Partielle Musiksynchronisation

Meinard Müller[1], Daniel Appelt[2]

[1] *Max-Planck-Institut für Informatik, Campus E1-4, D-66123 Saarbrücken, Email: meinard@mpi-inf.mpg.de*

[2] *Institut für Informatik III, Universität Bonn, Römerstr. 164, D-53117 Bonn, Email: daniel.appelt@gmx.net*

Einleitung

Oft enthalten digitale Musikbibliotheken verschiedene Versionen oder Interpretationen eines Musikstücks. Die Entwicklung von Synchronisationstechniken zur automatischen Verlinkung solcher unterschiedlicher Varianten ist in Hinblick auf eine effiziente Navigation in inhomogenen Musikdatenbeständen von großer Bedeutung. Ziel der Synchronisation ist es, zu einer bestimmten Position innerhalb einer Interpretation die entsprechende Stelle innerhalb einer anderen Version zu bestimmen. Für einen aktuellen Überblick und weitere Literaturangaben zur Musiksynchronisation verweisen wir auf [1]. Bisherige Synchronisationsverfahren setzen voraus, dass sich die zu verlinkenden Datenströme bis auf interpretatorische Unterschiede in Dynamik, Klang und Tempoverlauf im Wesentlichen entsprechen. Eine wichtige Aufgabe besteht nun darin, eine semantisch sinnvolle Synchronisation auch dann zu gewährleisten, wenn die Versionen nur in Teilen übereinstimmen. Häufig auftretende Variationen beinhalten dabei das Auslassen von Wiederholungen, das Einfügen zusätzlicher Teile wie Soli oder Kadenzen, oder eine unterschiedliche Anzahl von Strophen oder Refrains in populärer Musik. In diesem Beitrag beschreiben wir ein neuartiges Synchronisationsverfahren, welches eine sinnvolle Verlinkung von Audioaufnahmen auch in Gegenwart struktureller Variationen erlaubt. Hierzu führen wir das Konzept pfadbeschränkter Änlichkeitsmatrizen ein, auf deren Basis eine partielle Verlinkung mittels eines effizient berechenbaren Optimierungsverfahren bestimmt werden kann. Unsere allgemeine Strategie besteht hierbei darin, möglichst lange, zusammenhängende Musikabschnitte zu verlinken und so eine Zerstückelung des Audiomaterials zu vermeiden.

Pfadbeschränkte Ähnlichkeitsmatrizen

Grundlage zur Synchronisation von Musikstücken bilden so genannte *Ähnlichkeitsmatrixen*. Hierzu werden zunächst die beiden zu synchronisierenden Audioaufnahmen in geeignete Merkmalsfolgen $V := (v^1, v^2, \ldots, v^N)$ und $W := (w^1, w^2, \ldots, w^M)$ transformiert. Zur Berechnung der $(N \times M)$-Ähnlichkeitsmatrix $\mathcal{S}$ wird dann jedes Merkmal v^n, $1 \leq n \leq N$, der ersten Folge mit jedem Merkmal w^m, $1 \leq m \leq M$ der zweiten Folge bezüglich eines geeigneten Ähnlichkeitsmaßes verglichen. Unser Verfahren basiert auf 12-dimensionalen Chromamerkmalen, die die lokale Energieverteilung (wir verwenden eine Auflösung von einem Chromavektor pro Sekunde) der Audiosignale auf die 12 Chromabänder widerspiegelt [1]. Hierbei beziehen sich die Chroma auf die zwölf Tonhöhenklassen C, C$^\sharp$, D, ..., H der wohltemperierten Stimmung. Folgen solcher Merkmale korrelieren stark mit dem Harmonieverlauf des zu Grunde liegenden Musikstücks und sind hochgradig invariant bezüglich Änderungen von Parametern wie Dynamik, Klangfarbe und Artikulation. Als Ähnlichkeitsmaß verwenden wir das Skalarprodukt der zu vergleichenden Chromavektoren. Damit ergibt sich $\mathcal{S}(n,m) := \langle v^n, w^m \rangle$. Im folgenden bezeichnen wir ein Tupel (n,m) auch als *Zelle* von $\mathcal{S}$ und den Wert $\mathcal{S}(n,m)$ als *Score* der Zelle. Abb. 1a zeigt die resultierende Ähnlichkeitsmatrix zweier (strukturell modifizierter) CD-Aufnahmen unterschiedlicher Interpretationen von Brahms' Ungarischem Tanz Nr. 5. Die erste Version (vertikale Achse) hat die musikalische Form $A_1^1 B_1^1 B_2^1 C^1 A_2^1 B_3^1 B_4^1 D^1$ und die zweite Version (horizontale Achse) die musikalische Form $A_1^2 A_2^2 B_1^2 B_2^2 A_3^2 B_3^2 D^2$. Die meisten der bisherigen Synchronisationsverfahren basieren auf Techniken des Dynamic Time Warping (DTW), wobei jedem Element der einen Folge ein Element der anderen Folge zugeordnet wird. Dieses Verfahren ist jedoch problematisch, falls Elemente der einen Folge keine geeigneten Entsprechungen in der anderen Folge haben. Dies kann im Fall struktureller Unterschiede zu fehlgeleiteten Zuordnungen führen, siehe Abb. 1a.

Beim Vorliegen struktureller Unterschiede werden also partielle Synchronisationstechniken benötigt, die unter Zulassung von Auslassungen nur die sich entsprechende Teile der Audioaufnahmen verlinken. Die grundlegende Idee beim Entwurf solcher Verfahren basiert auf der Beobachtung, dass Paare von ähnlichen Teilfolgen in den beiden Versionen in Form von Pfaden mit hohem Score in der Ähnlichkeitsmatrix sichtbar werden, die in Richtung der Hauptdiagonalen verlaufen. Als Beispiel betrachten wir den Pfad in Abb. 1f, der von der Zelle $(1, 18)$ zur Zelle $(67, 69)$ führt. Dieser Pfad offenbart die Ähnlichkeit der beiden Audiosegmente, die dem musikalischen Teil $A_1^1 B_1^1 B_2^1$ in der ersten und dem Teil $A_2^2 B_1^2 B_2^2$ in der zweiten Version entsprechen. Als Grundlage für die partielle Synchronisation verwenden wir sogenannte *pfadbeschränkte Ähnlichkeitsmatrizen*, deren Konstruktion in mehreren Schritten erfolgt. Zunächst wird die diagonal verlaufende Pfadstruktur der Ähnlichkeitsmatrix mittels geeigneter Filterungstechniken herausgearbeitet [1], siehe Abb. 1b. Aus der so verbesserten Ähnlichkeitsmatrix werden nun die Pfade mit hohem Score sukzessive mit einer Greedy-Strategie unter Verwendung geeigneter Schwellwerte extrahiert. Die extrahierte Pfadstruktur wird dann in eine pfadbeschränkte Ähnlichkeitsmatrix transformiert, wobei alle zu den extrahierten Pfaden gehörigen Zellen einen Score von Eins und die übrigen Zellen einen Score von Null erhalten. Abb. 1d zeigt eine verfeinerte

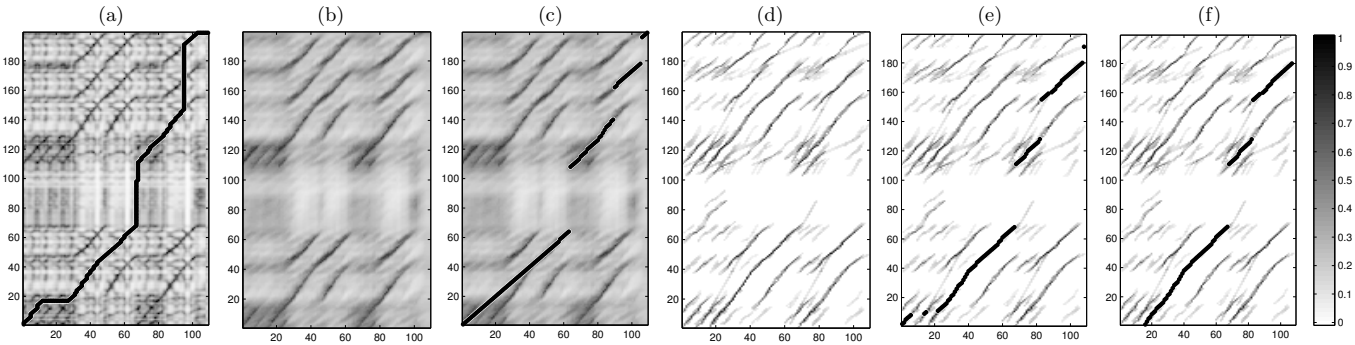

Abbildung 1: (a) Klassische Ähnlichkeitsmatrix $\mathcal{S}$ mit globalem Zuordnungspfad. **(b)** Geglättete Ähnlichkeitsmatrix. **(c)** Geglättete Ähnlichkeitsmatrix mit Score-maximimierendem Match. **(d)** Pfadbeschränkte Ähnlichkeitsmatrix $\mathcal{S}^{\mathrm{pb}}$. **(e)** $\mathcal{S}^{\mathrm{pb}}$ mit Score-maximimierendem Match. **(f)** $\mathcal{S}^{\mathrm{pb}}$ mit bereinigetem Match.

Version einer solchen pfadbeschränkten Martix, die wir im folgenden mit dem Symbol $\mathcal{S}^{\mathrm{pb}}$ bezeichnen. Bei dieser Verfeinerung wird die extrahierte Pfadstruktur stablisiert, indem Ungenauigkeiten in den Pfadlängen ausgeglichen, lückenhafte Pfade ergänzt und bedeutungslose Pfadfragmente verworfen werden. Hierbei kommen Clusteringmethoden in Verbindung mit einem Transitivitätsschritt zum Einsatz [2]. Weiterhin wird die Null-Eins Scorebewertung durch eine kontinuierliche Bewertung zwischen Null und Eins ersetzt, durch die eine Art Wahrscheinlichkeit einer Pfadzugehörigkeit einer Zelle ausgedrückt wird. Dabei ist es wichtig, dass weiterhin nur diejenigen Zellen einen positiven Score aufweisen, die zur Pfadstruktur gehören.

Partielles Matching

Unser Ziel besteht darin, möglichst lange und zusammenhängende Segmente der beiden Audioaufnahmen zu verlinken. Im Fall, dass es für ein Segment der einen Audioaufnahme kein semantisch sinnvolles Pendant in der anderen Aufnahme gibt, soll dieses Segment unverlinkt bleiben. Im folgenden sei ein *Match* eine Folge $\mu = (\mu_1, \ldots, \mu_L)$ mit $\mu_\ell = (n_\ell, m_\ell) \in [1:N] \times [1:M]$ für $\ell \in [1:L]$, so dass $1 \leq n_1 < n_2 < \ldots < n_L \leq N$ und $1 \leq m_1 < m_2 < \ldots < m_L \leq M$. Ein Match induziert eine partielle Zuordnung, wobei jedes Element der einen Folge höchstens einem (oder auch keinem) Element der anderen Folge zugeordnet wird. Der *Score* von μ bezüglich der Ähnlichkeitsmatrix $\mathcal{S}^{\mathrm{pb}}$ ist dann als $\sum_{\ell=1}^{L} \mathcal{S}^{\mathrm{pb}}(v^{n_\ell}, w^{m_\ell})$ definiert. Ein Score-maximierender Match kann nun mittels dynamischer Programmierung effizient berechnet werden und stellt unser Verlinkungsergebnis dar. Die Verwendung der pfadbeschränkten Ähnlichkeitsmatrizen $\mathcal{S}^{\mathrm{pb}}$ bei diesem Verfahren ist von entscheidender Bedeutung, da hierdurch die möglichen Zuordnungen eines Matches per se auf semantisch sinnvolle Zuordnungen eingeschränkt werden. Ein Score-maximierender Match besteht nur aus Zellen mit positivem Score, die damit zur Pfadstruktur gehören, siehe Abb. 1e. Würde stattdessen eine konventionelle Ähnlichkeitsmatrix zu Grunde gelegt werden, könnte der Match zu semantisch unsinningen Zuordnungen führen, siehe Abb. 1c.

Im allgemeinen kann ein Score-maximierender Match μ eine größere Anzahl kurzer Wegzusammenhangskomponenten aufweisen, die zu einer Zerstückelung des Audiomaterials führen. In einem letzten Schritt wird daher der optimale Match überarbeitet, indem zum einen die längeren Komponenten von μ noch weiter geeignet verlängert werden (was allerdings das Vorliegen geeigneter Pfade in $\mathcal{S}^{\mathrm{pb}}$ erfordert), und zum anderen die kurzen Komponenten von μ verworfen werden, siehe Abb. 1f. Der bereinigte Match stellt das Endresultat der partiellen Musiksynchronisation dar. Für technische Details verweisen wir auf [2].

Experimente

Im folgenden beschreiben wir eines unserer Experimente (siehe `http://www-mmdb.iai.uni-bonn.de/projects/partialSync/` für repräsentative Beispiele). Hierzu wurden Synchronisationspaare gebildet, die aus zwei unterschiedlichen Interpretationen desselben zu Grunde liegenden Musikstücks bestehen. Die Aufnahmen eines Paares wurden durch zufälliges Auslassen und Einfügen von gekennzeichneten Segmenten strukturell modifiziert. Für diese beiden Versionen wurde dann ein Match berechnet, dessen Komponenten analysiert wurden. Hierbei wird eine Komponente als *korrekt* erachtet, wenn diese sich musikalisch entsprechende Teile miteinander verlinkt. Bei einem Testlauf mit 128 verschiedenen Synchronisationspaaren (mit einer Gesamtzahl von 318 Komponenten), wurden 87% aller Komponenten als semantisch korrekt eingestuft. Weitere Experimente haben gezeigt, dass sich diese Rate verbessern lässt, wenn mehrere mittels unterschiedlicher Merkmalsauflösungen (z. B. 1 Hz und 2 Hz) berechnete Synchronisationsresultate verbunden werden. Als eine zukünftige Aufgabe soll untersucht werden, inwieweit durch den Einsatz automatisiert berechneter musikalischer Strukturen [1] weitere Verbesserungen erzielt werden können.

Literatur

[1] M. Müller, *Information Retrieval for Music and Motion*, Monograph, Springer, 2007, 318 pages.

[2] M. Müller and D. Appelt, *Path-Constrained Partial Music Synchronization*, Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, 2008.