Beatles vs. Bach: Merkmalsextraktion im Phasenraum von Audiodaten

Ingo Mierswa
Universität Dortmund
Fachbereich Informatik
Lehrstuhl für Künstliche Intelligenz
Email: mierswa@ls8.cs.uni-dortmund.de

Abstract

Gute Ergebnisse bei der Klassifikation von Audiodaten durch maschinelle Lernverfahren setzen die Extraktion aussagekräftiger Merkmale voraus. Es werden Ansätze vorgestellt, mittels Techniken der Zeitreihenanalyse einen Satz solcher Merkmale zu erzeugen. Insbesondere die Anwendung einer Phasenraumtransformation stellt die Basis zur Extraktion neuer Merkmale dar. Dieser Satz von Merkmalen ist invariant gegenüber einer zeitlichen Verschiebung der Musikstücke und robust bezüglich ihrer Qualität. Zudem erlauben die generierten Merkmale nicht nur eine musikalische Interpretation, sondern erhöhen darüberhinaus auch die Performanz bei der Klassifikation nach Genres.

1 Einleitung

Die Eingabe eines Lernverfahrens ist in einer Repräsentationssprache L_E dargestellt. L_E beschreibt im Folgenden eine Instanzmenge E, wobei jede Instanz $e \in E$ ein Vektor von Ausprägungen einer Anzahl Merkmale darstellt. Ein Lernverfahren versucht für solche Instanzen eine Hypothese aus einer Hypothesensprache L_H zu finden, die Rückschlüsse über die Natur der Daten sowie die Anwendung auf neue Instanzen erlaubt [Mitchell, 1996; Witten and Frank, 2000].

Audiodaten können als univariate Zeitreihen aufgefaßt werden. Zu jedem Zeitpunkt i ist die Elongation a_i datiert. Sei L_E von der Gestalt, dass eine Instanz $e \in E$ für jeden der n Zeitpunkte einer Zeitreihe ein Merkmal besitzt. Damit ist jedes Musikstück eine Instanz aus E.

Ein Lernverfahren liefert auch in diesem Fall für eine Menge von Instanzen eine Hypothese. Jedoch verschenkt man auf diese Weise Information, da eine inhärente Eigenschaft von Reihen nicht beachtet wird: Die Werte liegen geordnet vor [Pyle, 1999]. Sei L'_E eine Sprache, deren Elemente diese Information berücksichtigen. Gesucht ist dann eine Transformation von L_E zu L'_E [Morik, 2000]. Dieser Vorgang stellt die Extraktion aussagekräftiger und robuster Merkmale dar. Sie müssen invariant sein gegenüber Translationen in der Zeit und verschiedene Versionen oder Qualitäten desselben Musikstückes sollten zumindest nicht qualitativ zu unterschiedlichen Ausprägungen der Merkmale führen.

Im Folgenden werden in Abschnitt 2 Ansätze erläutert, Erkenntnisse aus dem Bereich der Mididateien und der digitalen Sprachverarbeitung auf die Merkmalsextraktion aus Audiodaten zu übertragen. In Abschnitt 3 werden Musikdaten als Zeitreihen eingeführt und die Datenmenge diskutiert. In den nächsten Abschnitten wird die Extraktion robuster Merkmale aus Audiodaten beschrieben und in Abschnitt 7 Ergebnisse der Anwendung auf reale Daten aufgezeigt.

2 Merkmalsextraktion aus Audiodaten

Bisherige Ansätze zur Merkmalsextraktion aus Audiodaten behandelten aufgrund technischer Restriktionen hauptsächlich Mididateien. Bei diesen sind alle hörbaren Noten nach dem 12-Ton-System numeriert. Jede Note in einem Musikstück umfaßt die Tonhöhennummer, den Startzeitpunkt, die Dauer und die Lautstärke [Loy, 1989]. Der erste Schritt war der Übergang von einstimmigen zu polyphonen Stücken [Pickens, 1996]. Erst in jüngster Zeit konnte man auch Audiodaten in Form von akustischen Wellen verarbeiten. Es hat sich allerdings gezeigt, dass die direkte Übertragung von Erkenntnissen aus dem Bereich der Mididateien auf Audiodaten nicht zu guten Performanzergebnissen führte. Weitere Ursprünge der Merkmalsextraktion aus Audiodaten liegen in der Repräsentation und Verarbeitung von Sprachsignalen [Brillinger and Irizarry, 1998]. Man beschränkte sich zunächst auf die Klassifikation einzelner Töne, die Zuordnung zu Instrumenten oder auf die Unterscheidung zwischen männlicher und weiblicher Stimme [Irizarry, 2001]. Die jüngsten Ergebnisse bei der Behandlung von Audiodaten liefern eine Vielzahl von Merkmale, die historisch aus den einzelnen Ansätzen zusammengewachsen sind [Tzanetakis, 2002; Tzanetakis *et al.*, 2001]. Viele der bisher extrahierten Merkmale leiden jedoch unter einer hohen Anfälligkeit gegenüber Zeittranslationen oder verminderter Qualität des Ausgangsmaterials.

3 Musikstücke als Zeitreihen

Musikstücke sind akustische Wellen und damit reellwertige Funktionen der Zeit. Um ein Lied zu digitalisieren wird die Welle in viele kleine Abschnitte zerstückelt. Dies approximiert die kontinuierliche Funktion, wobei die Güte der Approximation und somit die Qualität der Digitalisierung von der Anzahl der Abschnitte abhängt (*sampling rate* [Hz]). Jedem dieser Zeitpunkte (*sample points*) eines Liedes wird die Elongation der Welle an diesem Zeitpunkt zugeordnet. Diese Diskretisierung der akustischen Welle erlaubt die Betrachtung eines Musikstückes als univariate Zeitreihe. Abbildung 1 zeigt eine solche Wellenform. Auf der X-Achse ist der Zeitverlauf des Liedes dargestellt und auf der Y-Achse die jeweilige Elongation.

Angenommen, wir wollen Lieder mit einer Länge von drei Minuten klassifizieren. Wie viele Merkmale hat jede

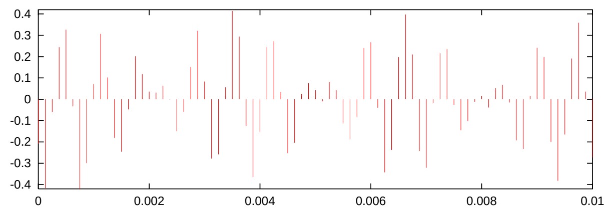


Abbildung 1: Beispielhafte Wellenform eines Liedausschnittes. Die äquidistanten Samples sind deutlich zu erkennen.

Instanz aus E? Die *sampling rate* der Wellenform in Abbildung 1 ist 44100 kHz, jeder sample point enthält die Daten für die beiden Stereokanäle [Roads, 1996]. Für ein Lied von drei Minuten ergeben sich

$$44100kHz \cdot 2 \cdot 180s \approx 16 \cdot 10^9$$

Werte. Durch eine Transformation zu L'_E, der Sprache mit wenigen extrahierten Merkmalen pro Instanz, ist eine Verbesserung der Vorhersageergebnisse [Ritthoff *et al.*, 2002; Liu and Motoda, 1998] sowie eine starke Kompression der Daten zu erwarten.

4 Merkmale in der Zeitdimension

Zunächst werden Merkmale vorgestellt, die sich ohne Basistransformation in einen anderen Raum extrahieren lassen. Die Interpretationen dieser Merkmale beziehen sich stets auf das Tempo und die Lautstärke eines Liedes, da die X-Achse den Zeitverlauf widerspiegelt und die Y-Achse die Elongation. Im Folgenden bezeichnen wir Zeitreihen der Länge n mit $\{x_i\}_{i\in\{1,\ldots,n\}}$. Außerdem definieren wir die Abbildung $index$, welche für jedes x_i der Zeitreihe die Stelle auf der X-Achse liefert.

4.1 Mittlere Lautstärke

Die durchschnittliche Auslenkung einer Lautsprechermembran ist ein Grad für die mittlere Lautstärke. Da die Elongation jedoch um den Wert 0 schwankt und wir eine negative Auslenkung als genauso stark empfinden wie eine positive, berechnen wir statt des arithmetischen Mittels das absolute arithmetische Mittel:

$$LS(\{x_i\}_{i\in\{1,\ldots,n\}}) = \frac{1}{n}\sum_{i=1}^{n}|x_i|$$

Dies ist ein Maß für die Lautstärke, gemittelt über die gesamte Länge des Musikstückes. Eine Invarianz gegen Verschiebungen der Zeit erreicht man durch Auslassen der nicht von 0 verschiedenen Werte, insbesondere zu Beginn und am Ende eines Liedes.

4.2 Extremadifferenz

Die Differenzen zwischen den Extrema eines Liedes liefern einen ersten Hinweis auf die Geschwindigkeit. Ein einfacher Algorithmus zur Findung von Extremwerten [Ribbrock and Kurth, 2002] liefert wiederholt angewandt besonders herausstechende Werte, die übrigen Werte seien nach der Anwendung 0. Sei $E_x = \{x_i | x_i \neq 0\}$ die Menge der Extrema der Zeitreihe $\{x_i\}_{i\in\{1,\ldots,n\}}$, also derjenigen Werte x_i, die nach Anwendung des Algorithmus nicht 0 sind. O.B.d.A. seien diese in der Reihenfolge ihrer ursprünglichen Indizes mit $e_1, \ldots, e_m$ benannt. Der durchschnittliche

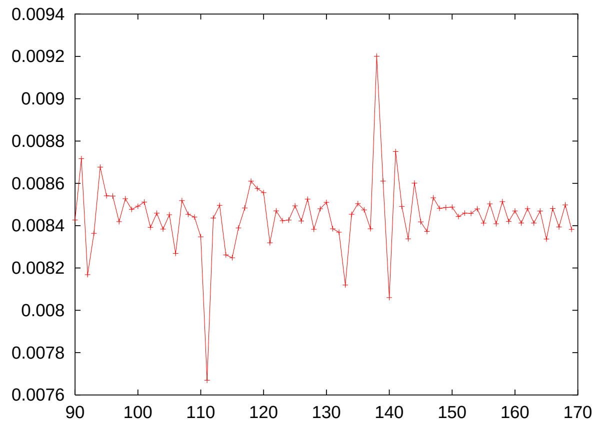


Abbildung 2: Die Werte der gemittelten absoluten Differenzen für Geschwindigkeiten zwischen 90 und 170 bpm. Die minimale Differenz bei 111 bpm ist deutlich zu sehen.

Abstand der Extrema ist dann

$$AVG_{diff}(E_x) = \frac{1}{|E_x|}\sum_{i=2}^{|E_x|}(index(e_i) - index(e_{i-1}))$$

Die Varianz der Differenzen errechnet sich kanonisch:

$$VAR_{diff}(E_x) =$$
$$\frac{1}{|E_x|}\sum_{i=2}^{|E_x|}((index(e_i) - index(e_{i-1})) - AVG_{diff}(E_x))^2$$

Die Positionen der Extrema selber sind nicht invariant gegenüber Translationen, wohl aber die Differenzen und damit auch der Durchschnitt sowie die Varianz der Differenzen.

4.3 Autokorrelation durch Phasenverschiebung

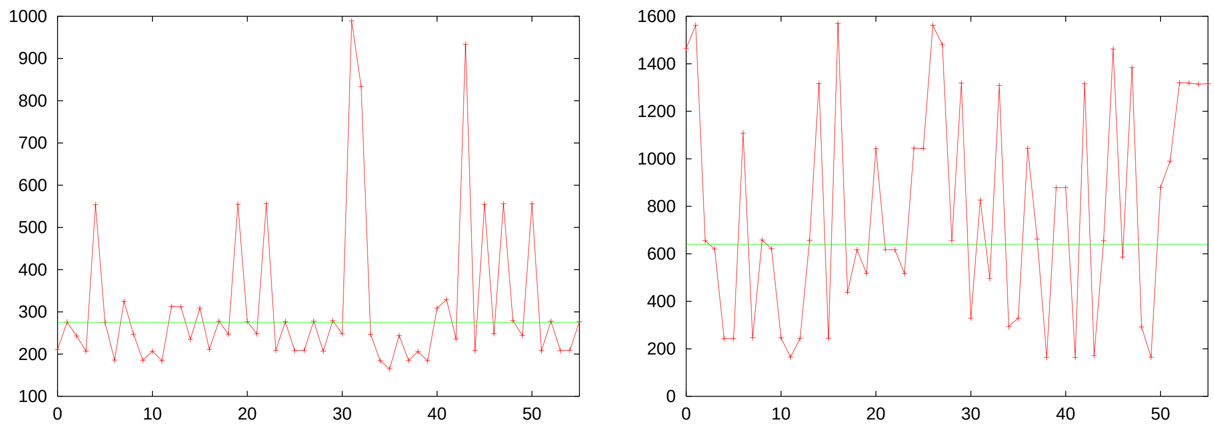
Es gibt verschiedene Ansätze das Tempo eines Liedes und den Rhythmus zu bestimmen [Dixon, 2001; Cemgil *et al.*, 2000]. Sie alle sind in der Lage, die genaue Position des ersten Schlages eines Taktes vorherzusagen. Diese Information ist ohnehin nicht invariant gegenüber der Zeit und daher stelle ich ein einfacheres Verfahren vor, welches die Geschwindigkeit alleine bestimmt.

Das Vorgehen basiert auf einer Phasenverschiebung der Musikstücke. Sei T die erwartete Zahl von Schlägen pro Takt und SR die *sample rate* des Stückes, also die Anzahl von Werten pro Sekunde. Iterativ wird nun das gesamte Stück um T Schläge nach rechts verschoben. Dies bedeutet für eine Geschwindigkeit von X Schlägen pro Minute (bpm: beats per minute) eine Verschiebung um

$$v = T \cdot SR \cdot \frac{60}{X}$$

Werte. Für jede Verschiebung berechnet man nun die gemittelte quadratische Differenz AD zum Original und erhält so ein Maß, wie gut das verschobene Stück autokorreliert. Iteriert man dieses Vorgehen für $X \in [start, ende]$, so erhält man für jede mögliche Geschwindigkeit zwischen $start$ und $ende$ einen Wert der Autokorrelation. Die gesuchte Geschwindigkeit entspricht der Stelle mit der minimalen Differenz, d. h. maximaler Korrelation:

$$AD_{min}(\{x_i\}_{i\in\{1,\ldots,n\}}) =$$
$$\min\left\{\frac{1}{n-v}\sum_{i=1}^{n-v}(x_i - x_{i+v})^2 \;\middle|\; X \in [start, end]\right\}$$



(a) Pop

(b) Klassik

Abbildung 3: Die Abbildungen (a) und (b) zeigen das Ergebnis einer gefensterten Fouriertransformation. Zu jedem Fenster wird die stärkste Frequenz bestimmt und aufgetragen, die horizontale Linie gibt die durchschnittliche Frequenz an. Bei dem klassischen Stück aus Abbildung (b) ist deutlich die höhere Durchschnittsfrequenz und die größere Varianz zu beobachten.

und schließlich

$$T(\{x_i\}_{i\in\{1,\dots,n\}}) = index(AD_{min}(\{x_i\}_{i\in\{1,\dots,n\}}))$$

Tests ergeben, dass dieses Verfahren für 85% aller Lieder die richtige Geschwindigkeit liefert und zudem mit Pausen innerhalb der Lieder umgehen kann.

Abbildung 2 zeigt die Werte der absoluten Differenzen für Geschwindigkeiten zwischen 90 und 170 bpm. Die Varianz dieser Korrelationstransformation gibt die Sicherheit an, mit der das Tempo bestimmt wurde.

5 Merkmale in der Frequenzdimension

Eine Basistransformation in den Raum der harmonischen Schwingungen nennt man Fouriertransformation. Sie liefert für jede der Basisschwingungen unterschiedlicher Frequenz und Phase den Anteil der Basisschwingung an der Gesamtfunktion [Schlittgen and Streitberg, 1997]. Das Ergebnis ist also ein Graph (Frequenzspektrum), der für jede Frequenz die Intensität der entsprechenden Grundschwingung an der Gesamtfunktion angibt [Tipler, 1991]. Die Transformierte einer Funktion f ist gegeben durch

$$FT(f) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f \cdot e^{i\nu x} dx$$

für eine feste Frequenz ν. Merkmale des Frequenzspektrums beziehen sich auf Melodien, Harmonien oder Instrumentierungen.

5.1 Peaks im Frequenzspektrum

Die höchsten Peaks eines Frequenzspektrums entsprechen den stärksten Frequenzen eines Liedes. Diese können durch besonders häufige Töne oder auch besonders häufige Oberschwingungen der Instrumente entstanden sein. Sie sind charakteristisch für die Klangfarbe eines Musikstückes. Die k höchsten Peaks eines Frequenzspektrums können mit einem Divide-and-Conquer Ansatz gefunden werden. Dieser sucht zunächst das Maximum und bestimmt damit den höchsten Peak. Sodann werden rekursiv die Peaks links und rechts bestimmt. Die Werte des aktuellen Peaks werden dabei ausgeschlossen. Die Intensität, die Frequenz und die Breite der k stärksten Peaks werden dann als Merkmale bereitgestellt.

5.2 Stärkste Frequenz in Zeitfenstern

Wendet man eine Fouriertransformation auf das gesamte Musikstück an, erhält man lediglich die Information, wie intensiv jede Frequenz insgesamt auftritt, nicht aber, an welchen Stellen des Liedes dieses geschieht. Daher ist die Berechnung der höchsten Peaks der Fouriertransformierten naturgemäß unabhängig von zeitlichen Verschiebungen.

Mit diesem Vorgehen verschenkt man jedoch die Information der zeitlichen Ordnung. Wir verschieben daher ein Zeitfenster mit Schrittweite s und Breite w über die gesamte Zeitreihe und berechnen in jedem Fenster

$$y_j = index(\max(FT(\{x_i\}_{i\in\{j\cdot s,\dots,j\cdot s+w\}})))$$

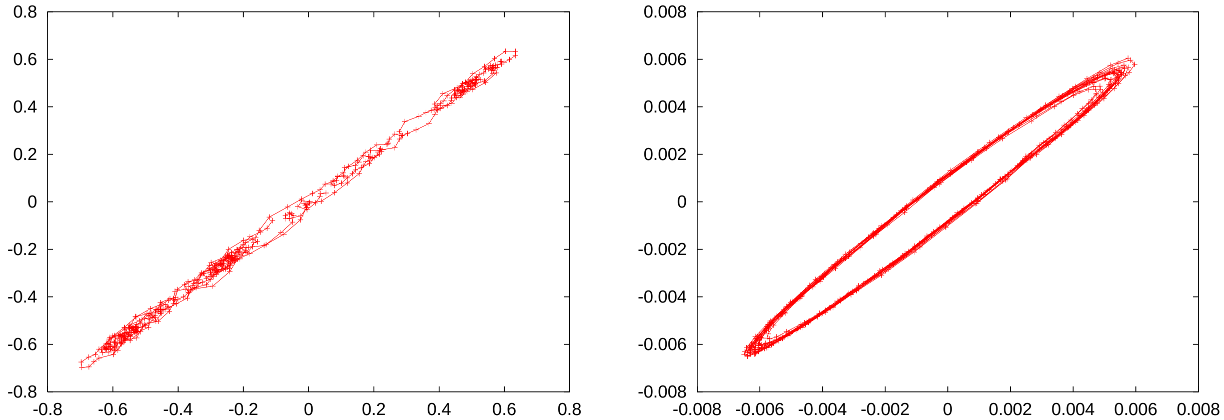
Diese y_j bilden erneut eine Zeitreihe $\{y_j\}_{j\in\{1,\dots,n/s-1\}}$, wobei der Wert eines jeden y_j die in dem betreffenden Zeitfenster stärkste Frequenz bezeichnet. Abbildung 3 zeigt das Ergebnis dieser Transformation für ein Klassik- und für ein Popstück. Man sieht deutlich die höheren Frequenzwerte und die größere Varianz von Frequenzen bei dem klassischen Stück. Als Merkmale bildet man folglich den Durchschnitt von $\{y_j\}_{j\in\{1,\dots,n/s-1\}}$ sowie die Varianz der Reihe. Die gefensterte Fouriertransformation ist also eine Rücktransformation in den Zeit-Raum. Die Varianz gibt zwar die zeitliche Veränderlichkeit an, jedoch sind die beiden Merkmale Durchschnitt und Varianz unabhängig vom genauen Zeitpunkt der auftretenden Ereignisse.

6 Merkmale im Phasenraum

Weitere Merkmale können durch eine Transformation der Musikstücke in den Phasenraum gewonnen werden. Diese Transformation erfolgt durch Bilden von Vektoren, deren Komponenten zeitlich verzögerte Ausschnitte der ursprünglichen Reihe $\{x_i\}_{i\in\{1,\dots,n\}}$ sind [Takens, 1981]:

$$\mathbf{p}_i = (x_i, x_{i+d}, x_{i+2d}, \dots, x_{i+(m-1)d})$$

Dabei stellt d die zeitliche Verzögerung (delay) und m die Dimension des entstehenden Phasenraums dar. Die Menge



(a) Pop

(b) Klassik

Abbildung 4: Die Abbildungen (a) und (b) zeigen das Ergebnis einer Phasenraumtransformation mit $d = 1$ und $m = 2$. Bei dem klassischen Stück aus Abbildung (b) ist deutlich die höhere Zirkularität zu beobachten.

$P_{d,m} = \{\mathbf{p}_i | i = 1, \ldots, n - (m-1)d\}$ bildet die Transformierte der Serie $\{x_i\}_{i \in \{1,\ldots,n\}}$ im Phasenraum. Man verläßt mit dieser Transformation den Raum der Zeit und betrachtet die Abhängigkeit der Werte zu den verschiedenen Zeitpunkten untereinander. Experimente zeigen, dass eine Transformation ohne Verzögerung, also mit $d = 1$, in einen 2-dimensionalen Raum eine gute Ausgangsbasis für die Extraktion weiterer Merkmale aus Musikdaten sind.

6.1 Grad der Zirkularität

Abbildung 4 zeigt das Ergebnis der Transformation eines typischen Popmusikstückes sowie eines klassischen Stückes in den 2-dimensionalen Phasenraum. Auf den ersten Blick fällt auf, dass das klassische Stück wesentlich rundere, harmonischere Kurven im Phasenraum beschreibt, während das populäre Lied zackigere Bahnen aufweist. Wir wollen diesen Unterschied durch zwei Merkmale messen, die den Grad der Zirkularität wiedergeben sollen.

Zunächst soll der durchschnittliche Winkel zwischen den Teilstücken bestimmt werden. Für jeden Punkt $\mathbf{p}_i$ existieren zwei Verbindungen zu anderen Knoten:

$$\begin{aligned} \mathbf{s}_{iv} &= \mathbf{p}_{i-1} - \mathbf{p}_i \\ \mathbf{s}_{in} &= \mathbf{p}_{i+1} - \mathbf{p}_i \end{aligned}$$

Die Kante $\mathbf{s}_{iv}$ verläuft von $\mathbf{p}_i$ zu seinem Vorgänger und die Kante $\mathbf{s}_{in}$ zu dem Nachfolger. Der zwischen $\mathbf{s}_{iv}$ und $\mathbf{s}_{in}$ eingeschlossene Winkel sei α. Abbildung 5 zeigt die verwendeten Vektoren. Mittels der geometrischen Interpretation des Skalarproduktes läßt sich α bestimmen durch

$$\alpha = \arccos \frac{\langle \mathbf{s}_{iv}, \mathbf{s}_{in} \rangle}{|\mathbf{s}_{iv}| \cdot |\mathbf{s}_{in}|}$$

Der Durchschnitt über die Winkel zwischen allen Teilstücken bildet somit ein weiteres Merkmal:

$$AVG_{PR-\alpha} = \frac{1}{|P_{d,m}| - 2} \sum_{i=2}^{|P_{d,m}|-1} \arccos \frac{\langle \mathbf{s}_{iv}, \mathbf{s}_{in} \rangle}{|\mathbf{s}_{iv}| \cdot |\mathbf{s}_{in}|}$$

Wie auch zuvor wird die Varianz der Winkel im Phasenraum ebenfalls als Merkmal ermittelt.

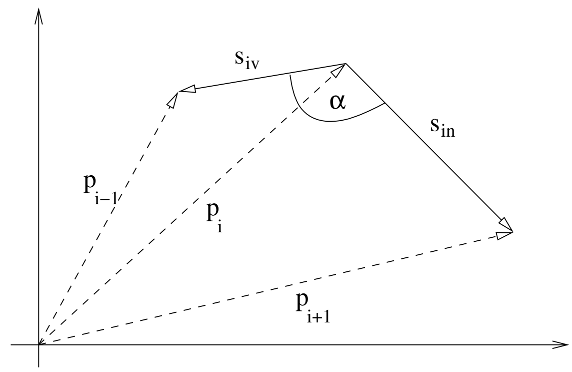


Abbildung 5: $\mathbf{p}_{i-1}$, $\mathbf{p}_i$ und $\mathbf{p}_{i+1}$ sind drei aufeinanderfolgende Vektoren im Phasenraum. Sie definieren die Teilstücke $\mathbf{s}_{iv}$ und $\mathbf{s}_{in}$, welche den Winkel α einschließen.

6.2 Durchschnittliche Länge

Bei der Betrachtung der Phasenraumdiagramme fällt außerdem auf, dass die Längen der Teilstücke zwischen den Vektoren unterschiedlich stark schwanken. Es ist daher sinnvoll, die durchschnittliche Länge der Teilstücke

$$AVG_{PR-L} = \frac{1}{|P_{d,m}| - 1} \sum_{i=2}^{|P_{d,m}|} |\mathbf{s}_{iv}|$$

und ihre Varianz ebenfalls als Merkmal aufzunehmen.

7 Anwendung auf reale Musikdaten

In den vorherigen Abschnitten wurden die folgenden Merkmale vorgestellt und beschrieben, wie sie aus Musikdaten $\{x_i\}_{i \in \{1,\ldots,n\}}$ extrahiert werden:

- Mittlere Lautstärke,
- durchschnittliche Extremadifferenz und Varianz,
- Tempo sowie die Varianz der Autokorrelation,
- k höchste Peaks nach Fouriertransformation,
- Durchschnitt und Varianz der stärksten Frequenzen im Zeitverlauf,

- Durchschnitt und Varianz der Winkel nach Phasenraumtransformation,
- Durchschnitt und Varianz der Abstände nach Phasenraumtransformation.

Da jeder Peak des Frequenzspektrums die drei Merkmale Höhe, Frequenz und Breite des Peaks erzeugt, ergeben sich insgesamt $11 + 3k$ Merkmale aus jedem Musikstück. In den folgenden Experimenten hat sich $k = 5$ bewährt, so dass insgesamt 26 Merkmale extrahiert wurden. Gemessen an den über 16 Milliarden Werten für ein Lied von drei Minuten ist somit ein großer Kompressionsfaktor erreicht worden.

7.1 Experimente

Die Implementierung der dargestellten Methoden erfolgte in einem generischen Framework zur Behandlung von Wertereihen wie in [Morik and Liedtke, 2000] gefordert. Die Experimente wurden mit Hilfe der Experimentierumgebung YALE [Fischer *et al.*, 2002] durchgeführt. Vor jedem Differenzfilter wurden die Daten mit einem Tiefpass gefiltert, vor jeder Fouriertransformation eine Hanning-Fensterfunktion angewendet. Vor der Extraktion der Merkmale wurde aus jedem Lied ein zufälliges Sample der Länge 60 Sekunden gewählt, die Extraktion dauerte auf einem Athlon 1600+ dann ca. 20 Sekunden.

Es wurden zwei Datensätze verwendet. Der Datensatz KLASSIK/POP besteht aus jeweils 100 Liedern aus den Genres Klassik und Pop. Der Datensatz POP/TECHNO besteht aus insgesamt 160 Musikstücken, jeweils 80 aus den Genres Pop und Techno. Beide Datensätze lagen als MP3-Dateien mit 128 kbits/s vor. Die Trennung des Datensatzes KLASSIK/POP wird als relativ einfache Klassifikationsaufgabe angesehen. Ungleich schwerer ist die Trennung von POP/TECHNO, da sich diese beiden Musikgattungen in vielen Punkten ähnlich sind.

7.2 Merkmalsselektion

Mittels eines genetischen Algorithmus [Ritthoff *et al.*, 2002] wurden die besten Merkmale selektiert. Bei dem Datensatz KLASSIK/POP wurden die durchschnittliche Extremadifferenz, die drei niedrigsten der Peaks sowie die Breite der beiden übrigen Peaks entfernt. Bei TECHNO/POP wurde stattdessen das Merkmal Tempo deselektiert sowie die Höhe der vier niedrigsten Peaks entfernt.

Die Anwendung eines einfachen 1-R-Lerners [Holte, 1993] lieferte für die beiden Datensätze unterschiedliche Merkmale. Im Fall KLASSIK/POP wurde die Varianz der Abstände nach einer Phasenraumtransformation als das beste trennende Merkmal bestimmt. Die Anwendung der einfachen Regel klassifizierte bereits 184 der 200 Instanzen korrekt. In der Domäne POP/TECHNO wurde hingegen die Varianz der Extremadifferenz als Merkmal gewählt, was bereits 121 der 160 Instanzen korrekt klassifizierte.

Es existiert also kein Merkmal, das für beide Datensätze entfernt wurde. Insgesamt wird deutlich, dass für beide Klassifikationsaufgaben die Merkmale einen unterschiedlichen Stellenwert haben. Daher sollte eine Merkmalsselektion aus dem Satz aller Merkmale vor der Hypothesenbildung erfolgen.

7.3 Ergebnisse

Als Lernverfahren kamen k-nearest-neighbors (k-NN), Naive Bayes, C4.5 [Quinlan, 1993] sowie eine Support Vector Machine (SVM) [Joachims, 1999; Rüping, 2000] zum Einsatz. Die Performanzmaße Accuracy (acc), Precision (pre) und Recall (rec) wurden anhand einer 10-fachen Kreuzvalidierung evaluiert.

		KLASSIK/POP	**POP/TECHNO**
C4.5	acc	$98,33\% \pm 3,08$	$86,87\% \pm 7,20$
	pre	$97,51\% \pm 5,47$	$85,25\% \pm 9,61$
	rec	$98,46\% \pm 4,29$	$93,32\% \pm 6,70$
SVM	acc	$95,38\% \pm 5,10$	$80,78\% \pm 9,94$
	pre	$99,55\% \pm 2,75$	$82,61\% \pm 11,03$
	rec	$92,19\% \pm 8,05$	$78,95\% \pm 12,74$
k-NN	acc	$96,86\% \pm 5,16$	$81,89\% \pm 5,86$
	pre	$98,39\% \pm 5,00$	$83,90\% \pm 9,42$
	rec	$94,31\% \pm 10,43$	$81,57\% \pm 10,38$
Bayes	acc	$96,92\% \pm 5,10$	$75,89\% \pm 10,34$
	pre	$97,57\% \pm 4,29$	$73,67\% \pm 9,99$
	rec	$95,31\% \pm 9,90$	$84,50\% \pm 10,89$

Tabelle 1: Ergebnisse.

Tabelle 1 zeigt die Ergebnisse der verschiedenen Lernläufe. Auffällig ist, dass die Wahl des Lernverfahrens kaum Einfluß auf die Güte des Ergebnis hat. Offensichtlich ist es für alle Verfahren etwa gleich schwierig, die Datensätze anhand der gegebenen Merkmale zu trennen.

Bei der Klassifikation des Datensatzes KLASSIK/POP wurden für Accuracy und Recall Werte von über 98% (C4.5) erreicht, die beste Precision von $99,55\%$ brachte eine Support Vector Machine mit linearem Kernel. Es stellte sich heraus, dass eine Erhöhung der Qualität auf CD-Niveau keine weitere Verbesserung der Ergebnisse mit sich bringt, eine kleine Verschlechterung ist erst mit einer schlechteren Kodierung als 96 kbits/s zu messen. Eine genaue Untersuchung der Lernergebnisse bei verschiedenen *sample rates* und Kompressionstechniken steht noch aus.

8 Zusammenfassung und Ausblick

Ein Satz von Merkmalen aus Audiodaten wurde beschrieben und ihre Extraktion erläutert. Dabei wurden Techniken der Zeitreihenanalyse benutzt, insbesondere die Transformation in den Phasenraum erlaubt die Generierung neuartiger Merkmale. Diese sind invariant gegenüber einer zeitlichen Verschiebung der Musikstücke. Es hat sich herausgestellt, dass die Merkmale robust genug sind, um auch mit Rauschen im für den menschlichen Hörer vertretbaren Maße umzugehen. In einem zweiten Schritt wurde für jeden Datensatz getrennt eine Merkmalsselektion durchgeführt.

Die Extraktion der vorgestellten Merkmale erhöht die Performanz für die jeweiligen Klassifikationsaufgaben. Accuracy, Precision und Recall liegen für den Datensatz KLASSIK/POP nah bei 100%, für POP/TECHNO bei 85% bis 93%.

Die besten Merkmale für die beiden Klassifikationsaufgaben sind Varianzen von Größen in verschiedenen Dimensionen. Im Fall KLASSIK/POP handelte es sich um die Varianz der Abstände nach einer Phasenraumtransformation und für POP/TECHNO wurde die Varianz der Extremadifferenz als herausragendes Merkmal gewählt. Die Varianz spiegelt die Veränderlichkeit der Daten über die Zeit wider, ohne selbst eine Abhängigkeit zu dem genauen Zeitpunkt des Auftretens der Änderungen zu besitzen.

In Zukunft soll anhand von Playlists die Präferenz eines Benutzers gelernt werden. Das System kann dem Benutzer dann Vorschläge für ihm unbekannte Musik machen. Der Einsatz einer speziellen SVM Variante für Rankings [Joachims, 2002] könnte die Performanz weiter steigern. Desweiteren ist eine automatisierte Auswahl der optimalen Vorverarbeitungsschritte denkbar, je nach bisheriger Vorverarbeitung wird die nächste Methode ausgewählt. Zu guter Letzt gibt es erste Ansätze, Intervalle in den verschiedenen Dimensionen zu bestimmen und aus den Intervallen weitere Merkmale zu generieren.

9 Danksagung

Dieser Artikel stellt Ergebnisse meiner Diplomarbeit am Lehrstuhl für künstliche Intelligenz der Universität Dortmund vor. Ich danke Prof. Dr. Katharina Morik und Dipl.-Inform. Michael Wurst für die ausgezeichnete Betreuung sowie Dipl.-Inform. Ralf Klinkenberg für zahlreiche Erklärungen und Ermutigungen.

Literatur

[Brillinger and Irizarry, 1998] R. Brillinger and R. A. Irizarry. An investigation of the second- and higher- order spectra of music. *Signal Processing*, 65:161–179, 1998.

[Cemgil *et al.*, 2000] A. Cemgil, B. Kappen, P. Desain, and H. Honing. On tempo tracking: Tempogram representation and Kalman filtering. In *Proceedings of the International Computer Music Conference*, pages 352–355, 2000.

[Dixon, 2001] Simon Dixon. An interactive beat tracking and visualisation system. In *Proceedings of the International Computer Music Conference*, pages 215–218, 2001.

[Fischer *et al.*, 2002] Simon Fischer, Ralf Klinkenberg, Ingo Mierswa, and Oliver Ritthoff. YALE: Yet Another Learning Environment. Technical Report CI-136/02, Universität Dortmund, Lehrstuhl Informatik VIII, Juni 2002.

[Holte, 1993] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90, 1993.

[Irizarry, 2001] Rafael A. Irizarry. Local Harmonic Estimation in Musical Sound Signals. *Journal of the American Statistical Association*, 96(454), 2001.

[Joachims, 1999] Thorsten Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT Press, 1999.

[Joachims, 2002] Thorsten Joachims. Evaluating Retrieval Performance Using Clickthrough Data. In *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, 2002.

[Liu and Motoda, 1998] H. Liu and H. Motoda. *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer, 1998.

[Loy, 1989] G. Loy. Musicians make a standard: the MIDI phenomenon. *Computer Music Journal*, 9(4), 1989.

[Mitchell, 1996] M. Tom Mitchell. *Machine Learning*. McGraw Hill, New York, USA, 1996.

[Morik and Liedtke, 2000] Katharina Morik and Harald Liedtke. Learning about time. Technical report, Univerität Dortmund, Lehrstuhl Informatik VIII, 2000. MiningMart Deliverable No. 3.

[Morik, 2000] Katharina Morik. The Representation Race – Preprocessing for Handling Time Phenomena. In *Proceedings of ECML 2000*, 2000.

[Pickens, 1996] Jeremy Pickens. A survey of feature selection techniques for music information retrieval. Technical report, Center of Intelligent Information Retrieval, Department of Computer Science, University of Masschusetts, 1996.

[Pyle, 1999] Dorian Pyle. *Data Preparation for Data Mining*, chapter Series Variables. Morgan Kaufmann, 1999.

[Quinlan, 1993] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Diego, CA, USA, 1993.

[Ribbrock and Kurth, 2002] A. Ribbrock and F. Kurth. A full-text retrieval approach to content-based audio identification. In *International Workshop on Multimedia Signal Processing*, 2002.

[Ritthoff *et al.*, 2002] Oliver Ritthoff, Ralf Klinkenberg, Simon Fischer, and Ingo Mierswa. A Hybrid Approach to Feature Selection and Generation Using an Evolutionary Algorithm. Technical Report CI-127/02, Universität Dortmund, Lehrstuhl Informatik VIII, Februar 2002.

[Roads, 1996] Curtis Roads. *The Computer Music Tutorial*. MIT Press, 1996.

[Rüping, 2000] Stefan Rüping. *mySVM - Manual*. Universität Dortmund, Lehrstuhl Informatik VIII, Oktober 2000.

[Schlittgen and Streitberg, 1997] Rainer Schlittgen and Bernd H.J. Streitberg. *Zeitreihenanalyse*. Oldenbourg Verlag München, 1997.

[Takens, 1981] F. Takens. Detecting strange attractors in fluid turbulence. *Dynamical Systems and Turbulence*, 1981.

[Tipler, 1991] Paul A. Tipler. *Physics for Scientists and Engineers*. Worth Publishers, 1991.

[Tzanetakis *et al.*, 2001] George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals. In *In Proceedings of the Int. Symposium on Music Information Retrieval (ISMIR)*, pages 205–210, 2001.

[Tzanetakis, 2002] George Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD thesis, Computer Science Department, Princeton University, June 2002.

[Witten and Frank, 2000] Ian H. Witten and Eibe Frank. *Data Mining*. Morgan Kaufmann, San Diego, CA, USA, 2000.