Automatisierte Synchronisation von Audio- und Partiturdaten

Meinard Müller, Frank Kurth, Tido Röder, Michael Clausen

Institut für Informatik, Uni Bonn, Römerstr. 164, D-53117 Bonn, Email: {meinard,frank,roedert,clausen}@iai.uni-bonn.de

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, notenbasierte Partiturdaten oder MIDI-Daten. Die Inhomogenität und Komplexität solcher Daten führen zu großen, weitgehend noch ungelösten Problemen bei der automatisierten Datenerschließung, etwa für Anwendungen wie die inhaltsbasierte Suche und Navigation in digitalen Musikbibliotheken. In diesem Kontext spielt die sogenannte *Musiksynchronisation* eine wichtige Rolle, bei der es um die automatische Verlinkung von Daten unterschiedlicher Formate geht, siehe Abb. 1.

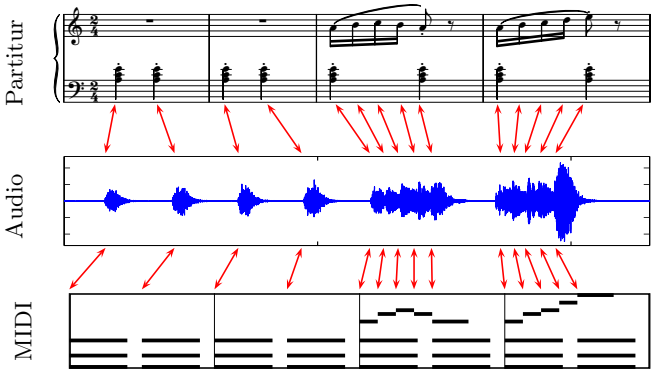

Abbildung 1: Verlinkung von Musikdaten in unterschiedlichen Formaten (Partitur, Audio, MIDI), die dasselbe Musikstück (die ersten vier Takte der Etüde Nr. 2, op. 100, F. Burgmüller) repräsentieren.

In diesem Beitrag studieren wir das Szenario, 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 Position im Audiodatenstrom die entsprechende Stelle 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 rein symbolischen Partiturdaten grundlegend von den wellenformbasierten Audiodaten unterscheiden, stellt sich die Audio-Partitur-Synchronisation als ein schwieriges Problem dar. Auf der einen Seite besteht die Partitur aus notenbasierten Parametern wie Tonhöhen, Einsatzzeiten und Tonlängen, welche großen interpretatorischen Spielraum hinsichtlich des Tempos, der Dynamik oder der Ausführung von Notengruppen wie Trillern zulassen. Auf der anderen Seite kodiert eine CD-Aufnahme alle Parameter, die zur Rekonstruktion der akustischen Realisation (Wellenform) benötigt werden – die zugrundeliegenden Notenparameter sind allerdings nicht explizit gegeben. Daher gehen die meisten bisherigen 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 [1, 2, 3]. Der Arbeit [2] folgend gehen wir nun auf einige Grundideen genauer ein.

Extraktion

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 verwenden wir in unserem Verfahren eine kleine Anzahl von semantisch ausdrucksstarken Merkmalen, 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. Hierzu wird unter Verwendung fortgeschrittener Filtertechniken (Filterbank bestehend aus 88 elliptischen IIR-Filtern) das Audiosignal in 88 Bänder (gemäß den Klaviertönen) zerlegt. Mittels energiebasierter Verfahren werden dann für jedes Band Kandidaten für Einsatzzeiten berechnet, siehe auch Abb. 2.

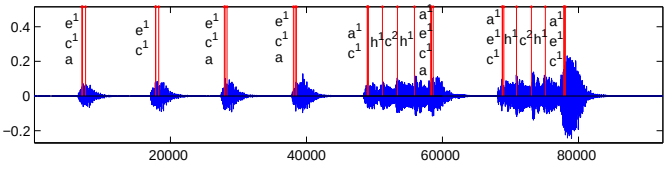

Abbildung 2: Extraktion von notenähnlichen Parametern aus dem Audiodatenstrom.

Im Fall polyphoner Musik stellt die Extraktion von Notenparametern ein extrem schwieriges Problem dar. Selbst für die Klasse polyphoner Klaviermusik, auf die wir uns im folgenden beschränken, 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 in Hinblick auf eine *Musiktranskription* unzureichend sein mögen, ermöglichen sie dennoch im allgemeinen eine ausgezeichnete *Musiksynchronisation*.

Synchronisation

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 verwenden wir ein Verlinkungsmodell, welches sich von klassischen auf „dynamic time warping“ (DTW) basierenden Methoden, siehe z. B. [3], unterscheidet. Um eventuellen Unstimmigkeiten zwischen dem Partitur- und Audiodatenstrom, bedingt z. B. durch interpretatorische Abweichungen oder fehlerhafte Extraktion, Rechnung zu tragen, erzwingen wir nicht die Zuordnung aller Partitur- bzw. Extraktionsparameter, sondern erlauben auf beiden Seiten auch unverlinkte Ereignisse – ganz nach dem Motto: „Besser keine Zuordnung als eine schlechte Zuordnung.“ Darüber hinaus lassen wir uns bei der Definition des lokalen Ähnlichkeitsmaßes von folgendem einfachen aber weitreichenden Prinzip leiten: Die Partitur gibt uns vor, wonach im Audiodatenstrom zu suchen ist: Bei der Verlinkung werden also nur Extraktionsparameter berücksichtigt, die sich in der Partitur widerspiegeln. Für die technischen Details verweisen wir auf [2].

Experimente

Unser Verfahren wurde in MATLAB implementiert und anhand zahlreicher Beispiele polyphoner Klaviermusik unterschiedlicher Komplexität getestet, einschließlich der Etüden op. 100 von Burgmüller, der Etüden op. 10 von Chopin und einiger Klaviersonaten von Beethoven. Zur Demonstration und Bewertung der Synchronisationsergebnisse wurden diese *sonifiziert*. Hierzu sei erinnert, daß ein Audio-Partitur-Synchronisationsergebnis einer Zuordnung der musikalischen Einsatzzeiten der Partiturnoten mit den physikalischen Einsatzzeiten der entsprechenden Ereignisse im Audiodatenstrom entspricht. Für jede verlinkte Partiturnote wurde nun ein kurzer Sinuston der vorgegebenen Tonhöhe generiert, wobei die physikalische Einsatzzeit entsprechend der obigen Zuordnung gewählt wurde. Schließlich wurde ein Stereo-Datenstrom erzeugt, welcher im linken Kanal eine Mono-Version des Audiodatenstroms und im rechten Kanal den Sinus-generierten Datenstrom enthält. Die so erzeugte Sonifikation[1] der Resultate zeigt, daß unser Verfahren für die eingeschränkte Musikklasse polyphoner Klaviermusik gute Synchronisationsergebnisse hoher Auflösung erzielt, die für Anwedungen 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.

[1] Die Sonifikationsergebnisse sind auch unter der Adresse `www-mmdb.iai.uni-bonn.de/download/sync/` verfügbar.

Ankerkonfigurationen

Der Synchronisationsalgorithmus kann beträchtlich beschleunigt werden, falls im Vorfeld der eigentlichen DP-Berechnung schon eine kleine Anzahl an geeigneten Zuordnungen musikalischer und physikalischer Einsatzzeiten bekannt ist. So benötigt z. B. die DP-Berechnung für eine 173 Sekunden dauernde CD-Aufnahme der Etüde Nr. 3, op. 10 von Chopin 423 Sekunden unter Verwendnung von 8.9 MB Speicher. Die Vorkenntnis einer Zuordnung bei Sekunde 99 beschleunigt die DP-Berechnung auf 222 Sekunden und benötigt nur noch 3.2 MB Speicher. Theoretisch hat die Vorkenntnis einer einzigen Zuordnung in der Mitte des Stücks eine Halbierung der DP-Laufzeit und eine Viertelung des Speicherplatzbedarfs zur Folge. Basierend auf dieser Beobachtung führen wir das Konzept der *Ankerkonfigurationen* ein, welche man sich typischerweise als Notenobjekte mit besonders auffallenden dynamischen oder spektralen Eigenschaften vorstellen kann, wie z. B. ein isoliert gespielter Fortissimo-Akkord oder eine lange Pause. Die Entsprechungen solcher Notenobjekte im Audiodatenstrom können nun in einem Vorverarbeitungsschritt effizient, d. h. linear in Zeit und Speicher, identifiziert werden. Die restlichen Zuordnungen können dann mittels mehrerer kurzer DP-Berechnungen zwischen den *Ankerzuordnungen* bestimmt werden.

Ausblick

Die automatisierte Musikdatenerschließung stellt ein aktuelles Forschungsgebiet mit noch vielen ungelösten und interessanten Problemstellungen dar. Die Schwierigkeit liegt insbesondere in der Komplexität und Mannigfaltigkeit von Musikdaten begründet – nicht nur hinsichtlich unterschiedlichster Datenformate, sondern auch hinsichtlich der Gattung (z. B. Pop, Klassik, Jazz), der Instrumentation (z. B. Orchester, Klavier, Schlagzeug, Stimme) und vielen weiteren Parametern (z. B. Dynamik, Tempo, Klangfarbe). Für die Zukunft planen wir unter anderem, das Problem der Musiksynchronisation für allgemeinere Musikklassen effektiv und effizient zu lösen. Hierbei soll ein System entstehen, welches verschiedenartige, konkurrierende Strategien vereinigt, anstatt sich auf eine Strategie festzulegen.

Literatur

[1] Vlora Arifi, Michael Clausen, Frank Kurth, and Meinard Müller. Synchronization of Music Data in Score-, MIDI- and PCM-Format. In Walter B. Hewlett and Eleanor Selfridge-Fields, editors, *Computing in Musicology*. MIT Press, 2004.

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

[3] Ferréol Soulez, Xavier Rodet and Diemo Schwarz, Improving Polyphonic and Poly-instrumental Music to Score Alignment. Proc. of the *4th ISMIR*, Baltimore, Maryland, 2003.