Rhythmusbasierte Audiomerkmale und Anwendungen in der Musikerkennung

Frank Kurth, Thorsten Gehrmann, Meinard Müller

Institut für Informatik III, Universität Bonn, 53117 Bonn, Deutschland, Email: {frank,gehrmann,meinard}@iai.uni-bonn.de

Einleitung

Die Beschreibung von Audiosignalen durch Merkmale mit zeitlich lokalem Bezug ist in vielen Anwendungen im Bereich des Music Information Retrieval von großer Bedeutung. Meist kommen hier spektrale Merkmale zum Einsatz, die etwa die Harmonizität, die Tonhöhe oder die Klangfarbe des Signals an einer Zeitposition beschreiben. In unserem Beitrag stellen wir demgegenüber eine neue Klasse von Signalmerkmalen vor, die in Anlehnung an den musikalischen Rhythmusbegriff entworfen wurden. Als Anwendung zeigen wir, wie diese rhythmusbasierten Merkmale zur Erkennung zeitlich stark skalierter Audiosignale eingesetzt werden können.

Rhythmusbasierte Audiomerkmale

Die im folgenden vorgestellten Merkmale sind angelehnt an die musikalischen Begriffe *Tempo* (Toneinsätze oder *Beats* pro Zeiteinheit), *Rhythmus* (relatives zeitliches Verhältnis der Töne) und *Metrum* (Betonungsschema einer Tonabfolge).

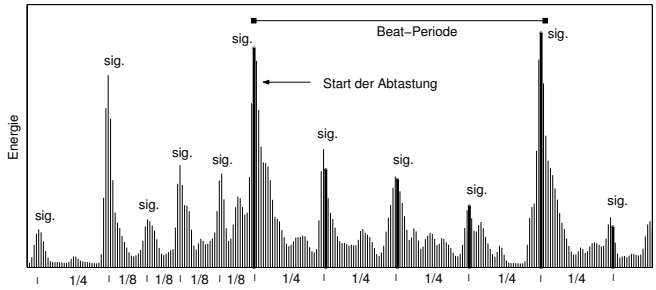

Abbildung 1: Novelty-Kurve, extrahierte signifikante Maxima, Beatperiode und Abtastpunkte der Metrummerkmale.

Zur Merkmalsextraktion wird ein Audiosignal zunächst mit einer gefensterten Fouriertransformation analysiert und durch Bildung des ℓ^1-Abstands der Beträge benachbarter Spektralvektoren die Novelty-Kurve bestimmt, siehe Abb. 1. Zur Tempoextraktion wird die Novelty-Kurve mit einer M-Band Resonator-Filterbank analysiert. Da Peaks in der Novelty-Kurve großen Änderungen des lokalen Spektrums entsprechen und auf Noteneinsätze hindeuten, wird hier angenommen, dass die Eigenfrequenz desjenigen Resonatorfilters die Tempofrequenz angibt, bei der dabei der stärkste Resonanzeffekt auftritt. Zu einem Zeitpunkt t bestimmt man zunächst ein *Beatspektrogramm* $b_t \in \mathbb{R}^M$, in dem jeder Eintrag der Summe eines Resonator-Ausgangs in einer Umgebung von 20 Sekunden um t entspricht, siehe Abb. 2 (oben). Es zeigt sich, dass bei der Schätzung des lokalen Tempos aus der Maximalstelle von b_t eine starke Verwechslungsgefahr des tatsächlichen Tempos mit dem doppelten oder halbierten Tempo besteht; ein Phänomen, das sogar bei der menschlichen Beurteilung des Tempos häufig auftritt. Aus diesem Grund schlagen wir vor, Tempofrequenzen unter Verdopplung zu identifizieren, was eine (logarithmische) Unterteilung des Beatspektrogramms in $O = M/N$ *Tempooktaven* induziert (Abb. 2 oben). Durch Aufsummierung der Energien der bis auf Verdopplung gleichen Resonanzfrequenzen von b_t über alle Tempooktaven erhält man das *zyklische Beatspektrum* (Abb. 2 unten). Dem folgenden Anwendungsbeispiel liegen die Parameter $M = 90$, $N = 30$ und $O = 3$ zu Grunde. Dieses Vorgehen ist an die Berechnung der sogenannten Chroma-Merkmale, wo Energien von Frequenzbändern zu oktavgleichen Tonhöhen aufsummiert werden, angelehnt [1]. Eine robustere Schätzung $T(t)$ des lokalen *Tempos modulo Verdopplung* erhält man nun aus dem Maximum des zyklischen Beatspektrums und reziprok hierzu die lokale *Beatperiode* $B(t) = 1/T(t)$ als Schätzung der Zeitdauer zwischen zwei aufeinanderfolgenden Beatpositionen. Da $T(t)$ laut Konstruktion eine Tempo-Äquivalenzklasse bezeichnet, wählen wir vor der Bestimmung von $B(t)$ einen geeigneten Tempo-Repräsentanten. Während die Temposchätzung äquidistant für jede Abtastposition der Novelty-Kurve durchgeführt wird, erfolgt die Bestimmung rhythmischer und metrischer Merk-

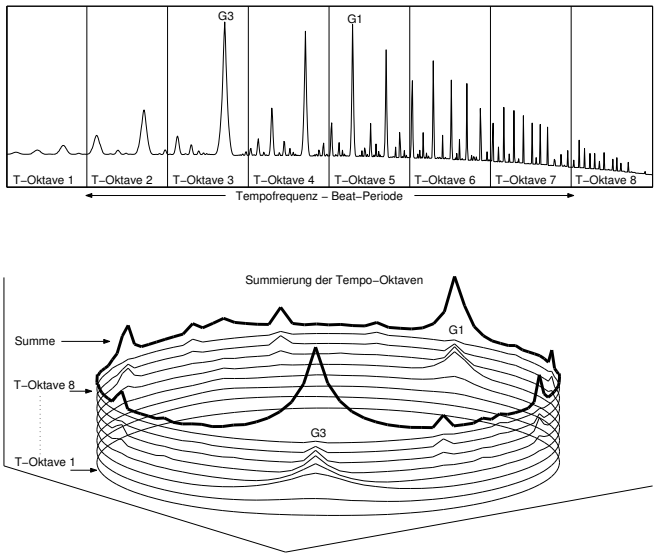

Abbildung 2: Beatspektrogramm mit 8 Tempooktaven (oben) und Berechnung des zyklischen Beatspektrums durch Summieren über Tempooktaven (unten).

male nur bezüglicher einzelner Einsatzzeiten. Zur Einsatzzeitenschätzung werden aus der Novelty-Kurve signifikante Maximalstellen extrahiert, siehe Abb. 1. Hier ist das Signifikanzkriterium so gewählt, dass die extrahierten Maxima invariant gegenüber zeitlichen Signalskalierungen sind [2]. Für jede solche Einsatzzeit t wird der lokale Rhythmus $R(t)$ als Verhältnis der Intervallänge $|t'-t|$ zur nächsten Einsatzzeit t' und $B(t)$ bestimmt. Zur Schätzung des lokalen Metrums wird die Novelty-Kurve beginnend bei t mit Schittweite $B(t)/4$ an 6 Stellen abgetastet. Durch Quotientenbildung benachbarter Abtastwerte mit anschließender Quantisierung resultiert so eine Metrumschätzung $M(t) \in \mathbb{R}^5$ und insgesamt der lokale Merkmalsvektor $(T(t), R(t), M(t))$ an der Stelle t.

Anwendung in der Audioidentifikation

Audioidentifikation bezeichnet die Aufgabe eine konkrete, eventuell verrauschte Audioaufnahme eines Musikstücks automatisch zu erkennen. Erfolgreiche Ansätze zur Lösung des Identifikationproblems sind seit einigen Jahren bekannt [3, 4]. In Rundfunkszenarien, wo Musikstücke oft in vom Original abweichenden Geschwindigkeiten wiedergegeben werden, sind solche Identifikationsansätze nicht ohne weiteres anwendbar. Erweiterte Ansätze, wie etwa HMM-basierte Methoden [3], liefern hier bessere Ergebnisse, sind aber bei starken Geschwindigkeitsabweichungen problematisch. Es zeigt sich, dass sich die von uns vorgeschlagenen rhythmusbasierten Merkmale zusammen mit einem auf *Konstellationssuche* basierenden Ansatz zur indexbasierten Audioidentifikation [4] zur robusten Identifikation auch zeitlich stark deformierter Audiosignale einsetzen lassen. Dieser Ansatz beruht darauf, jedes Signal d einer Datenbank bekannter Stücke durch eine Menge $F[d] \subset \mathbb{Z} \times X$ robuster Merkmale der Form (t, x) zu beschreiben, wobei $x \in X$ eine Merkmalsklasse und $t \in \mathbb{Z}$ den Merkmalszeitpunkt bezeichnet. Somit werden Merkmale nur bestimmten Zeitpunkten zugewiesen, was der Merkmalsmenge $F[d]$ eine charakteristische Zeitkonstellation verleiht und ein Schlüssel für die Effizienz der in [4] beschriebenen Methoden zur Audioidentifikation ist. Bei der Identifikation wird ein unbekanntes Signalfragment q gleichfalls in eine Merkmalskonstellation $F[q]$ überführt und kann nun als Teilkonstellation von d ab Position τ *identifiziert* werden, falls sich die entsprechenden Merkmale vermöge eines Zeitshifts $F[q] + \tau \subseteq F[d]$ ineinander überführen lassen. Die Erweiterung dieses Ansatzes auf die Identifikation zeitlich skalierter Signale erfolgt nun, indem jedem Merkmal zusätzlich zur Zeitkomponente noch eine Intervalldauer zugeordnet wird. Merkmale sind dann von der Form $(t, d, f) \in \mathbb{Z} \times \mathbb{R} \times X$ und die Audioidentifikation erfolgt nun durch zusätzliche Ermittlung eines Skalierungsfaktors s, so dass $s \cdot F[q] + \tau \subseteq F[d]$, also die durch Zeitverschiebung *und* Intervallskalierung modifizierten Merkmale der Anfrage in $F[d]$ enthalten sind, vgl. [2]. Die hier skizzierten Merkmale werden dazu, mit obigen Bezeichnungen, in die Form $(t, B(t), (R(t), M(t)))$ gebracht. Als Intervalldauer werden somit die lokale Beatperiode und als Merkmalsklasse die Rhythmus- und Metruminformationen verwendet. Zum Test des skizzierten Identifikationsverfahrens wurde eine Datenbank aus 100 Audiostücken (7h Musik) unterschiedlicher Genres, resultierend in ca. 50.000 Merkmalen verwendet. Angefragt wurden jeweils 300 Ausschnitte der Länge 30 Sekunden. Tabelle 1 zeigt die Identifikationsrate abhängig vom Skalierungsfaktor. Die Robustheit des Verfahrens gegenüber unterschiedlichen Signal-Degradationen zeigt Tabelle 2.

Faktor [%]	97	94	89	84	79
Identifikationsrate [%]	98	98	98	95	87
Faktor [%]	103	106	112	119	126
Identifikationsrate [%]	99	98	98	94	90

Tabelle 1: Identifikationsraten bei Signalskalierungen.

Degradationenstyp	Identifikationsrate [%]
Verrauschung (SNR=18dB)	98
Verrauschung (SNR=6dB)	92
MPEG@128 kBit/s	100
MPEG@32 kBit/s	96
Mikrofon (Abstand 30cm)	91

Tabelle 2: Robustheit gegenüber Signal-Degradationen.

Ausblick

In aktuellen Arbeiten wendet man sich dem allgemeineren Problem des *Audio Matchings* zu, bei dem es darum geht, unterschiedliche Versionen des Musikstücks zu einem gegebenen Audioausschnitt in einer Musikdatenbank aufzufinden. Erfolgversprechende Ansätze im Bereich klassischer Musik [5], nutzen zum Matching den bei klassischen Stücken oft sehr chrakteristischen Harmonieverlauf aus. Unter Einbeziehung der rhythmusbasierten Merkmale könnten solche Matching-Ansätze auf weitere Musikgenres, bei denen rhythmisches Verhalten deutlicher hervortritt als Harmonieverläufe, erweitert werden.

Literatur

[1] Mark A. Bartsch und Gregory H. Wakefield, *Audio thumbnailing of popular music using chroma-based representations*, IEEE Trans. on Multimedia, vol. 7 (1), pp. 96–104, 2005.

[2] Thorsten Gehrmann *Robuste rhythmusbasierte Identifikation akustischer Signale*, Diplomarbeit, Universität Bonn, 2005.

[3] Pedro Cano, Eloi Battle, Harald Mayer und Helmut Neuschmied, *Robust Sound Modeling for Sound Identification in Broadcast Audio*, Proc. 112th AES Convention, Munich, Germany, 2002.

[4] Michael Clausen und Frank Kurth, *A Unified Approach to Content-Based and Fault Tolerant Music Recognition*, IEEE Trans. on Multimedia, vol. 6 (5), 2004.

[5] Meinard Müller, Frank Kurth und Michael Clausen, *Audio Matching via Chroma-based Statistical Features*, ISMIR, London, GB, 2005.