Vektorquantisierung chromabasierter Audiomerkmale

Frank Kurth[1], Meinard Müller[2], Sebastian Ewert[3], Michael Clausen[3]

[1] *FGAN-FKIE, 53343 Wachtberg-Werthhoven, Deutschland, Email: kurth@fgan.de*

[2] *Max-Planck-Institut für Informatik, 66123 Saarbrücken, Deutschland, Email: meinard@mpi-inf.mpg.de*

[3] *Universität Bonn, Institut für Informatik III, 53117 Bonn, Deutschland, Email: {ewerts,clausen}@iai.uni-bonn.de*

Einleitung

Chromabasierte Audiomerkmale haben sich in den letzten Jahren als ein mächtiges Werkzeug zur Analyse von Musiksignalen erwiesen. Insbesondere konnten durch den Einsatz von Chromamerkmalen große Fortschritte beim *Audiomatching* harmoniebasierter Musik erzielt werden. Das Ziel des Audiomatchings besteht darin, bei Anfrage eines kurzen Abschnitts einer CD-Aufnahme alle hierzu musikalisch ähnlichen Abschnitte innerhalb einer Kollektion von Musikaufnahmen zu identifizieren. Im Hinblick auf ein effizientes und auf große Datenmassen skalierendes Verfahren ist die Möglichkeit zur Quantisierung und die damit verbundene Indexierbarkeit der Chromamerkmale sehr wichtig. In diesem Beitrag stellen wir zwei Methoden zur Chromaquantisierung vor. Die erste Methode basiert auf einem Clusteringansatz für den wir den bekannten LBG-Algorithmus geeignet adaptieren. Die zweite Methode nutzt semantisches Vorwissen über den Merkmalsraum aus, das sich aus dem auf der wohltemperierten Stimmung basierenden Harmoniekonzept für westliche Musik ergibt. Abschließend vergleichen wir die aus den beiden Methoden resultierenden Codebücher im Rahmen des indexbasierten Audiomatchings.

Chromabasierte Merkmale

Im ersten Schritt der Merkmalsextraktion wird das zu transformierende Audiosignal in 88 Tonhöhenbänder (entsprechend den Noten A0 bis C8) 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 Energieverteilung der im Audiosignal vorkommenden Frequenzen auf die 12 Chromabänder widerspiegelt, siehe Abb. 1.

Die so erhaltenen Chromamerkmale sind durch die Identifikation modulo Oktaven gleicher Tonhöhenbänder robust unter Klangfarbenänderung und durch die Normalisierung zusätzlich 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 auf Einheitslänge normalisiert und die Datenrate der Vektorfolge wird durch Downsampling auf 1 Hz reduziert, siehe Abb. 1. Die resultierenden Merkmale liegen somit im ersten Quadranten $\mathcal{F} := \{v \in [0,1]^{12} \mid \|v\|_2 = 1\}$ der Einheitssphäre des $\mathbb{R}^{12}$. Sie werden im weiteren abkürzend mit *CENS* (**C**hroma **E**nergy distribution **N**ormalized **S**tatistics) bezeichnet, siehe [KM08].

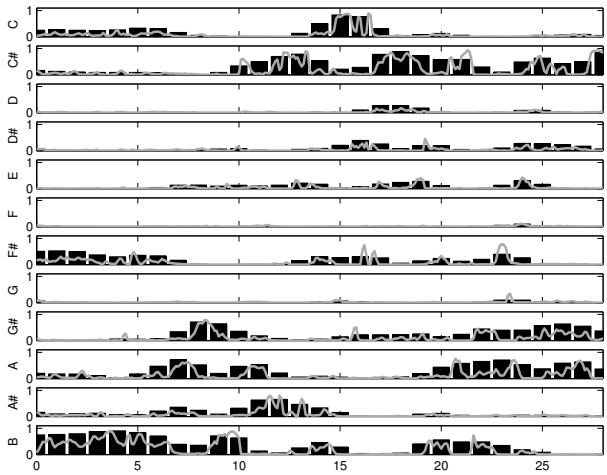

Abbildung 1: Takte 44–55 einer Interpretation von Vivaldis Frühling RV 269, Nr. 1. Helle Kurven: lokale Chroma-Energieverteilungen (10 Hz); dunkle Balken: CENS-Merkmalsfolge (1 Hz).

Quantisierung mittels LBG-Verfahren

Zur Quantisierung der CENS-Merkmale wird ein Codebuch $\{c_1, \ldots, c_R\} \subset \mathcal{F}$ geeigneter Größe R konstruiert. An Stelle eines CENS-Merkmals v wird dann nur noch der Index $i \in [1:R]$ eines Codebuchvektors c_i gespeichert, der minimalen Abstand zu v besitzt.

Die Konstruktion eines Codebuchs mittels des bekannten LBG-Verfahrens, das in Abb. 2 (links), anhand eines zweidimensionalen Beispiels mit sechs Vektoren (Kreise) und einer Codebuchgröße von $R = 2$ illustriert ist, geschieht wie folgt. Zunächst wird das Codebuch, etwa mit Zufallsvektoren, initialisiert. In Abb. 2 sind Codebuchvektoren durch Kreuze dargestellt. Der LBG-Algorithmus iteriert dann jeweils zwei Schritte. Im ersten Schritt wird jeder Vektor einem bezüglich eines geeigneten Abstandsmaßes (hier Euklidabstand) am nächsten gelegenen Codebuchvektor zugewiesen (in (1) durch Pfei-

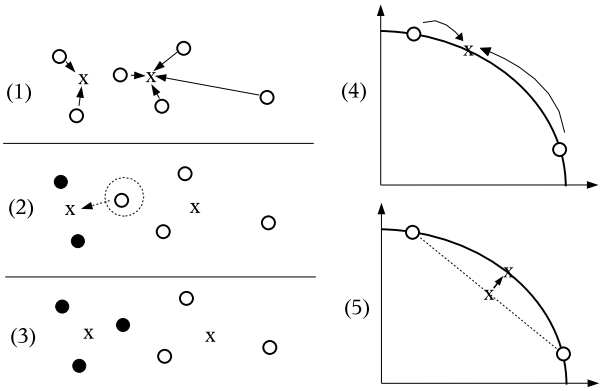

Abbildung 2: Links: LBG-Verfahren für 6 Vektoren und 2 Codebuchvektoren. Rechts: Modifizierte Neubestimmung von Codebuchvektoren für das LBG-Verfahren auf Sphären.

le skizziert). Die durch die Zuweisung resultierenden zwei Cluster von Vektoren sind in (2) durch schwarze und weiße Kreise hervorgehoben. Durch Mittelwertbildung aller zu einem Cluster gehöriger Vektoren wird im zweiten Schritt ein neuer Codebuchvektor für jedes Cluster bestimmt (Kreuze in (2)). Eine weitere Iteration beider Schritte liefert die in (3) dargestellten Cluster mit zugehörigen Codebuchvektoren. Hierbei wurde der in (2) gestrichelt markierte Vektor einem anderen Cluster zugewiesen. Die LBG-Iteration wird abgebrochen, wenn sich der mittlere Abstand der Vektoren zu den zugehörigen Codebuchvektoren nur noch in geringem Maße ändert.

Zur Erweiterung des LBG-Verfahrens auf die in $\mathcal{F}$ liegenden CENS-Merkmale verwenden wir als Abstandsmaß den Winkel zwischen zwei $\mathcal{F}$-Vektoren. Unser Verfahren arbeitet dann im wesentlichen analog zu LBG. Lediglich zur Neubestimmung der Codebuchvektoren wird zusätzlich zur Mittelung ein Projektionsschritt eingeführt, der in Abb. 2 (rechts), vereinfachend auf dem 2D-Einheitskreis illustriert ist: (4) zeigt die zwei zu einem Codebuchvektor gehörigen Vektoren. Die Neubestimmung des Codebuchvektors durch Mittelung liefert den in (5) skizzierten Vektor unterhalb des Einheitskreises. Die Projektion dieses Vektors auf den Einheitskreis liefert nun den neuen Codebuchvektor. Praxistests mit Codebuchgrößen $50 \leq R \leq 200$ zeigen für die CENS-Merkmale eine nur wenig langsamere Konvergenzgeschwindigkeit der modifizierten LBG-Variante im Vergleich zum klassischen LBG-Verfahren.

Quantisierung mittels Notenmodell

Ein Nachteil des LBG-basierten Ansatzes zur Codebuchbestimmung ist das hierfür benötigte repräsentative Trainingsmaterial und die potenzielle Gefahr des Overfittings. Unter Verwendung von Vorwissen über das typische zu erwartende Aussehen von CENS-Vektoren bei westlicher Musik in wohltemperierter Stimmung konstruieren wir nun ein vom Trainigsmaterial unabhängiges Codebuch. Beim Vorliegen harmonischer Musik erwarten wir eine Konzentration der Energie auf einige wenige Komponenten des CENS-Vektors. Ein vorliegender C-Dur-Akkord resultiert beispielsweise in CENS-Vektoren

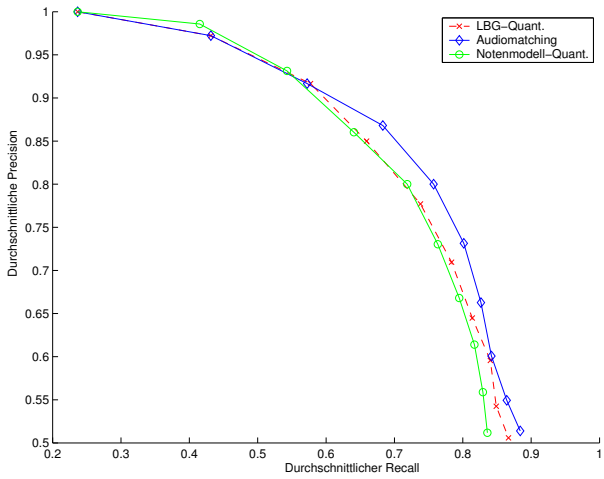

Abbildung 3: Precision-Recall-Diagramm beider Quantisierungsvarianten beim indexbasierten Audiomatching im Vergleich zum klassischen Audiomatching-Verfahren.

nahe $\frac{1}{\sqrt{3}}(1,0,0,0,1,0,0,1,0,0,0,0) \in \mathcal{F}$. Tests anhand von 55 Stunden klassischen Audiomaterials bestätigen, dass bei 95% der resultierenden CENS-Vektoren mehr als 50% der Energie in höchstens vier Komponenten enthalten ist. Bezeichnet δ_i den i-ten Einheitsvektor im $\mathbb{R}^{12}$, so definieren wir C_j als die Menge aller Vektoren $\frac{1}{\sqrt{j}}(\delta_{k_1} + \cdots + \delta_{k_j})$, $1 \leq k_1 < \cdots < k_j \leq 12$, mit genau j dominanten Komponenten. Motiviert durch unsere Experimente betrachten wir hier zunächst die Menge $C_1 \cup \cdots \cup C_4$ aller 793 Vektoren von bis zu vier dominanten Komponenten. Weitere Untersuchungen zeigen, dass an Stelle der, die einzelnen Noten eines Akkords repräsentierenden, Einheitsvektoren δ_i vielmehr mittels eines Obertonmodells konstruierte Vektoren $\tilde{\delta}_i$ verwendet werden sollten, was schließlich zum hier vorgeschlagenen modellbasierten Codebuch führt [KM08].

Anwendung: Audiomatching

Während der klassische Ansatz zum Audiomatching die unquantisierten CENS-Merkmale einer Anfrage mit den Merkmalen der Musikkollektion vergleicht, kann basierend auf der hier vorgestellten CENS-Quantisierung ein (Datenbank-) Index aus Codebuchindizes erstellt werden, mittels dessen ein stark beschleunigtes Audiomatching erreicht werden kann [KM08]. Abb. 3 zeigt das gemittelte Precision-Recall-Diagramm für 36 Audiomatching-Anfragen auf einer Datenkollektion von 110 Stunden klassischer Musik. Während der klassische Ansatz leicht bessere Ergebnisse liefert, zeigen die beiden indexbasierten Varianten mit Codebuchgrößen von jeweils 200 (LBG-Quantisierung) und 793 (Quantisierung mittels Notenmodell) vergleichbar gute Ergebnisse bei gleichzeitiger Beschleunigung des Audiomatchings um (von der Anfragelänge abhängige) Faktoren von 15-20.

Literatur

[KM08] Kurth, F., Müller, M.: Efficient Index-based Audio Matching. IEEE Trans. on Audio, Speech, and Language Processing **16**(2), 2008, 382–395.