VERGLEICH STATISTISCHER KLASSIFIKATOREN ZUR ERMITTLUNG MUSIKALISCHER ASPEKTE

Stephan Hübler, Matthias Wolff*, Matthias Eichner***
**TU Dresden - Institut für Akustik und Sprachkommunikation*
*** mufin GmbH*
stephan.huebler@ias.et.tu-dresden.de

Kurzfassung: Menschen beschreiben Musik mit einfachen, signalfernen, bedeutungsvollen Aspekten. Deren automatische Gewinnung aus dem Musiksignal bildet die Grundlage für eine Annotation von Musiktiteln und damit einer Vielzahl von weiteren Anwendungen unter anderem Musikempfehlungssysteme, Navigieren durch Musiksammlungen sowie deren Visualisierung und das Bilden von Kaufempfehlungen. Gegenstand der Untersuchung sind sechs Aspekte (Instrument Density, Music Color, Percussiveness, Sing Detection, Style, Tempo) mit jeweils zwei bis zehn Ausprägungen. Aus statistischer Sicht werden Merkmalvektoren, hinter denen sich Ausprägungen eines Aspektes verbergen, beobachtet. Dabei handelt es sich um verschiedene Low- und Midlevel-Merkmale. Da Musiksignale Informationen in ihrer zeitlichen Struktur tragen, wird insbesondere der zeitliche Verlauf in die Untersuchung mit einbezogen. Zunächst wird eine Klassifikationsentscheidung pro Merkmalvektor eines Musiktitels getroffen, welche dann zu einer Gesamtentscheidung bezüglich der Ausprägung eines Aspektes für den Song führt.

Es werden die Ergebnisse für die Klassifikation von perzeptuellen musikalischen Aspekten eines Musikstückes mittels statistischer GMM und HMM basierender Klassifikatoren präsentiert.

1 Einleitung

Ein Teilgebiet der Musikanalyse ist die Klassifikation von Musiksignalen mit bedeutungsvollen musikalischen Aspekten [5]. Sie sind Ausgangspunkt vielfältiger Verarbeitungsmöglichkeiten wie semantische Suche nach Audiomaterial oder der Ähnlichkeitsgenerierung von Musiktiteln. Es gilt die Entscheidung des Menschen, welcher ein Musikstück mit Aspekten wie den hier verwendeten beschreibt, nachzubilden. Die Beurteilung von Musik und damit auch die Annotation von Musik ist subjektiv, was in unterschiedlichem Musikgeschmack begründet liegt [4]. Dennoch ist es möglich, eine Schnittmenge in der Beurteilung zu finden. Ziel ist die automatische Einordnung von Musikstücken bezüglich der möglichen Aspektausprägung.

Die vorliegende Studie ist eingebettet in das Empfehlungssystem AudioGen der Firma mufin[1] und wurde an der TU Dresden mithilfe des hauseigenen Experimentiersystems dlabpro durchgeführt. Schwerpunkt sind die verschiedenen Möglichkeiten der Klassifikation durch die Gruppe der Markoverkenner. Die anderen Elemente der Aspektgewinnung, von der Auswahl der Signale, Hörerurteile, Merkmalextraktion, Sekundäranalyse bis hin zur Entscheidungsfunktion und Auswertung, entsprechen dem bestehenden AudioGen System [1] und sind nicht Gegenstand der Betrachtung. Die Klassifikationsergebnisse beziehen sich ausschließlich auf die hier

[1] www.mufin.com

verwendeten Klassifikatoren, welche nicht Bestandteil des AudioGen-Systems sind. Musik besitzt eine zeitliche Struktur durch die Anordnung von Tönen und Klängen über der Zeit. Deswegen ist insbesondere die Verwendung von Hidden-Markov-Modellen, wie sie in der Sprachverarbeitung eingesetzt werden, Ausgangspunkt einer möglichen Verbesserung der Klassifikation.

2 Aspekte

Gegenstand der Untersuchung sind die sechs Aspekte Instrument Density, Music Color, Percussiveness, Tempo, Sing Detect und Style (Tabelle 1). Bei den ersten vier handelt es sich um perzeptuelle Beschreibungen, daher um die Wahrnehmung durch den Menschen.

- Unter Instrument Density ist die perzeptuelle Wahrnehmung der Instrumentendichte zu verstehen, daher ob viele (full) oder wenige (sparse) Instrumente innerhalb des Musikstückes gleichzeitig zum Einsatz kommen.
- Music Color beschreibt die wahrgenommene Klangfarbe mit dunkel (dark) oder hell (bright).
- Ob ein Musikstück als perkusiv (percussive) oder nicht (nonpercussive) wahrgenommen wird, hängt vom Einsatz von Rhythmusinstrumenten wie Schlagzeug, Percussions oder ähnlichem ab.
- Jeder Titel wird von dem Hörer in eine der drei Tempokategorien schnell (fast), mittel (midtempo) oder slow (langsam) eingeordnet. Die Grenzen sind nicht BPM-basierend, sondern entsprechen der Wahrnehmung durch den Menschen.
- Sing Detect macht eine Aussage, ob in dem Musikstück gesungen wird oder nicht.
- Es kommen zehn Genreausprägungen (Styles) zum Einsatz, wobei jedes Musikstück maximal einem zugeordnet wird.

3 Aspektgewinnung

Ausgangspunkt für die Gewinnung von Aspekten ist ein Pool von Musiksignalen mit der dazugehörigen Beschriftung der Aspektausprägungen, welche von einer Hörergruppe erstellt wurde. Anschließend werden Merkmale aus dem Signal extrahiert, eine sekundäre Merkmalanalyse durchgeführt und mittels einer Entscheidungsfunktion jedem Titel eine Aspektausprägung zugeordnet. Zur Auswertung stehen dann Verwechslungsmatrizen zur Verfügung. Die Aspektgewinnung ist in Abbildung 1 dargestellt. Gegenstand der vorliegenden Untersuchung sind die verschiedenen Möglichkeiten der Klassifikation. Es wird zunächst ein kurzer Überblick über die anderen Elemente der Aspektgewinnung gegeben.

3.1 Signale und Hörerurteile

Der verwendete Datensatz der Firma mufin umfasst 836 Musikdateien aus zehn unterschiedlichen Genres mit jeweils weiteren Subgenres. Es handelt sich um einen umfassenden Querschnitt der existierenden Musik von Klassik, Pop, Schlager, Jazz bis zu Hip Hop und Techno. Jedes Lied wurde von einer Hörergruppe mit jeweils einer Ausprägung pro Aspekt annotiert. Dabei gab es auch immer die Möglichkeit, einem Stück keine der zur Verfügung stehenden Aspektausprägungen zuzuordnen, wenn sich die Hörer nicht eindeutig entscheiden konnten.

Aspekt	Ausprägungen
Instrument Density	Sparse
	Full
Music Color	Bright
	Dark
Percussiveness	Percussive
	Nonpercussive
Tempo	Fast
	Midtempo
	Slow
Sing Detect	Singer
	No Singer
Style	Classical
	Country
	Dance
	Electronica
	Jazz
	Latin
	Rap
	Rock
	Soul
	Speech

Tabelle 1 - Übersicht der Aspektausprägungen der verwendeten Aspekte

Auf dieser Grundlage können Unterdatensätze für das Training des jeweiligen Aspektes gezogen werden. Um eine statistisch belastbare Aussage der Klassifikationsergebnisse zu erhalten, sind möglichst viele Daten nötig. Daher wurden die Klassifikatoren mit einer fünffachen Kreuzvalidierung ausgewertet, wo jeweils 20% der Songs das Testset stellen.

3.2 Primär- und Sekundäranalysator

Jeder Aspekt beruht auf einer eigenen Zusammenstellung von Audiomerkmalen aus der Tabelle 2. Der Pool von möglichen Features wurde vom Fraunhofer-Institut mit der Überlegung erstellt, ein Musikstück möglichst mit unterschiedlichen Sichtweisen aus den Bereichen Klang und Rhythmus zu repräsentieren [1]. Die Merkmale werden außerdem in [3] ausführlich erläutert. Als Sekundäranalysator werden jeweils die n ersten LDA-transformierten Vektorelemente verwendet.

3.3 Entscheidungsfunktion und Auswertung

Die Unterscheidungsfunktion des Klassifikators erbringt eine Klassifikationsentscheidung pro Merkmalvektor. Der gesamte Titel wird der Klasse (Aspektausprägung) zugeordnet, welcher die Mehrheit der Merkmalvektoren angehören. Die dazugehörige Entscheidungsfunktion ist in Formel 1 mit s ... Entscheidung, c ... Klasse, k ... Merkmalvektorindex, K ... Anzahl Merkmalvektoren (pro Song), s_k ... Klasse eines Merkmalvektors und δ ... KRONECKER-Symbol gegeben.

$$s = arg\max_{c} \sum_{k=1}^{K} \delta(s_k, c) \tag{1}$$

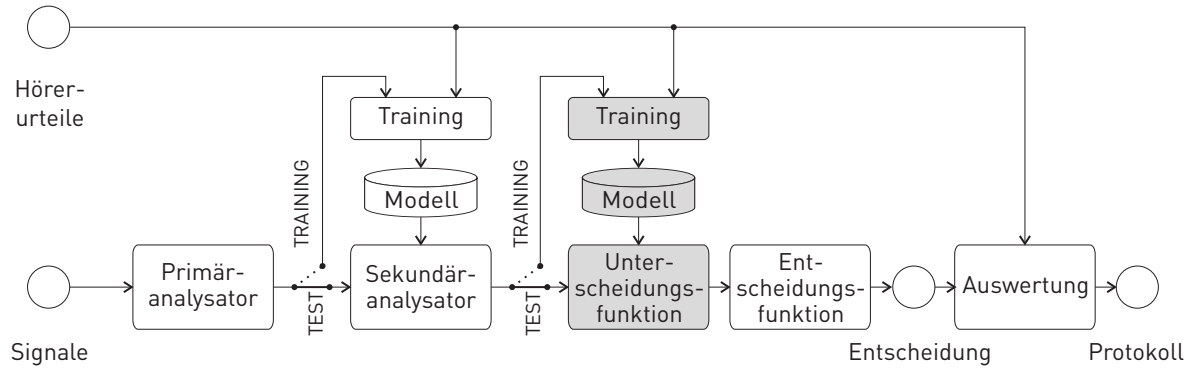

Abbildung 1 - Übersichtsdarstellung der Aspektgewinnung - Gegenstand der Untersuchungen ist der grau hinterlegte Klassifikator

Nach Vorliegen der Klassifikationsentscheidung aller Titel im Testset ist es möglich, eine Verwechslungsmatrix nach Formel 2 anzugeben. Dabei gibt $H(s,c)$ die absolute Häufigkeit an, wie oft für Klasse s entschieden wurde, wenn Klasse c vorliegt.

$$H = \Big(H(s,c)\Big) = \begin{matrix} \begin{pmatrix} H(1,1) & H(1,2) & \cdots & H(1,C) \\ H(2,1) & H(2,2) & \cdots & H(2,C) \\ \vdots & \vdots & \ddots & \vdots \\ H(C,1) & H(C,2) & \cdots & H(C,C) \end{pmatrix} & \downarrow s \\ \rightarrow c & \end{matrix} \tag{2}$$

Die Auswertung der Verwechslungsmatrix erfolgt durch die Angabe der korrekten Klassifikationsentscheidungen (correctness) nach Formel 3. Die Correctness ist ein vergleichbarer Wert bezüglich der Qualität der jeweiligen Aspektklassifikatoren. Außerdem wird das 95%-Konfidenzintervall angegeben.

$$Cor = \frac{\sum_{c=1}^{C} H(c,c)}{\sum_{c=1}^{C} \sum_{s=1}^{C} H(s,c)} \tag{3}$$

4 Experimente

Die Reihenfolge der Experimente geht mit der Komplexität der Modelle einher. So werden zunächst Gaussian Mixture Models bezüglich ihrer Eignung für die vorliegende Klassifikationsaufgabe getestet. Sie ermöglichen es, jedem Merkmalvektor eine andere Aspektausprägung zuzuordnen. Als nächste Stufe werden Hidden-Markov-Modelle mit einer festen Links-Rechts-Struktur wie in der Sprachverarbeitung getestet. Ausgangspunkt sind dafür 3 bzw. 10 Zustände, was dazu führt, dass im Testmaterial mindestens 3 bzw. 10 aufeinanderfolgende Merkmalvektoren einer Klasse zugeordnet werden müssen. Im letzten Experiment, dem Topologietraining, erhält das Modell vollkommene Freiheit, die zeitliche Struktur einer Aspektausprägung über den ganzen Song abzubilden.

4.1 Gausian Mixture Model (GMM)

Mithilfe der Merkmalvektoren, welche einer Klasse zugeordnet sind, werden zunächst die Parameter einer mehrdimensionalen Normalverteilung geschätzt. Es entsteht ein Gaussian Mixture

Merkmal	Dimension
Klang	
Nulldurchgänge	1
spezifische Lautheit Log	12
spezifische Lautheit Norm	12
Mel Frequency Cepstrum Coefficients	16
Audio Spectrum Flatness Measure	16
Audio Spectrum Crest Factor	16
Audio Spectrum Centroid	12
Rhythmus	
Ausschnitt der Autokorrelationsfunktion	70
Statistiken über Autokorrelationsfunktion	6
Onset Density	19
Percussiveness	19
Modulation	
ZRC Modulation	12
MFCC Modulation	16
	16
SFM Modulation	12
	12

Tabelle 2 - Eine Auswahl der im AudioGen-System verwendeten Signaleigenschaften mit der jeweiligen Dimension

Model durch das Teilen und Verschieben der Normalverteilung bezüglich der Richtung ihrer größten Varianz. Die Parameter der neu entstandenen Normalverteilungen werden durch den EM-Algorithmus neu berechnet. Dieser Vorgang der Modellbildung wird wiederholt durchgeführt. In Abbildung 2 sind die GMMs für den Aspekt Instrument Density für die zwei möglichen Klassen in Form eines HMM dargestellt. Jeder Kante ist eine Normalverteilung zugeordnet, alle Kanten zusammen entsprechen der Realisierung des Gaussian Mixture Modells.

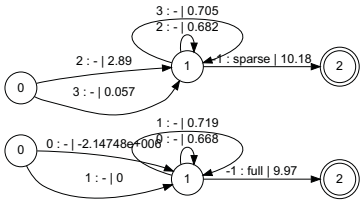

Abbildung 2 - Darstellung der GMMs für den Aspekt Instrument Density nach einer Teilung der Normalverteilungen. Die Kanten sind mit Eingabesymbol (Nummer der Normalverteilung), Ausgabesymbol und Übergangswahrscheinlichkeit beschriftet.

4.2 Hidden-Markov-Model (HMM) - Links-Rechts-Struktur

Die GMMs betrachten jeden Merkmalvektor einzeln und ordnen ihn einer der möglichen Klassen zu. Im Gegensatz dazu ist es mit HMM möglich, die zeitliche Abfolge der Merkmalvektoren mit in die Klassifikationsentscheidung für jeden Frame einzubeziehen. Zunächst wird eine fes-

te Struktur der HMMs vorgegeben. Bei der Verwendung von drei Zuständen (HMM-3), wie in Abbildung 3 dargestellt, werden mindestens drei aufeinanderfolgende Merkmalvektoren der gleichen Klasse zugeordnet. Es können aber auch wesentlich mehr aufeinanderfolgende Merkmalvektoren dieser Klasse zugeordnet sein, da es möglich ist, innerhalb des HMMs beliebig lange in einem Zustand zu bleiben. Für die Initialisierung dieser Modelle wurde jeweils der komplette Trainingstitel in 3 Teile aufgeteilt und den entsprechenden Kanten zugeordnet. Jede Kante besitzt danach eine mehrdimensionale Gaußverteilung. Für die Adaptierung des Modells an das Trainingsmaterial werden auch hier die Normalverteilungen geteilt und deren Parameter durch den EM-Algorithmus neu berechnet - es entstehen auf diese Weise neue Kanten. Das Training der HMM erfolgt mit Viterbisuche.

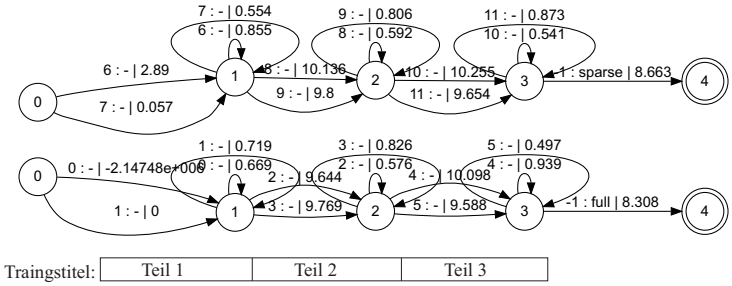

Abbildung 3 - Darstellung der HMM mit 3 Zuständen für den Aspekt Instrument Density nach einer Teilung der Normalverteilungen. Für die Initialisierung werden die Traingsdaten in 3 Teilen auf die Kanten aufgeteilt. Die Kanten sind mit Eingabesymbol (Nummer der Normalverteilung), Ausgabesymbol und Übergangswahrscheinlichkeit beschriftet.

Bei dem Experiment HMM-10 wird ein Modell mit 10 initialen Zuständen verwendet und dementsprechend das Trainingsmaterial auf 10 Zustände aufgeteilt.

4.3 Hidden-Markov-Model (HMM) - Topologietraining

Die Links-Rechts-Struktur aus Experiment 2, wie sie in der Spracherkennung Anwendung findet, ist wahrscheinlich nicht die adäquate Abbildung der zeitlichen Struktur eines Musikstückes. Möglicherweise ergeben sich über den Song typische Folgen von Merkmalvektoren, welche eine Aspektausprägung charakterisiert. Dabei wird insbesondere die Möglichkeit zugelassen, nicht nur aufeinanderfolgende Merkmalvektoren der gleichen Normalverteilung zuzuordnen, sondern auch der gleichen Folge von Normalverteilungen. Diese Schleifen innerhalb des Modells gehen dann über mehrere Zustände. Ein auf diese Weise trainiertes Modell ist in Abbildung 4 dargestellt. Da über die zeitliche Struktur der Aspektausprägung kein Vorwissen vorhanden ist, wird ein automatisches Topologietraining verwendet.

Ausgehend von einem Zustand wird ein Trainingsfahrplan mit den zwei Hauptverfahren „Teilen“ und „Pruning“ durchgeführt. Beim „Teilen“ wird die Hälfte der Normalverteilungen (Kanten) geteilt und für diese ein neuer Zustand als Zielknoten eingeführt. Dafür werden die Normalverteilungen mit der größten Varianz gewählt. Auf diese Weise erhält das Modell die Freiheit, neue Wege zu bilden. Beim „Pruning“ werden die Kanten entfernt, welche bei der Viterbidekodierung der Trainingsdaten nicht oder wenig benutzt werden. Es werden nur Kanten entfernt, welche das Modell nicht zerstören. Eine Zerstörung liegt vor, wenn es keinen durchgehenden Weg vom Start- zum Endknoten gibt. Die Verfahren zum Topologietraining werden in [2] näher erläutert.

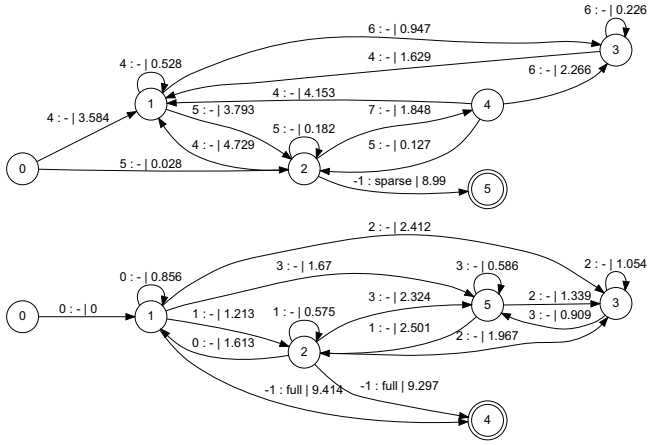

Abbildung 4 - Darstellung des HMM nach dem Topologietraining für den Aspekt Instrument Density. Die Kanten sind mit Eingabesymbol (Nummer der Normalverteilung), Ausgabesymbol und Übergangswahrscheinlichkeit beschriftet.

5 Ergebnisse und Schlussfolgerungen

Die Erkennungsraten nach Formel 3 mit dem dazugehörigen Vertrauensbereich sind in Tabelle 3 angegeben. Die Mustererkennung in Musikstücken ist mittels der hier verwendeten Markov-Modelle prinzipiell möglich. Es kann sich kein Klassifikator als alleiniger Gewinner für alle Aspekte durchsetzen. In der Tabelle 3 sind alle Klassifikatoren fett gedruckt, deren Correctness oberhalb des unteren Vertrauensbereiches des jeweils besten Klassifikators pro Aspekt liegt. So setzt sich lediglich für den Aspekt Instrument Density das Topologiemodell alleine durch, für alle anderen Aspekte kommen mindestens zwei Klassifikatoren in Frage. Alle getesteten Klassifikatoren bringen bei dem Aspekt Perussivness die gleich hohe Erkennungsrate. Das Klassifikationsproblem scheint sehr einfach zu sein, da alle verwendeten Klassifikatoren zu einer sehr hohen Erkennungsrate von $95,8 \pm 2,2$ kommen. Die perzeptuelle Tempoerkennung eines Stückes funktioniert mit den hier verwendeten Parametern der Aspektgewinnung für alle Modelle mit höchstens $62,4 \pm 3,4$ Erkennungsrate schlecht. Es zeigt sich, dass die Wahl des richtigen Modelles aspektabhängig ist.

Aspekt	**Kls.**	**Dat.**	**GMM**	**HMM-3**	**HMM-10**	**TOPO**
InstrDensity	2	219	75,7±5,7	72,6±5,9	82,6±5,0	**87,2±4,4**
MusicColor	2	373	78,8±4,1	**82,5±3,8**	**85,2±3,6**	80,4±4,0
Perc	3	315	**95,8±2,2**	**95,8±2,2**	**95,8±2,2**	**95,8±2,2**
SingDetect	2	745	**89,1±2,2**	**89,1±2,2**	**90,7±2,0**	86,9±2,4
Style	10	541	**77,5±5,6**	**77,9±7,9**	**79,1±3,4**	73,0±3,7
Tempo	3	741	58,8±3,5	**59,6±3,5**	**62,4±3,4**	**60,1±3,5**

Tabelle 3 - Correctness der verschiedenen Experimente mit Vertrauensbereich; die getesteten Klassifikatoren sind nicht Bestandteil des AudioGen-Systems; **Fett:** Klassifikatoren deren Correctness oberhalb des unteren Vertrauensbereiches des besten Klassifikators pro Aspekt liegt

Für den Einsatz in der Praxis ist die Größe der Modelle, daher die Anzahl der Parameter, von entscheidender Bedeutung. Die Parameteranzahl hängt unter anderem von der Anzahl der Klassen, der Größe der Merkmalvektoren und von der Anzahl an Zuständen und Kanten im Modell

ab. Diesbezüglich unterscheiden sich die entwickelten Modelle zum Teil stark. So wird durch das Topologietraining in den meisten Fällen die minimale Anzahl von Parametern erreicht.

Aspekt	Kls.	Dat.	GMM	HMM-3	HMM-10	TOPO
InstrDensity	2	219	354	2114	3522	**102**
MusicColor	2	373	1410	**266**	**442**	170
Perc	3	315	**706**	**68**	**222**	109
SingDetect	2	745	**11266**	**8450**	**1762**	137
Style	10	541	**235530**	**176650**	**147210**	1674
Tempo	3	741	1539	**9219**	**15363**	**150**

Tabelle 4 - Parameteranzahl pro Modell; **Fett:** Klassifikatoren deren Correctness oberhalb des unteren Vertrauensbereiches des besten Klassifikators pro Aspekt liegt; vgl. Tabelle 3

Mit den durchgeführten Experimenten wurden die Aspekte auf eine mögliche zeitliche Struktur bezüglich des gesamten Songs untersucht. Bei Instrument Density ergibt sich eine deutliche Verbesserung durch die Einbeziehung einer automatischen Strukturaufdeckung. Teilweise funktioniert, die Links-Rechts-Struktur eins HMM besser als GMM. Hier wurde der gesamte Trainingstitel auf das Modell zeitlich abgebildet. Im Testsong mussten dann mindestens drei (HMM-3) oder zehn (HMM-10) aufeinanderfolgende Merkmalvektoren einer Ausprägung zugeordnet werden. In weiteren Untersuchungen soll der Titel in Segmente wie Takt oder Strophe/Refrain unterteilt werden, da sich zeitliche Strukturen der hier verwendeten Aspekte wohl nicht über den ganzen Song erstrecken. Mögliches Verbesserungspotential liegt in den anderen Teilen der Aspektgewinnungskette aus Abbildung 1, welche nicht Bestandteil dieser Studie waren. Es konnte gezeigt werden, dass der Einsatz von Hidden-Markov-Modellen die Klassifikation für einige Aspekte verbessert.

Literatur

[1] BASTUCK, C.: *An Extensible and Multiperspective Approach for Music Similarity*. In: *MIREX 2007*, Sep 2007. Research Institut: Frauenhofer IDMT.

[2] DUCKHORN, F.: *Optimierung von Hidden-Markov-Modellen*. Diplomarbeit, TU Dresden, Institut für Akustik und Sprachkommunikation, Sep. 2007.

[3] HUEBLER, S.: *Suchraumoptimierung zur Identifizierung ähnlicher Musikstücke*. Diplomarbeit, TU Dresden, Institut für Akustik und Sprachkommunikation, Dez. 2008.

[4] LEVITIN, D.: *This is Your Brain on Music - Understanding a Human Obsession*. Atlantic Books, 2008.

[5] SKOWRONEK, J., M. MCKINNEY, S. VAN DE PAR und J. BREEBAART: *Anwendungsbeispiele für Musikanalyse-Algorithmen*. In: *Fortschritte der Akustik - DAGA 2008*, Bd. 34, S. 543–544, 2008. Research Institut: Philips Research Laboratories Eindhoven.