Vergleich von PCA- und Autoencoder-basierter Dimensionsreduktion von Merkmalssequenzen für die effiziente Musiksuche

Frank Zalkow, Meinard Müller

International Audio Laboratories Erlangen, 91058 Erlangen Deutschland, Email: frank.zalkow@audiolabs-erlangen.de

Einleitung

Die Problemstellung des Audio Matching verfolgt das Ziel anhand eines Audiobeispiels relevante Dokumente in einer Datenbank zu finden [6]. Typischerweise enthält diese Datenbank viele Musikstücke in jeweils unterschiedlichen Einspielungen. Bei Anfrage eines 10- bis 30-sekündigen Audioausschnitts sollen dann alle musikalisch entsprechenden Passagen in allen verfügbaren Einspielungen gefunden werden.

Aktuelle Verfahren zum Audio Matching basieren auf sogenannten Chromamerkmalen, welche die lokale Energie in den zwölf chromatischen Tonhöhenklassen repräsentieren [7, Kapitel 3]. Sie korrelieren daher stark mit dem Harmonieverlauf des zugrundeliegenden Musikstücks und sind robust gegenüber Änderungen in Klangfarbe und Dynamik. Eine gewisse zeitliche Robustheit wird zudem durch eine grobe Merkmalsrate von einem Merkmalsvektor pro Sekunde erreicht. Kurze Sequenzen von 20 dieser 12-dimensionalen Merkmalsvektoren werden als *Shingles* bezeichnet. Ein Vergleich auf Shingle-Ebene hat sich als geeignetes Mittel für das Audio Matching erwiesen [5]. Bei großen Datenbanken ist ein Abgleich von 240-dimensionalen Merkmalssequenzen jedoch äußerst kostspielig. Eine Dimensionsreduktion der Shingles ist somit notwendig, um diese Technik auf große Datenbanken anzuwenden.

Eine solche Dimensionsreduktion kann beispielsweise mit linearen Methoden, wie der Hauptkomponentenanalyse (PCA), oder nicht-linearen Methoden, wie Autoencodern, erreicht werden. In diesem Beitrag präsentieren wir erste Ergebnisse unserer Experimente zum Vergleich der Dimensionsreduktion mit PCA und Autoencodern. Wir deuten an, wie die Musiksuche beschleunigt werden kann, ohne bedeutend an Retrievalqualität zu verlieren.

Szenario und Vorarbeiten

Immer größere Musikdatenbestände erfordern flexible, inhaltsbasierte Suchverfahren für Audiodaten [2]. Diese Arbeit behandelt die Musiksuche beziehungsweise das Audio Matching [6] für klassische Musik. Das Szenario ist in Abbildung 1 skizziert: Gegeben ist eine Datenbank $\mathcal{D}$ mit Audioaufnahmen klassischer Musik $D \in \mathcal{D}$. Es wird eine Suchanfrage $Q \in \mathcal{Q}$ angefragt, welche aus einem kurzen Audioausschnitt besteht, beispielsweise ein kurzer Ausschnitt aus einer Aufnahme von Ludwig van Beethovens 5. Sinfonie. Die Retrievalaufgabe besteht darin, alle Aufnahmen in der Datenbank zu identifizieren, die dem Musikstück der Anfrage entsprechen. In dem genannten Beispiel sollten also idealerweise verschiedene Aufnahmen von Ludwig van Beethovens 5. Sinfonie zurückgegeben werden, etwa Orchesteraufnahmen mit Dirigenten wie Wilhelm Furtwängler und Leonard Bernstein und Klaviertranskriptionen mit Pianisten wie Glenn Gould und Konstantin Scherbakov. Da die Interpretation der Anfrage nicht jenen der Datenbank entsprechen muss, ergeben sich verschiedenen Herausforderungen beim Retrieval: Es kann Unterschiede im Tempo, der Instrumentierung und der Artikulation zwischen der Anfrage und den relevanten Datenbankdokumenten geben.

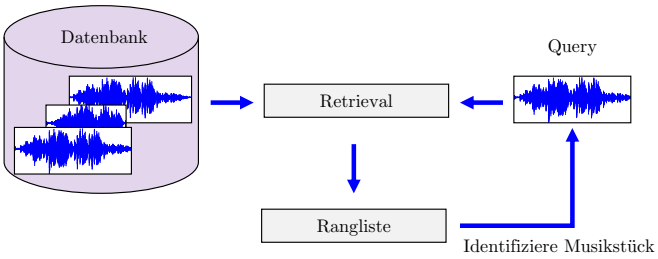

Abbildung 1: Retrieval Verfahren.

Ein Ansatz zur Lösung dieser Aufgabe ist, alle Datenbankdokumente sowie die Suchanfrage in Chromamerkmale zu überführen und diese zu vergleichen. Diese Merkmale repräsentieren die lokale Energie in den zwölf chromatischen Tonhöhenklassen $\{\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{H}\}$ [7, Kapitel 3]. Sie korrelieren daher stark mit dem Harmonieverlauf des zugrundeliegenden Musikstücks und weisen eine gewisse Robustheit gegenüber Änderungen in Klangfarbe und Dynamik auf. Idealerweise würde man die resultierenden Merkmalssequenzen lokal alignieren um die Tempounterschiede zwischen der Anfrage und den relevanten Datenbankdokumenten auszugleichen, beispielsweise mittels *Subsequence Dynamic Time Warping* [7, Kapitel 7]. Für große Datenbanken ergibt sich dabei allerdings eine hohe Laufzeit. Um die Suche zu beschleunigen, wurde in einer vorherigen Arbeit [5] für dieses Problem ein *Shingeling*-Ansatz vorgestellt: Bei einem Shingle $S \in \mathbb{R}^{12 \times L}$ handelt es sich um eine kurze Sequenz von Merkmalsvektoren der Länge $L \in \mathbb{N}$, die zum Vergleich benutzt wird. Die Anfrage besteht dabei aus einem einzigen Shingle. Für den Vergleich der Anfrage mit den Shingles der Datenbank muss ein geeignetes Ähnlichkeits- beziehungsweise Abstandsmaß festgelegt werden, beispielsweise der euklidische Abstand. Der Abstand zwischen dem ähnlichsten Shingle eines Datenbankdokuments und der Anfrage wird als Abstandsmaß für das Dokument herangezogen. Als Retrievalergebnis erhält man dann eine Rangliste durch die Sortierung der Datenbankdokumente anhand deren Abstandsmaße. Für die Shingles hat sich

Tabelle 1: Datensatz.

Komponist	Werk	Satz	Aufn.	hh:mm:ss
Beethoven	Op. 67	1	10	01:12:07
		2	10	01:44:53
		3	10	01:02:53
		4	10	01:48:00
Chopin	Op. 17 Nr. 4		63	04:32:56
	Op. 24 Nr. 2		64	02:26:38
	Op. 30 Nr. 2		33	00:46:52
	Op. 63 Nr. 3		87	03:07:08
	Op. 68 Nr. 3		51	01:25:58
Vivaldi	RV 315	1	7	00:37:40
		2	7	00:17:23
		3	7	00:20:40
			359	19:23:08

eine Sequenzlänge von $L = 20$ bei einer groben Merkmalsrate von einem Hertz als geeignet herausgestellt. Die resultierende Dimensionalität der Shingles $d \in \mathbb{N}$ ist also $d = 12 \times L = 240$. Die Merkmale sind sogenannte CENS-Merkmale [8]: Dabei handelt es sich um Chromamerkmale, bei denen eine zeitliche Glättung und eine Unterabtastung vorgenommen wurde. Die Frames sind jeweils ℓ^2-normalisiert. Um trotz des Verzichts auf lokale Aligninerungstechniken Tempounterschiede zwischen der Anfrage und den relevanten Datenbankdokumenten auszugleichen, werden verschiedene gedehnte und gestauchte Varianten der Anfrage benutzt. Hierfür haben sich die Stretching-Parameter $T = \{0.8, 1, 1.25\}$ als sinnvoll erwiesen.

In Vorarbeiten [5] wurde bereits die Dimensionalität der einzelnen Chromavektoren separat mittels Hauptkomponentenanalyse (PCA) reduziert. Dabei hat sich herausgestellt, dass sich die Retrievalergebnisse kaum verschlechtern, wenn man die Dimensionalität von 12 auf bis zu 8 reduziert (also $d = 160$) und nur moderat verschlechtert bei einer Dimensionalität von bis zu 4 (also $d = 80$). Das Ziel unserer Arbeit ist es, die Dimensionalität des Shingles $d \in \mathbb{N}$ weiter zu verringern, ohne bedeutend an Retrievalqualität zu verlieren. Anders als in der Vorarbeit wird dabei nicht die Dimensionalität der einzelnen Chromavektoren separat verringert, sondern der gesamten Merkmalssequenz insgesamt.

Datensatz

Tabelle 1 zeigt eine Übersicht über den verwendeten Datensatz. Er wurde so erstellt, dass er weitgehend ähnlich zu jenem der Vorbeiten [5] ist. Insgesamt gibt es etwa 19,5 Stunden Audiomaterial und 359 Aufnahmen, wobei die einzelnen Werke keinesfalls balanciert sind. Den Großteil der Daten machen Mazurken von Frédéric Chopin aus, welche im Rahmen des Mazurka-Projekts[1] gesammelt wurden [9].

Dimensionsreduktion

Die Hauptkomponentenanalyse (PCA) ist eine gängige lineare Methode zur Dimensionsreduktion, bei der die Daten in einen niederdimensionalen orthogonalen Unterraum projiziert werden. Hierbei besteht das Ziel die Varianz der resultierenden Dimensionen zu maximieren [1, Kapitel 12].

Ein Autoencoder [3, 4] ist eine nicht-lineare Methode zur Dimensionsreduktion. Es handelt sich um ein künstliches neuronales Netzwerk, das zum Ziel hat, eine Repräsentation für einen Datensatz zu lernen, aus welchem die ursprünglichen Daten wieder rekonstruieren werden können. Typischerweise hat ein Autoencoder eine symmetrische Struktur, bei der die mittlere Schicht eine geringere Dimensionalität als der Netzwerkeingang hat. Die Ausgabe dieser Schicht wird auch *Code* genannt. Die Hälfte des Autoencoders, die zum Code führt, wird als *Encoder* bezeichnet und entspreched heißt die zweite Hälfte *Decoder*. Dieser Decoder hat die Aufgabe, den Netzwerkeingang aus dem Code zu rekonstruieren, welchen man sich somit als eine komprimierte Repräsentation des Netzwerkeingangs vorstellen kann. Die Struktur des verwendeten Autoencoders ist in Abbildung 2 skizziert.

Autoencoder mit nur jeweils einer Schicht im Encoder und Decoder und mit linearen Aktivierungsfunktionen lernen eine Projektion, die ähnlich zur PCA ist, ohne dass die Projektion orthogonal sein muss. Durch die nicht-linearen Aktivierungsfunktionen sowie das Vorhandensein mehrerer Schichten können Autoencoder als mächtigere Verallgemeinerung der PCA aufgefasst werden [4, Kapitel 14].

Experimente und Ergebnisse

Für unsere Experimente wollen wir Suchanfragen aus der gesamten Datenbank verwenden. Um diese abzudecken, wählen wir je 10 äquidistant verteilte Shingles aus den 359 Aufnahmen (siehe Tabelle 1). Wir führen das Retrieval der 3590 Shingles mit der gesamten Datenbank durch und erhalten für jede Anfrage eine Rangliste. Damit berechnen wir Evaluatierungsmaße, die gängig im Bereich des Information Retrieval sind. **MAP** (Mean Average Precision) ist ein Maß, dass die Retrievalqualität anhand der gesamten Rangliste anzeigt [7, Kapitel 7]. **P@2** steht für den Anteil der relevanten Dokumente innerhalb der ersten beiden Aufnahmen der Rangliste. Dieses Evaluatierungsmaß wurde gewählt, weil P@1 für unsere Experimente nicht aussagekräftig ist. Dies liegt daran, dass alle Anfragen der Datenbank entstammen und daher der erste Treffer stets identisch mit der Anfrage ist. **P@r** steht für den Anteil der relevanten Dokumente innerhalb der ersten r Aufnahmen der Rangliste, wobei $r \in \mathbb{N}$ die Anzahl der relevanten Dokumente für die betreffende Anfrage ist. Die genannten Maße werden über alle Anfragen hinweg gemittelt.

Wir wollen verschiedene Aspekte bei unseren Experimenten untersuchen: Zum einen möchten wir prüfen, wie weit die Reduktion der Shingle-Dimensionalität möglich ist, ohne die Retrieval-Ergebnisse zu stark zu verschlechtern. Hierzu betrachten wir die Werte $d \in \{12, 30, 120, 240\}$. Zum anderen möchten wir für den Autoencoder die Loss-Funktionen Mean Square Error und Binary Crossentro-

[1] `http://www.mazurka.org.uk/`

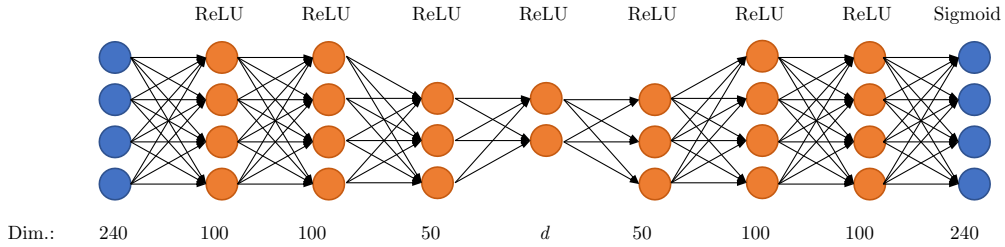

Abbildung 2: Visualisierung der genutzten Autoencoder-Struktur.

Tabelle 2: Retrievalergebnisse für **(a)** keine Merkmalsreduktion, **(b)** Merkmalsreduktion mit PCA und **(c)** Merkmalsreduktion mit Autoencoder.

		MAP	P@2	P@r	MSE	Epochen
(a)	$d = 240$	0,975	0,995	0,955	0,0	
(b)	$d = 12$	0,933	0,977	0,890	0,022	
	$d = 30$	0,965	0,993	0,936	0,013	
	$d = 120$	0,975	0,996	0,956	0,001	
(c)	$d = 12$, $\mathcal{L}$=BC	0,953	0,976	0,924	0,014	538
	$d = 12$, $\mathcal{L}$=MSE	0,926	0,971	0,879	0,015	707
	$d = 30$, $\mathcal{L}$=BC	0,972	0,994	0,950	0,011	513
	$d = 30$, $\mathcal{L}$=MSE	0,921	0,983	0,869	0,016	254
	$d = 120$, $\mathcal{L}$=BC	0,978	0,994	0,962	0,009	605
	$d = 120$, $\mathcal{L}$=MSE	0,974	0,994	0,953	0,009	923

py $\mathcal{L} \in \{\text{MSE}, \text{BC}\}$ vergleichen. Desweiteren möchten wir untersuchen, ob der Rekonstruktionsfehler (Mean Square Error, gemittelt über alle Shingles hinweg) und die Qualität der Retrievalergebnisse korrelieren.

Tabelle 2a zeigt die Ergebnisse ohne jegliche Dimensionsreduktion. Der hohe P@2 Wert von 0,995 zeigt an, dass es sich bei den ersten beiden Dokumenten der Rangliste für die meisten Anfragen um das richtige Musikstück handelt.

Tabelle 2b zeigt die Ergebnisse für die Dimensionsreduktion mit PCA für $d \in \{12, 30, 120\}$. Die Retrievalergebnisse werden mit steigender Dimensionalität besser, wobei sie für $d = 120$ ebenso gut sind wie die Ergebnisse ohne Dimensionsreduktion (P@2: 0,996). Wie zu erwarten, korreliert der Rekonstruktionsfehler (MSE) negativ mit der Dimensionalität und damit auch den Retrievalergebnissen.

Tabelle 2c zeigt die Ergebnisse für die Dimensionsreduktion mit Autoencoder. Auch hier lässt sich beobachten, dass die Retrievalergebnisse mit steigender Dimensionalität tendenziell besser werden. Für alle Experimente zeigt sich, dass das Training mit $\mathcal{L} = \text{BC}$ bessere Ergebnisse erzielt als das Training mit $\mathcal{L} = \text{MSE}$. Der Autoencoder scheint einen gewissen Vorteil gegenüber traditionelleren Methoden aufzuweisen. Beispielsweise ist bei der geringsten Dimensionalität $d = 12$ mittels PCA der P@r-Wert 0,890. Für den Autoencoder hingegen ist bei gleicher Dimensionalität der beste Wert 0,924. Für $d = 120$ ist der beste P@r-Wert des Autoencoders sogar 0,962, wobei der entsprechende Wert ohne Dimensionsreduktion 0,955 ist. Der Autoencoder bestätigt auch, dass ein höherer Rekonstruktionsfehler tendenziell mit schlechteren Retrievalergebnissen einhergeht. Dass dieser Zusammenhang nicht zwangsläufig besteht, zeigt sich allerdings, wenn man für $d = 12$ den Rekonstruktionsfehler der PCA (MAP: 0,933 und MSE: 0,022) mit dem des Autoencoders bei $\mathcal{L} = \text{MSE}$ (MAP: 0,926 und MSE: 0,015) vergleicht.

In weiteren Experimenten hat sich zudem herausgestellt, dass die Standardisierung der Merkmale auf den Mittelwert 0 und die Standardabweichung 1 keinen wesentlichen Einfluss auf die Retrievalergebnisse hat. Ein Grund dafür könnte sein, dass die CENS-Merkmale bereits spaltenweise normalisiert sind und sich daher in einem begrenzten Wertebereich bewegen.

Der Gewinn an Retrievalqualität durch den Autoencoder ist noch mäßig, aber die Ergebnisse geben uns die Hoffnung, dass es sinnvoll ist, diese Forschungsidee weiterzuverfolgen.

Ausblick

Wir haben in diesen Beitrag das Potential zweier Methoden zur Dimensionsreduktion für das Beschleunigen der Musiksuche für klassische Musik vorgestellt. Insbesondere haben wir die lineare Methode der Hauptkomponentenanalyse (PCA) sowie nicht-lineare Autoencoder verglichen. Wir haben Ergebnisse diskutiert und dabei gezeigt, dass die traditionelle PCA gute Ergebnisse erzielt, aber mit Autoencodern durchaus kleinere Verbesserungen erreicht werden können. Dies deutet darauf hin, dass Deep Learning durchaus Potential für diese Aufgabenstellung hat. In diesem Beitrag wurde ein erster Schritt vorgestellt. Darauf aufbauend müssen zukünftig weitere Aspekte erforscht werden, wie zum Beispiel das systematische Variieren der Autoencoder-Topologie und das Durchführen der Experimente auf größeren Datensätzen unter Anwendung von Kreuzvalidierungsverfahren.

Danksagung: Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft unterstützt (DFG MU 2686/11-1). Die International Audio Laboratories Erlangen sind ein Zusammenschluss der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) und dem Fraunhofer-Institut für Integrierte Schaltungen IIS. Die Autoren bedanken sich bei der Open-Source-Community, insbesondere den Entwicklern von Python[2], Librosa[3], Tensorflow[4] und Keras[5].

[2] `https://www.python.org`
[3] `https://github.com/librosa/librosa`
[4] `https://www.tensorflow.org`
[5] `https://keras.io`

Literatur

[1] Bishop, C. M.: Pattern recognition and machine learning. Springer, New York (2006).

[2] Casey, M. A., Veltkap, R., Goto, M., Leman, M., Rhodes, C. & Slaney, M.: Content-Based Music Information Retrieval: Current Directions and Future Challenges. In: Proceedings of the IEEE (2008), **96**, Nr. 4, S. 668–696.

[3] Charte, D., Charte, F., García, S., del Jesus, M. J. & Herrera, F.: A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. In: Information Fusion (2018), **44**, S. 78–96.

[4] Goodfellow, I., Bengio, Y. & Courville, A.: Deep Learning. MIT Press, Cambridge and London (2016). `http://www.deeplearningbook.org`.

[5] Grosche, P. & Müller, M.: Toward Characteristic Audio Shingles for Efficient Cross-Version Music Retrieval. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), S. 473–476. Kyoto, Japan (2012).

[6] Kurth, F. & Müller, M.: Efficient Index-Based Audio Matching. In: IEEE Transactions on Audio, Speech, and Language Processing (2008), **16**, Nr. 2, S. 382–395. URL `http://dx.doi.org/10.1109/TASL.2007.911552`.

[7] Müller, M.: Fundamentals of Music Processing. Springer Verlag (2015).

[8] Müller, M., Kurth, F. & Clausen, M.: Chroma-Based Statistical Audio Features for Audio Matching. In: Proceedings of the IEEE Workshop on Applications of Signal Processing (WASPAA), S. 275–278. New Paltz, NY, USA (2005).

[9] Sapp, C. S.: Comparative analysis of multiple musical performances. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), S. 497–500. Vienna, Austria (2007).