, die einen Überblick der wichtigsten Instrumentalwerke der klassischen Musik geben [1]. Diese meist viertaktigen, einstimmig gesetzten Melodielinien hinterlassen beim Hörer häufig einen bleibenden Eindruck und eignen sich daher als kompakte Beschreibungen der entsprechenden Werke. In diesem Beitrag untersuchen wir, inwieweit sich die symbolisch codierten Themen verwenden lassen, um entsprechende Audioaufnahmen in einer Musikdatenbank zu identifizieren. Hierzu passen wir gängige Verfahren der automatisierten Musiksuche systematisch an, um unterschiedlichen Herausforderungen im Kontext klassischer Musik gerecht zu werden. Bei der Identifikation der Themen in Audioaufnahmen muss man zum einen mit globalen und lokalen Tempounterschieden und zum anderen mit Transpositionen und Abweichungen in der Stimmung umgehen können. Weiterhin können die einstimmigen Themen in den Audioaufnahmen in einem mehrstimmigen Kontext vorliegen. In unseren Experimenten zeigen wir, wie man die Suchergebnisse systematisch verbessern kann. Als eine weitere interessante Anwendung der inhaltsbasierten Musiksuche diskutieren wir, wie sich diese Techniken einsetzen lassen, um vertiefende Einblicke in die musikalischen und akustischen Eigenschaften der zugrundeliegenden Musikaufnahmen zu erhalten.

1 Aufgabenstellung

Durch zunehmende Digitalisierung von Musikdaten aller Art sind in den letzten Jahren umfangreiche, oft unstrukturierte Musikdatenbeständen enstanden [9]. 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 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 [3, 8, 10, 11, 12, 14].

In diesem Beitrag betrachten wir ein solches Suchszenario, bei dem ein Audiodatenbestand anhand von musikalischen Themen durchsucht werden soll. Ausgangspunkt unsere Studie ist das Buch „A Dictionary of Musical Themes“ von Barlow und Morgenstern [1], das den Notentext von ca. 10 000 musikalischen Themen für wichtige Instrumentalwerke der westlichen klassischen Musik enthält. Die dort gesammelten Themen sind zudem im Internet in digitaler Form (als MIDI-Dateien) verfügbar [13]. Die Themen sind einstimmig notiert und bestehen meist aus vier Takten. Als Beispiel zeigt Abbildung 1 ein Thema aus Beethovens 5. Sinfonie in c-Moll Op. 67, auch bekannt als das „Schicksalsmotiv“.

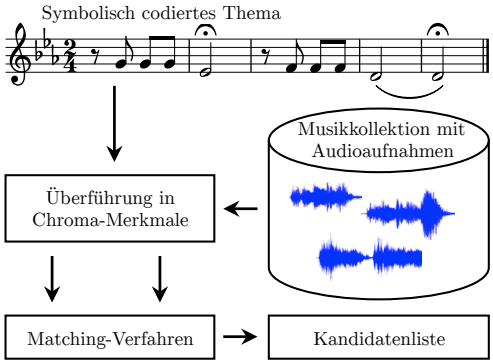

Abbildung 1: Überblick des Identifikationsverfahrens. Das symbolisch codierte Thema und die Audioaufnahmen der Musikkollektion werden jeweils in Chroma-Darstellungen überführt. Mittels eines DTW-basiertem Matching-Verfahrens werden die Themen in der Musikkollektion lokalisiert und als Kandidatenliste ausgegeben.

Zudem nehmen wir an, dass eine Musiksammlung von Gesamtaufnahmen der Musikstücke vorliegt. Dieser Audiodatenbestand kann im Allgemeinen sehr umfangreich und unstrukturiert sein. In unserem Szenario soll der Audiodatenbestand rein inhaltsbasiert (d. h. ohne die Verwendung von text-basierten Metadaten) durchsucht werden. Hierbei dienen die symbolisch codierten Themen als Suchanfrage. Das Ziel besteht darin, die zu dem angefragten Thema zugehörigen Musikaufnahmen zu identifizieren. Bei der Bearbeitung dieser Suchaufgabe stellen sich die folgenden Herausforderungen:

- **Crossmodalität.** Während es sich bei den Anfragen um symbolisch codierte Notentextdaten (MIDI) handelt, besteht der Datenbestand aus akustischen Musikaufnahmen.
- **Stimmung.** Die Stimmung von Instrumenten, Ensembles und Orchestern kann von Aufnahme zu Aufnahme variieren.
- **Transposition.** Die Aufnahme kann von der im Notentext festgelegten Originaltonart abweichen, z. B. durch Transposition in eine andere Stimmlage.
- **Tempounterschiede.** Die Interpretation eines Notentexts eröffnet große musikalische Freiheiten, die zu lokalen Schwankungen und globalen Abweichun-

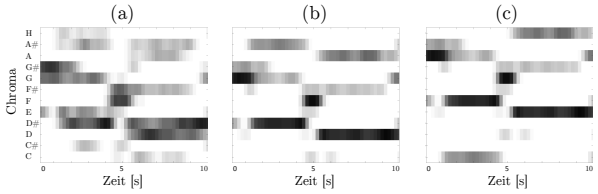

Abbildung 2: **(a)** Chroma-Darstellung einer Audioaufnahme mit Abweichung von der angenommenen Stimmung. **(b)** Chroma-Darstellung nach Kompensation der Stimmungsabweichung. **(c)** Transponierte Chroma-Darstellung.

gen im Tempo führen können. Weiterhin ist das Tempo des symbolisch codierten Themas nicht gegeben und muss geschätzt werden.

- **Polyphonie.** Das einstimmig notierte Thema wird in der Aufnahme häufig durch zusätzliche Begleitstimmen und andere Melodiestimmen überlagert.

Hinzu kommen noch Variabilitäten bezüglich der Instrumentierung, der Klangfarbe oder der Dynamik. In Abschnitt 2 beschreiben wir zunächst ein in der Literatur bekanntes Matching-Verfahren (siehe auch Abbildung 1). In unseren Experimenten (Abschnitt 3) zeigen wir, wie sich die Suchergebnisse durch geeignete Anpassungen und Erweiterung des Verfahrens systematisch verbessern lassen.

2 Matching-Verfahren

Unser Suchverfahren basiert auf einem in [8] beschriebenen Verfahren, bei dem zunächst die Anfrage und die Audioaufnahme in eine gemeinsame Merkmalsdarstellung transformiert und dann mit einem Matching-Verfahren verglichen werden. Ähnliche Verfahren wurden in der Literatur zum Abgleich von polyphonen Notentextdaten und Audioaufnahmen beschrieben [4, 5, 15].

2.1 Chroma-Darstellung

Im Musikkontext werden insbesondere Chroma-Merkmale mit großem Erfolg für unterschiedliche Such- und Analyseaufgaben eingesetzt [2, 6, 8]. 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 einen crossmodalen Vergleich (siehe Abbildung 3).

Bei der Umwandlung einer Audioaufnahme in eine Chroma-Darstellung kommen Filterbanktechniken zum Einsatz, die auf der Kenntnis der Stimmung basieren (wobei der Kammerton A4 der Frequenz 440 Hz entspricht). Bei einer Abweichung von der angenommenen Stimmung kann die resultierende Chroma-Darstellung große „Verschmierungseffekte“ aufweisen (siehe Abbildung 2a). Durch eine vorgeschaltete Schätzung der Stimmung können diese Effekte kompensiert werden (siehe Abbildung 2b und [6]). Weiterhin können durch zyklische

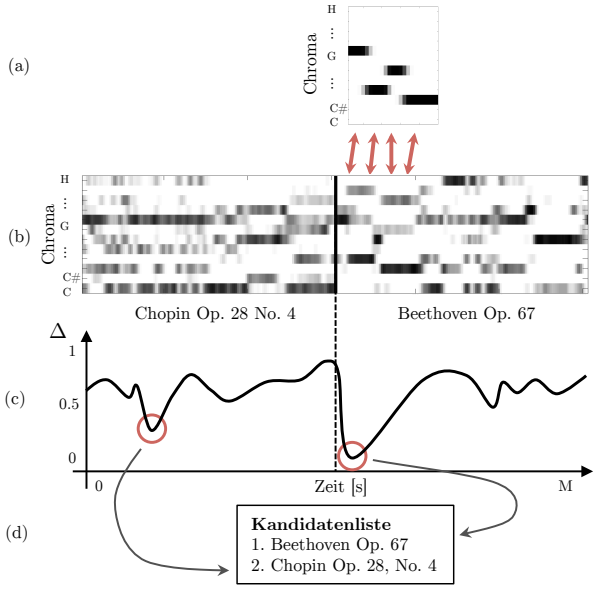

Abbildung 3: Illustration des Matching-Verfahrens. **(a)** Chroma-Darstellung der Anfrage. **(b)** Chroma-Darstellung von Aufnahmen der Musikkollektion. **(c)** Matching-Kurve Δ. **(d)** Sortierte Liste der Matching-Kandidaten.

Verschiebungen der Chroma-Merkmale Transpositionen simuliert werden (siehe Abbildung 2c und [7]). Diese Anpassungen sind in unserem Suchszenario von großer Bedeutung (siehe Abschnitt 3).

2.2 Matching-Verfahren

Wir beschreiben nun ein Verfahren zum Abgleich des angefragten Themas und der Audioaufnahmen der Musikkollektion. In einem ersten Schritt werden zunächst die Chroma-Darstellungen aller Audioaufnahmen berechnet. Sei $Y := (y_1, y_2, \ldots, y_M)$ die resultierende Folge der Chroma-Merkmale aller Aufnahmen (wobei wir der Einfachheit halber annehmen, dass die gesamte Kollektion durch eine Folge repräsentiert wird). Weiterhin wird die symbolische Anfrage in eine mit $X := (x_1, x_2, \ldots, x_N)$ bezeichnete Chroma-Darstellung überführt. Die Länge $M \in \mathbb{N}$ der Folge Y ist dabei wesentlich größer als die Länge $N \in \mathbb{N}$ der Anfragefolge X.

Die Aufgabe besteht nun darin, die Anfrage X lokal mit Teilfolgen von Y zu vergleichen. Um mögliche Tempounterschiede zu kompensieren, verwenden wir eine als „Subsequence Dynamic Time Warping“ (SDTW) bekannte Alignierungstechnik (siehe [8, Kapitel 4]). Intuitiv gesprochen werden hierbei die zu vergleichenden Folgen zeitlich so verzerrt, dass die Abfolge der jeweiligen Merkmale möglichst gut übereinstimmt. Der Grad der möglichen Verzerrung kann dabei durch eine Schrittweitenbedingung Σ festgelegt werden. In unserem Experiment verwenden wir die beiden Schrittweitenbedingungen $\Sigma_1 := \{(1,0),(0,1),(1,1)\}$ und $\Sigma_2 := \{(2,1),(1,2),(1,1)\}$ (siehe [8, Kapitel 4]).

Durch Anwendung von SDTW erhält man eine Matching-Funktion $\Delta : [1:M] \to \mathbb{R}$, die für jeden Zeitpunkt $m \in [1:M] := \{1, 2, \ldots, M\}$ angibt, wie gut die Anfragefolge X mit einer Teilfolge von Y beginnend mit

m übereinstimmt. Je kleiner der Wert $\Delta(m)$, desto größer die Übereinstimmung von X mit der Teilfolge.

Das Matching-Verfahren wird durch Abbildung 3 anhand unserer Beethoven-Anfrage („Schicksalsmotiv“) illustriert. Hierbei besteht die Merkmalsfolge der Musikkollektion exemplarisch aus Aufnahmen zweier Musikstücke (Chopin, Beethoven). Die resultierende Matching-Funktion Δ ist in Abbildung 3c zu sehen. Die lokalen Minima dieser Funktion deuten auf mögliche Trefferkandidaten hin. Zur Bestimmung einer Kandidatenliste werden die Zeitpositionen der lokalen Mininma mit aufsteigendem Δ-Wert bestimmt.

3 Experimente

Die nun folgenden Experimente sind eine Erweiterung der Untersuchungen in [4]. Unser Korpus setzt sich aus einstimmigen Themen und mehrstimmigen Audioaufnahmen zusammen. Die Basis bildet die Sammlung von Barlow und Morgenstern [1]. Die dort gesammelten Themen sind im Internet als MIDI-Dateien verfügbar [13] und dienen in dieser Form im weiteren Verlauf als Anfragen. Die Audioaufnahmen der Musikkollektion wurden den Themen manuell zugeordnet und stammen von kommerziellen Tonträgern. Für alle Experimente gilt, dass wir nicht an der exakten Position der Anfrage in einer Musikaufnahem interessiert sind, sondern lediglich die zu einer Anfrage zugehörige Musikaufnahme identifizieren wollen. Die Experimente sind zweistufig aufgebaut: Zum Justieren der Parameter nutzen wir einen Datensatz ($\mathcal{D}_1$), der aus 177 Melodieanfragen und einer Musikkollektion von 100 Audioaufnahmen besteht. In einem zweiten Schritt werden die ermittelten Parameter auf einem größeren Datensatz ($\mathcal{D}_2$) angewendet, um die Skalierbarkeit des Systems zu testen. Dieser Datensatz umfasst 2046 Melodieanfragen und eine Musikkollektion von 1113 Audioaufnahmen. $\mathcal{D}_1$ ist dabei eine Teilmenge von $\mathcal{D}_2$.

Das Ergebnis jeder Suchanfrage von Musikaufnahmen ist eine aufsteigend nach Δ-Kosten sortierte Liste. Die tatsächliche Position der zu identifizierenden Audioaufnahme in dieser Liste bezeichen wir als Rang. Als Evaluationsmaß verwenden wir den mittleren Rang und dessen Standardabweichung. Die Intention der folgenden, iterativen Parameteroptimierung ist es, möglichst viele musikalische Eigenschaften abzubilden, um die Identifikation zu verbessern (d. h. den Rang zu reduzieren). Ein Rang von 1 ist der Optimalfall und bedeutet, dass die gesuchte Musikaufnahme zuoberst in der Liste steht. Die Extraktion der Frequenz-Merkmale aus der Anfrage und den Audioaufnahmen erfolgt mit einer zeitlichen Auflösung von 10 Hz. Die Weiterverarbeitung zu Chroma-Merkmalen erfolgt durch eine Fensterung (Hanning-Fenster) über neun aufeinanderfolge Merkmals-Vektoren und einer anschließenden Halbierung der Abtastrate auf 5 Hz. Die Verwendung dieser Merkmale wird in den Experimenten als *Chroma* gekennzeichnet.

Tabelle 1 fasst die Ergebnisse der Parameteroptimierung auf dem Datensatz $\mathcal{D}_1$ zusammen. Mit einem mittleren Rang von 19,58 und einer Standardabweichung von 29,11

Tabelle 1: Ergebnisse für den Datensatz $\mathcal{D}_1$ (177 Melodieanfragen und 100 Audioaufnahmen). Aufgetragen sind der mittlere Rang und die Standardabweichung für unterschiedliche Systemparameter: Σ = SDTW-Schrittweite, S = Schätzung der Stimmung, T = Berücksichtigung von Transpositionen, 10 s = Länge der Anfrage ist auf 10 s fixiert.

	Chroma	Chroma+S	Chroma+S+T
Σ_1	19,58±29,11	16,45±26,44	18,01±23,70
Σ_2	17,25±28,45	14,30±25,61	13,16±17,38
Σ_2+10 s	17,30±28,85	14,28±26,34	12,53±18,81

Rängen dient das Experiment *Chroma + Σ_1* als Ausgangpunkt für weitere Optimierungen.

In einem zweiten Experiment *Chroma + S + Σ_1* aktivieren wir die Schätzung der Stimmung der Audioaufnahme, um die in in Abbildung 2 angedeuteten Effekte zu reduzieren. Dies verbessert den mittleren Rang auf 16,45 (SD = 26,44).

Die Verwendung der Schrittweite Σ_1 erlaubt dem Matching-Verfahren höchste Flexibilität in der Alignierung von Anfrage und Audioaufnahme. Das SDTW-Verfahren kann die Anfrage dabei so stark deformieren, dass sie zu vielen Stellen in der Musikkollektion passt und damit die Anzahl der „falschen“ Treffer (*false-positives*) ansteigt. Um diese zeitliche Flexibilität zugunsten der Erhaltung der Melodieanfrage einzugrenzen, verwenden wir im folgenden die Schrittweite Σ_2. Die Eingrenzung besteht darin, dass das Tempo der betrachteten Teilsequenz der Musikkollektion zu jedem betrachteten Zeitpunkt nur noch minimal halb oder maximal doppelt so groß wie das Tempo der Melodieanfrage sein darf. Dieser typische „Trade-off“ zwischen zeitlicher Flexibilität und Robustheit des Matching-Verfahren, bewirkt eine Verbesserung des mittleren Rangs auf 14,30 (SD = 25,61).

Das nächste Experiment erweitert das System um die Berücksichtigung von musikalischen Transpositionen. Musiktheoretisch sind Transpositionen von ± 6 Halbtönen möglich. Im folgenden betrachten wir Transpositionen von ± 2 Halbtönen, da sich experimentell herausgestellt hat, dass diese für unser Szenario ausreichend sind. Dazu werden die Chroma-Merkmale der Melodieanfrage zyklisch verschoben und erneut mit der Datenbank abgeglichen. Für jeden Transpositionsindex erhält man nun eine Matching-Funktion, wie in Abbildung 4a-c. Die Bildung einer transpositions-invarianten Matching-Funktion ist in Abbildung 4d dargestellt. Zu jedem Zeitpunkt wird diejenige Matching-Kurve verwendet, die die niedrigsten Kosten aufweist. Die Reduktion des mittleren Rangs auf 13,16 (SD = 17,38) deutet darauf hin, dass innerhalb der Musikkollektion Aufnahmen existieren, die von der notierten Tonart abweichen.

Die Tempi der verwendeten Melodieanfragen weichen oftmals stark von den tatsächlichen in den Audioaufnahmen auftretenden Tempi ab. Typischerweise ist die Länge einer Anfrage in den Audioaufnahmen fünf bis fünfzehn Sekunden lang. Wir verwenden daher eine konstante Länge von 10 s für alle Anfragen. Hierdurch erreichen wir eine leichte Verbesserung des mittleren Rangs auf 12,53 (SD = 18,81).

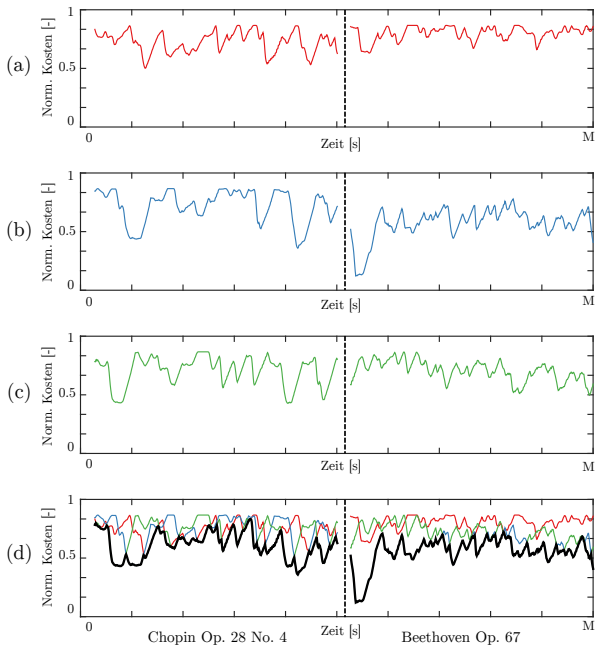

Abbildung 4: (a)-(c) Matching-Kurven für verschiedene Transpositionen. **(d)** Fenster-weise Minimierung der Matching-Funktionen resultiert in einer transpositionsinvarianten Matching-Funktion Δ.

Um die Skalierbarkeit zu untersuchen, wurde das Verfahren mit den Einstellungen *Chroma* + S + T + Σ_2 auf dem Datensatz $\mathcal{D}_2$ angewendet. Dies resultiert in einem mittleren Rang von 274,30 (SD = 308,42). Bei fester Anfragelänge von 10 s ergibt sich ein mittlerer Rang von 265,37 (SD = 308,08). Bei Verzehnfachung der Musikkollektion um den Faktor 20 wird der mittlere Rang um den Faktor 20 schlechter und weist eine hohe Standardabeweichung in der Verteilung der Ränge auf. Eine Betrachtung dieser Verteilung zeigt, dass mehr als die Hälfte der Anfragen innerhalb der ersten 200 Ränge liegen. Wir können mithilfe des aktuellen Systems große Musikkollektionen durchsuchen, die Ergebnisse bieten allerdings noch erheblichen Freiraum für weitere Optimierungen, die wir im folgenden kurz erläutern.

4 Ausblick

Das hier vorgestellte System wird uns als Ausgangspunkt für weitere Untersuchungen dienen. Insbesondere werden wir versuchen, die erstellten Datenbestände mit weiteren Annotationen anzureichern. So erhoffen wir uns z. B. durch Kenntnis der Tonart der Audioaufnahme und der Länge der enthaltenen Melodieanfrage, das Szenario kontrollierbarer zu gestalten, um auftretene Effekte gezielt und systematisch zu untersuchen. Den Grad der Polyphonie auf Seiten der Audioaufnahmen wollen wir mithilfe von Verfahren der Quellentrennung oder Verstärkung der dominanten Melodie reduzieren.

Danksagung
Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft unterstützt (DFG MU 2682/5-1). Die International Audio Laboratories Erlangen sind ein Zusammenschluss der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) und dem Fraunhofer-Institut für Integrierte Schaltungen IIS.

Literatur

[1] Barlow, H. & Morgenstern, S.: A Dictionary of Musical Themes. Überarbeitete Edtion, 3. Aufl. Crown Publishers, Inc. (1975).

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

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

[4] Ewert, S., Müller, M. & Clausen, M.: Musicmatching bei Variabilitäten in der Harmonik und Polyphonie. In: Proceedings of the 36th Deutsche Jahrestagung für Akustik (DAGA). Berlin, Germany (2010).

[5] Fremerey, C., Clausen, M., Ewert, S. & Müller, M.: Sheet Music-Audio Identification. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR), 645–650. Kobe, Japan (2009).

[6] Gómez, E. (2006): Tonal Description of Music Audio Signals. Dissertation, UPF Barcelona.

[7] Goto, M.: A Chorus-Section Detecting Method for Musical Audio Signals. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 437–440. Hong Kong, China (2003).

[8] Müller, M.: Information Retrieval for Music and Motion. Springer Verlag (2007).

[9] Müller, M.: Neue Entwicklungen im Bereich des Music Information Retrieval. In: Proceedings of the ITG-Fachtagung Sprachkommunikation. Bochum, Germany (2010).

[10] Müller, M., Goto, M. & Schedl, M. [Hrsg.] (2012): Multimodal Music Processing, *Dagstuhl Follow-Ups*, Bd. 3. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany.

[11] Orio, N.: Music Retrieval: A Tutorial and Review. In: Foundation and Trends in Information Retrieval (2006), **1**, 1: 1–90.

[12] Pickens, J., Bello, J. P., Monti, G., Crawford, T., Dovey, M., Sandler, M. & Byrd, D.: Polyphonic Score Retrieval Using Polyphonic Audio. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR). Paris, France (2002).

[13] Schwartz, J. T. & Schwartz, D. (2008): The Electronic Dictionary of Musical Themes. Website `http://www.multimedialibrary.com/barlow/`, zuletzt besucht am 12.01.2015.

[14] Serrà, J., Gómez, E. & Herrera, P. (2010): Audio cover song identification and similarity: background, approaches, evaluation and beyond. In: Ras, Z. W. & Wieczorkowska, A. A. [Hrsg.]: Advances in Music Information Retrieval, *Studies in Computational Intelligence*, Bd. 274, Kap. 14, 307–332. Springer, Berlin, Germany.

[15] Suyoto, I. S., Uitdenbogerd, A. L. & Scholer, F.: Searching Musical Audio Using Symbolic Queries. In: IEEE Transactions on Audio, Speech & Language Processing (2008), **16**, 2: 372–381.