Audio Matching für ähnlichkeitsbasierte Musiksuche

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

Einleitung

Inhaltsbasierte Analyse und Retrieval von Musikdokumenten stellen seit einigen Jahren einen sehr aktiven Forschungsbereich dar. In diesem Kontext 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 Ausschnitte 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. In diesem Beitrag stellen wir das Problem des sogenannten *Audio Matchings* vor. Ausgangspunkt ist hier eine große Musikdatenbank, die typischer Weise mehrere verschiedene Aufnahmen desselben Musikstücks enthält, wobei diese Aufnahmen im allgemeinen von unterschiedlichen Interpreten und in eventuell verschiedenen Besetzungen eingespielt wurden. Ist nun ein angefragter kurzer Audioausschnitts gegeben, so sollen automatisch alle entsprechenden Ausschnitte in allen in der Datenbank enthaltenen Interpretationen des zugrundeliegenden Musikstücks gefunden werden. Zur Lösung dieses Problems führen wir eine neue Klasse statistischer Chromamerkmale ein.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. Wir beschreiben ein Matchingverfahren, das selbst gegenüber signifikanten globalen Tempovariationen robust ist und beschreiben unsere Experimente auf einem Datenbestand von über 110 Stunden klassischer Musik.

CENS-Merkmalsfolgen

Im ersten Schritt der Merkmalsextraktion wird das zu transformierende Audiosignal unter Verwendung einer Multiraten-Filterbank aus elliptischen IIR-Filtern in 88 Tonhöhenbänder (gemäß der temperierten Stimmung) 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 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 Energieverteilunge der im Audiosignal vorkommenden Frequenzen auf die 12 Chromabänder widerspiegelt, siehe z.B. Abb. 1.

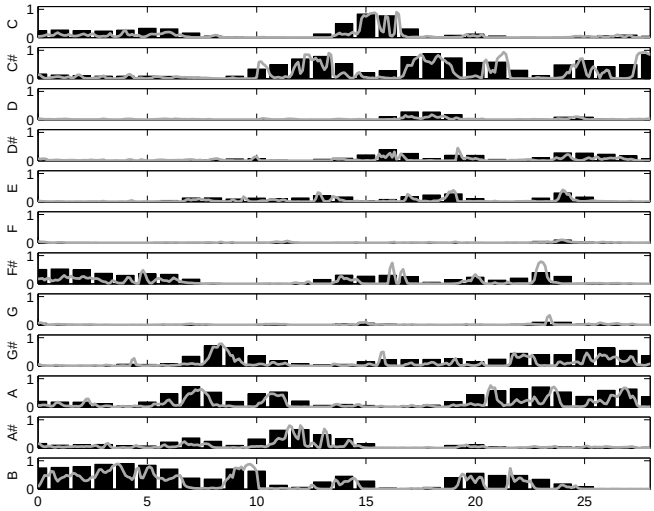

Abbildung 1: Die Abbildung zeigt die Takte 44 − 55 (Sekunden 112 − 140) einer Zukerman-Interpretation von Vivaldis Frühling RV 269, No. 1. Die hellen Kurven stellen die lokalen Chroma-Energieverteilungen (10 Hz) dar, während die dunklen Balken die CENS-Merkmalsfolge (1 Hz) zeigen.

Die so erhaltenen Chromamerkmale sind durch die Chroma-Identifikation robust unter Klangfarbenänderung und durch die Normalisierung 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 mit den Faktor 10 auf 1 Hz reduziert, siehe Abb. 1. Die so resultierenden Merkmale werden abkürzend mit *CENS* (**C**hroma **E**nergy distribution **N**ormalized **S**tatistics) bezeichnet, vgl. [1] für Details.

Audio Matching

Auf der Grundlage der CENS-Merkmalen gehen wir nun auf die wichtigsten Ideen eines robusten Verfahrens zum Audio-Matching ein und diskutieren typische Ergebisse anhand eines Vivaldi-Beispiels (vgl. `http://www-mmdb.iai.uni-bonn.de/projects/audiomatching` für weitere Beispiele). Unsere Testdatenbank besteht aus 1167 Stücken (112 Stunden) klassischer Musik diverser bekannter Komponisten. Dabei liegen für die meisten Stücke mehrere Interpretation vor. In einem Vorverar-

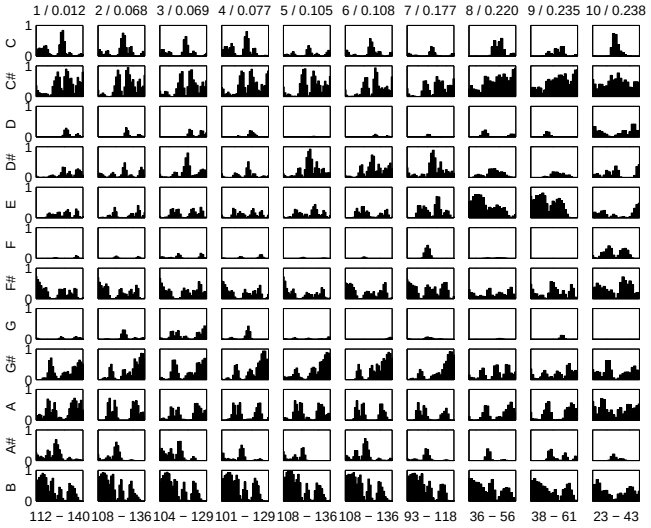

Abbildung 2: CENS Merkmalsfolge für die ersten 10 Treffer der Vivaldi-Anfrage aus Abb. 1.

beitungsschritt berechnen wir die CENS-Merkmalsfolgen aller Aufnahmen der Datenbank. Durch Konkatenation dieser Folgen (und unter Protokollierung der Dateigrenzen) können wir die gesamte Datenbank durch eine einzige Folge $\mathcal{D} := (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^N)$ von CENS-Merkmalen repräsentieren.

In unserem Szenario besteht eine typische Anfrage aus einem $10-30$-sekündigen Musikauschnitt. 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 definieren wir die Abstandsfunktion $\Delta : [1 : N - M+1] \rightarrow [0,1]$ durch $\Delta(i) := 1 - \frac{1}{M} \sum_{m=1}^{M} \langle \vec{v}^{i+m-1}, \vec{w}^m \rangle$. 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 der 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 D der Länge M. Unterschiedliche Interpretationen weisen häufig verschiedliche globale Tempi auf. Um diesen gerecht zu werden, erzeugen wir unterschiedliche Versionen Anfrage, 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 Anfrage wird dann eine separate Abstandsfunktion Δ^j, $j \in [1:8]$, berechnet. Seien nun $i_{\min} \in [1 : N - M + 1]$ und $j_{\min} \in [1 : 8]$ diejenige 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 schließen wir zur Vermeidung von Überschneidungen mit dem besten Treffer die in einer Umgebung von $i_{\min}$ liegenden Indizes für die weiteren Betrachtungen aus und fahren in analoger Weise fort, bis eine gewissse Anzahl an Treffern erreicht oder eine vorab festgelegte Abstandsschranke überschritten wird.

Unsere Datenbank enthält sieben verschiedenen Interpretationen des Vivaldi-Beispiels aus Abb. 1 (Abbado, Carmirelli, Lizzio, Mae, Nishizaki, Perlman, Zukerman). Für die besten sieben Treffers unseres Matching-Verfahrens erhält man genau die Ausschnitte in den sieben Interpretationen, die den Takten $44-55$ der Anfrage entsprechen. Abb. 2 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. 1) auf. Die Position des Ausschnitts 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 Ausschnitt 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 Ausschnitt 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.

Fazit und Ausblick

Die Grundidee das Audio-Matchings basiert darauf, die erwünschten Invarianzen in die Merkmale selbst zu integrieren, um auf diese Weise robuste und effiziente Matching-Verfahren 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. Für die Zukunft planen wir weiterhin, die Effizienz des Audio-Matchings durch geeignete Indexierung der CENS-Merkmale zu steigern.

Literatur

[1] Meinard Müller, Frank Kurth, and Michael Clausen, Audio Matching via Chroma-Based Statistical Features. Proc. of the *6th ISMIR*, London, GB, 2005.