Automatisierte Extraktion rhythmischer Merkmale zur Anwendung in Music Information Retrieval-Systemen

Dissertation zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau,

von Dipl.-Ing. Christian Uhle

Ilmenau, 29.11.2005

Betreuer:	Prof. Dr.-Ing. Karlheinz Brandenburg
Gutachter:	Prof. Dr.-Ing. Jürgen Wernstedt
	Dr.-Ing. Jürgen Herre

urn:nbn:de:gbv:ilm1-2007000402

Kurzfassung

Das Thema dieser Dissertation ist die Extraktion von Merkmalen, die rhythmische Eigenschaften von Audiosignalen beschreiben. Diese Merkmale sind für die Anwendung in Music Information Retrieval (MIR)-Systemen ausgewählt.

Obwohl in der Vergangenheit an der Extraktion rhythmischer Merkmale wie zum Beispiel Tempo und Taktart in großem Umfang gearbeitet wurde, erreichen aktuelle Verfahren nicht die Erkennungsleistung eines geübten Zuhörers. Eine der Ursache dafür wird in der Auswertung von Informationen auf unterschiedlichen Abstraktionsebenen beim Menschen vermutet, eine weitere bei der Berücksichtigung von musikalischem Vorwissen. Der hier beschriebene Ansatz orientiert sich an diesen Analysemechanismen.

Zur Identifikation von geeigneten Merkmalen und relevanten Aspekten der menschlichen Verarbeitung der Schallsignale werden Grundlagen aus Musiktheorie, Psychoakustik und Kognitionswissenschaft erklärt. Bekannte Verfahren zur Extraktion rhythmischer Merkmale werden in einer ausführlichen Darstellung des Standes der Technik anschließend erläutert.

Der Hauptteil der Arbeit enthält eine Zusammenstellung von Verfahren des maschinellen Hörens, die Informationen auf unterschiedlichen Abstraktionsebenen auswerten. Eine kompakte Darstellung der metrischen Struktur wird zur Ermittlung der metrischen Merkmale vorgestellt. Da einerseits die Auswertung von Low-level-Merkmalen die Anwendung von musikalischem Vorwissen nur in geringen Maß ermöglicht, und andererseits die Informationen auf höheren Abstraktionsebenen durch ihre Fehlerhaftigkeit die Erkennungsleistung in verschiedenen Situationen einschränken können, werden die Ergebnisse der verschiedenen Verfahren in Abhängigkeit ihrer Konfidenzmaße zu einem Gesamtergebnis zusammengefasst.

Die Extraktion von rhythmischen Merkmalen aus den Informationen maschinell detektierter perkussiver Instrumente stellt einen Fortschritt im Vergleich zu bekannten Arbeiten dar. Eine Segmentierung in charakteristische Abschnitte des Audiosignals, die zum Beispiel Strophe oder Refrain repräsentieren, wird als Vorverarbeitungsschritt zur Analyse vorgestellt und die dadurch erreichte signifikante Verbesserung der Erkennungsleistung nachgewiesen.

Die Leistungsfähigkeit der Verfahren wird anhand eines umfangreichen Testdatensatzes evaluiert und die Eignung der extrahierten Merkmale in einem MIR-System untersucht.

Abstract

This thesis describes the automated extraction of features for the description of the rhythmic content of musical audio signals. These features are selected with respect to their applicability in music information retrieval (MIR) systems.

While research on automatic extraction of rhythmic features, for example tempo and time signature has been in progress for some time, current algorithms still seem to be a long way from matching human recognition performance. Among the reasons of the difference between the performances of a machine listening system and a trained listener are the use of information on different levels of abstraction and musical knowledge in human cognition. The approach described here is influenced by these two principles of cognition.

In order to identify appropriate features and relevant aspects of human processing of audio signals the necessary knowledge of musicology, psychoacoustics and cognition science are described. Subsequently, the description of the state-of-the-art comprises known methods for the extraction of rhythmic features from musical audio signals.

The main part of the thesis contains a collection of machine-listening methods evaluating information on different levels of abstraction. A compact representation of metrical structure of musical audio signals is proposed. The evaluation of low-level features enables the application of musical knowledge to a minimal degree only. On the other hand it becomes apparent, that the processing of high-level features is prone to errors due to the propagation of the errors in the extraction process of this information. This motivates the joint evaluation of low- and high-level information depending on their reliability.

The extraction of rhythmic features from information of automated detected percussive instruments represents a technical progress compared to the state-of-the-art. The segmentation of the audio signals in characteristic and similar regions representing verse or chorus for example is introduced as a valuable pre-processing step. The achieved significant improvements of the recognition rate are proved with real-world test data.

The performances of the developed methods are evaluated using a large corpus of test data and the applicability of the extracted features for the use in an exemplary MIR-system is examined.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

1.1 Einleitung und Motivation

Durch die fortschreitende Entwicklung auf dem Gebiet der Rechen-, Netzwerk- und digitalen Medientechnik stehen heutzutage vielen Menschen multimediale Daten in großem Umfang zur Verfügung. Jedes Jahr werden weltweit mehr als zehntausend Compact Discs (CD) veröffentlicht und über hunderttausend Musikstücke urheberrechtlich registriert [UZ99]. Eine steigende Anzahl von Internet-Vertrieben bietet Künstlern die Möglichkeit, ihre Produktionen auf neuen Distributionswegen zu veröffentlichen.

Eine effektive Suche nach spezifischen Inhalten kann in großen Archiven nur mit geeigneten Werkzeugen durchgeführt werden. Diese Werkzeuge sollen den Nutzer dabei unterstützen, die gewünschten Daten unter Verwendung von semantisch bedeutungsvollen und intuitiven Repräsentationen zu referenzieren. Handelt es sich bei den zu suchenden Daten um Musik, spricht man in diesem Zusammenhang von *Music Information Retrieval* (MIR). Typische Anwendungen von MIR sind die Erstellung einer *Playlist*, die Musikstücke mit bestimmten Kriterien bevorzugt oder ausschließt, oder die Suche nach zu einer Vorgabe ähnlichen Musikstücken. Eine anstrebenswerte Suchanfrage an ein zukünftiges *Home Entertainment System* beschreibt Downie [Dow03a]:

„Imagine a world where you walk up to a computer and sing the song fragment that has plaguing you since breakfast. The computer accepts your off-key singing, corrects your request, and promptly suggests to you that „Camptown Race“ is the cause of your irritation. You confirm the computer's suggestion by listening to one of the many MP3 files it has found. Satisfied, you decline the offer to retrieve all extant versions of the song, including a recently released Italian rap rendition and an orchestral score featuring a bagpipe duet.“

Die Inhalte beschreibenden Informationen werden unter dem Begriff *Metadaten* zusammengefasst. Die Aktivitäten bei der Entwicklung von standardisierten Darstellungen von Metadaten, zum Beispiel der von der Motion Picture Experts Group (MPEG) gestaltete MPEG-7-Standard [MPE01] und das Resource Description Framework [RDF04] zeigen die Bedeutung, welche diesen Informationen in der heutigen Zeit beigemessen wird.

Geeignete Metadaten für eine intuitive Suche in Musikdatenbanken beschreiben neben Entstehung und Veröffentlichung eines Werkes die melodischen, rhythmischen und harmonischen Eigenschaften, die Instrumentierung und das Genre. Ein aktuelles MIR-

System wertet bis zu 400 Merkmale aus [Mus05]. Da diese Daten für die überwiegende Mehrheit von Musiktiteln nicht verfügbar sind, und eine manuelle Extraktion sehr aufwendig ist, ist die Entwicklung maschineller Extraktionsverfahren notwendig, um umfangreiche Archive zu erschließen.

Diese Verfahren erfüllen idealerweise spezielle Funktionalitäten eines analytischen Zuhörers. In der Literatur sind in diesem Zusammenhang die Begriffe *Music-Listening Systems* [Sch00] und *Automatische Transkription*[1] [Kla04] verbreitet.

1.2 Inhaltlicher Überblick

Im nächsten Kapitel werden Grundlagen aus Musiktheorie, Psychoakustik und Kognitionswissenschaft dargestellt. Ziel ist die Schaffung einer begrifflichen Grundlage, die Identifikation von rhythmischen Merkmalen und die Erlangung von Hinweisen zu deren Extraktion aus digitalisierten Audiosignalen.

Anschließend wird der Stand der Technik auf dem Gebiet der maschinellen Rhythmusanalyse zusammengefasst. Dazu gehört eine systematische Darstellung vorhandener Modelle zur Beschreibung der metrischen Struktur, eine Übersicht über bekannte Verfahren zur Detektion von Noten, zur Identifikation perkussiver Instrumente, zur Ermittlung der metrischen Struktur eines Musikstückes und weiterer Merkmale, die zur Beschreibung abstrakterer Konzepte, zum Beispiel Intensität und Perkussivität, entwickelt wurden.

Der neuartige Ansatz zur Extraktion der rhythmischen Eigenschaften eines Musiksignals wird in Kapitel 4 vorgestellt. Er basiert auf einer Kombination bekannter und neuer Verfahren zur robusten Ermittlung der metrischen Struktur und zur maschinellen Identifikation perkussiver Instrumente ohne Tonhöheninformation.

Darauf aufbauend dient die Extraktion von Spielmustern der perkussiven Instrumente zur Gewinnung von Merkmalen, die zum Einsatz in MIR-Anwendungen vorgeschlagen werden. Die Segmentierung des Musiksignals in charakteristische Abschnitte unterstützt die rhythmische Analyse des Audiosignals. Zur Verbesserung der Erkennungsleistung werden die Resultate und Konfidenzwerte unterschiedlicher Verfahren ausgewertet. Dieser Ansatz wird hier als „redundante“ Analyse bezeichnet.

Die Erkennungsleistung und Robustheit der Teilkomponenten werden in Kapitel 5 evaluiert. Bei der Beurteilung der Verfahren spielt die Unschärfe der Referenzdaten auf Grund ihrer perzeptuellen Natur eine große Rolle, die anhand eines informellen Hörtests veranschaulicht wird. Weiterhin wird die Eignung der Merkmale zur Anwendung in einem MIR-System beispielhaft untersucht.

Abschließend wird in Kapitel 6 eine Zusammenfassung des Inhaltes der Arbeit und in Kapitel 7 ein Ausblick zu zukünftiger Forschung gegeben.

[1] Klapuri definiert „Automatische Transkription“ als den Prozess der Analyse eines akustischen musikalischen Signals zur Extraktion von Parametern der Klänge, aus denen das Musikstück gebildet ist, und die zu seiner Reproduktion ausreichen.

2 Grundlagen

2.1 Musiktheorie

Das *Oxford Concise Dictionary* definiert Musik als „die Kunst, Klänge von Stimme(n) oder Instrumenten zu kombinieren, um Schönheit in Gestalt und Ausdruck von Gefühl zu erreichen“ [CK99]. In fast allen Kulturen besteht Musik aus „organisierten, strukturierten, rhythmischen Abfolgen und Überlagerungen von Tönen, die einem ganz begrenzten Repertoire bestimmter Tonhöhen aus gewissen Tonleitern entstammen“ [Roe00]. Grundbestandteile von Musik sind Melodie, Rhythmus, Harmonie, Klang und Lyrik. Veränderungen der Ausprägungen dieser Merkmale eines Musikstückes führen nicht zwingend zur Wahrnehmung eines komplett verschiedenen Musikstücks [Dow03b].

Zahlreiche Definitionen des Begriffes Rhythmus sind in der Literatur zu finden. A. W. de Groot hat bis zum Jahr 1932 bereits über fünfzig verschiedene Bedeutungen aufgelistet [Sac53], Spitzer spricht von über einhundert [Spi03]. Über längere Zeiträume wiederkehrende Ereignisse in Natur, Technik und Kunst werden im Sprachgebrauch als *rhythmisch* bezeichnet. Rhythmen bestimmen die zeitliche Organisation fast aller Lebensvorgänge.

Das Wort *Rhythmus* stammt von dem griechischen Verb *rheo* (fließen) ab und steht im ursprünglichen Sinn für Gliederung, Reihenfolge und Ordnung. Plato bezeichnete Rhythmus als *kineseos taxis*, die Ordnung der Bewegung, Aristoxenos hingegen als *taxis chronon*, die Ordnung der Zeit [Sac53]. Nach Sachs ist Rhythmus verbunden mit Bewegung, Periodizität und Wahrnehmbarkeit [Sac53]. Ähnlich, jedoch ohne Berücksichtigung der Wahrnehmbarkeit, definiert Fleissner Rhythmus als gegeben durch die identische Wiederkehr der zeitlichen Abfolge von Änderungen einer Variable [Fle01]. Heusler definiert Rhythmus als „Gliederung der Zeit in sinnlich fassbare Teile“ und stellt damit die Gruppierungsfunktion in den Vordergrund. Iyer definiert Rhythmus als jegliche wahrgenommene oder abgeleitete zeitliche Organisation in einer Serie von Ereignissen. Alleine die Perzeption von Rhythmus manifestiert ihn, auch ohne die Intention des Produzenten [Iye98]. Large und Kolen definieren Rhythmus gleichbedeutend mit rhythmischem Muster [LK94]. Diese kurze Übersicht verschiedener Definitionen ist ein erster Hinweis auf die Mehrdeutigkeit und Unschärfe der Konzepte, die in einer Arbeit über Rhythmus eine Rolle spielen.

In der Regel liefert Musik die längsten bewusst wahrgenommenen Schallstrukturen, die unserem Gehirn begegnen [Jou01] und zur Verarbeitung im menschlichen Gehirn in klei-

nere Blöcke unterteilt werden. Rhythmus stellt in der Musik das zeitliche Ordnungsprinzip dar, das sich zwei Komponenten zu Nutze macht: die Gruppierung und das Metrum.

Gruppierung Das Prinzip der Gruppierung ist bei allen kognitiven Prozessen[1] gegenwärtig [LJ83]. Ein Zuhörer gruppiert Ereignisse im auditiven Strom zu Einheiten wie Motiven, Phrasen, Themen und Sektionen, die zusammengefasst wahrgenommen werden [Bre90]. Kleinere, untergeordnete Einheiten werden dabei hierarchisch zu übergeordneten Einheiten verbunden [CM63, LJ83]. Die Gruppierung wird durch die Ausbildung von metrischen Strukturen unterstützt [LK94].

Es wird zwischen serieller beziehungsweise horizontaler und periodischer beziehungsweise vertikaler Gruppierung unterschieden. Serielle Gruppierung wird zum Beispiel durch Ähnlichkeit in Tonhöhe und Klang der aufeinander folgenden Ereignisse verursacht. Die zeitlichen Abstände zwischen den Gruppen sind häufig die größten zeitlichen Abstände in einer gegebenen Sequenz von Ereignissen. Auf die Mechanismen der seriellen Gruppierung wird im Abschnitt 2.3 näher eingegangen. Periodische Gruppierung wird durch relative Abstände von nicht aufeinander folgenden Ereignissen hervorgerufen [Par94]. Povel spricht in diesem Zusammenhang von der Organisation von Ereignissen (seriell) und der Organisation von Intervallen (periodisch) [Pov94].

Verschiedene Autoren betrachten beide Arten der Gruppierung als unabhängig voneinander [LJ83, Pov94]. Eine Verbindung zwischen serieller und periodischer Gruppierung wird in [CM63, Ben84] durch Hinweis auf den Zusammenhang von akzentuierten Ereignissen und seriellen Gruppen angedeutet. Da sowohl Akzente als auch serielle Gruppen zur metrischen Organisation beitragen, wird periodische Gruppierung durch serielle beeinflusst [CM63, Ben84].

Die relative Bedeutung beider Formen ist vom Alter und Training des Zuhörers abhängig, wobei in der Regel für Kinder die serielle und für ausgebildete Musiker die periodische Gruppierung von größerer Bedeutung ist [Par94].

Metrum Der Begriff Metrum stammt vom griechischen Wort *metron* (das Maß) ab. In der antiken Dichtung bezeichnete er die gleichmäßig wiederkehrende Folge von betonten und unbetonten Silben. Äquivalent dazu beschreibt das musikalische Metrum das Verhältnis von betonten und unbetonten Pulsen. Eine metrische Struktur unterstützt die Ausbildung von Erwartungen beim Zuhörer, indem sie einen zeitlichen Rahmen für das Auftreten von Ereignissen, zum Beispiel Harmoniewechsel, konstituiert [LK94].

Ebenso wie die Gruppierung ist das Metrum hierarchisch organisiert, das heißt es existieren verschiedene Metren auf geschichteten Ebenen [Yes76, LJ83]. Der beim parallelen Auftreten unterschiedlicher Metren in einem Musikstück vorliegende Sonderfall wird Polymetrik genannt.

[1] In der Psychologie ist der Begriff *Chunking* gebräuchlich [Jou01].

Metren werden häufig als *binär* oder *ternär* klassifiziert. Ein binäres Metrum liegt in Taktarten vor, deren Taktzähler einer Zweierpotenz entsprechen [GM02], und wenn die Pulsperioden der unter der Zählzeit liegenden metrischen Schichten dem Verhältnis 1:2 entsprechen. Aufgrund der teilweisen rhythmischen Ähnlichkeit eines $\frac{6}{8}$-Taktes und eines $\frac{3}{4}$-Taktes gelten Taktarten, deren Taktzähler ein Vielfaches von drei darstellen, als ternär, ebenso wie ein $\frac{4}{4}$-Takt mit triolischer Unterteilung auf der der Zählzeit untergeordneten metrischen Ebene.

Neben streng hierarchisch organisierten Metren mit äquidistanten Pulsen sind in Osteuropa, Südasien und Nordafrika Rhythmen verbreitet, deren zugrunde liegender Puls zwei unterschiedliche Abstände mit einem Verhältnis von $3 : 2$ aufweist.

In der Terminologie von Schloss werden hierarchisch organisierte Metren als *divisiv* und zusammengesetzte Metren als *additiv* bezeichnet. Als dritte Kategorie werden *mesoperiodische* Rhythmen mit polyrhythmischen Strukturen sowie Mischformen definiert [Sch85].

Akzente Lerdahl und Jackendoff unterscheiden drei Arten von Akzenten: phänomenale, strukturelle und metrische Akzente. Phänomenale Akzente sind Ereignisse im Musiksignal, die eine Betonung eines Momentes erzeugen und so eine wichtige Rolle für die Ausbildung rhythmischer Strukturen spielen. Ein struktureller Akzent ist im Kontext einer Kadenz das Ziel der tonalen Bewegung. Er wird durch melodische und harmonische Schwerpunkte hervorgerufen. Metrische Akzente werden durch betonte Zählzeiten gebildet [LJ83].

Die Betonung eines Ereignisses kann verschiedenen Ursprungs sein. Zu den phänomenalen Akzenten zählen melodische, harmonische, dynamische, agogische, tonische und Rubato-Akzente. Melodische Akzente können durch Platzierung eines Spitzentons erzeugt werden, harmonische Akzente durch markante Harmoniewechsel [Hem97]. Dynamische Akzente werden durch unterschiedliche Lautstärke, agogische Akzente durch Tonlängenunterschiede und tonische Akzente durch Tonhöhenunterschiede zwischen betonten und unbetonten Noten hervorgerufen. Rubato-Akzente entstehen durch Tempoänderungen wie *accelerando* (schneller werdend) und *ritardando* (langsamer werdend) [Lem95]. Klang, Textur und Phrasur sind weitere Mittel zur Akzentuierung [Yes76]. Ein musikalischer Rhythmus ist immer durch eine Abfolge von phänomenalen Akzenten gekennzeichnet [LK94].

Puls und Pulsserie Ein Puls ist eines aus einer Serie von zeitlich subjektiv äquidistanten identischen Ereignissen. Diese Ereignisse sind nicht unbedingt im Musiksignal enthalten, aber werden vom Hörer wahrgenommen. Die Wahrnehmung einer Pulsserie wird hervorgerufen durch Noteneinsätze, Akzente oder Wiederholungen von rhythmischen Mustern. Es wird zwischen betonten und unbetonten Pulsen unterschieden.

Pulsserien treten auf hierarchisch geschichteten Ebenen auf [CM63, Yes76]. In der Li-

teratur sind alternativ die Begriffe *rhythmic stratum* [Yes76], *level of motion*, *level of pulsation* [PK90] und *rhythmic level* [Ros92] gebräuchlich. Das Konzept der Schichtung hierarchischer Pulsserien ist in Abbildung 2.1 anhand eines Clave-Rhythmus[2] illustriert. Der notierte Rhythmus wird aufgrund seiner starken Verbreitung in der kubanischen Son-Musik als *Son-Clave* bezeichnet.

Den verschiedenen Pulsschichten werden vom Zuhörer unterschiedliche Bedeutungen beigemessen. Der das musikalische Tempo bestimmende Puls wird als Zählzeit, Beat oder primärer Puls bezeichnet. Der Notenwert der Zählzeit sollte etwa in der Mitte zwischen dem langsamsten und schnellsten Notenwert des Stückes liegen [Hem97]. Der Beat ist die fundamentale Einheit der zeitlichen Struktur von Musik [GM97a]. Er besitzt ein moderates Tempo und agiert als Referenz für die anderen Schichten [LJ83]. Ein musikalischer Beat muss nicht exakt isochron sein, um als Beat wahrgenommen zu werden. Im Allgemeinen wird ein Beat aber als intensiver empfunden, je gleichmäßiger er auftritt. Diese Tatsache führt neben anderen Faktoren dazu, dass afrikanische perkussive Musik zu einem stärkeren Empfinden eines Beats führt als europäische klassische Musik [Par94].

In der englischsprachigen Literatur sind weiterhin die Begriffe „downbeat“, „offbeat“ and „upbeat“ gebräuchlich, es finden sich jedoch unterschiedliche Definitionen, so dass im Folgenden nur der Begriff „offbeat“ verwendet wird, um zwischen Zählzeiten auftretende Pulse in binären Metren zu bezeichnen.

In westlicher klassischer Musik beträgt die mittlere Anzahl der Pulsschichten fünf: jeweils zwei Ebenen treten durchschnittlich über und unter der Zählzeitebene auf. Die metrische Struktur ist jedoch innerhalb eines Stückes variabel, wobei die der Zählzeit untergeordneten Ebenen stärker variieren.

Zwei Pulsschichten, deren Perioden nicht in einem ganzzahligen Verhältnis zueinander stehen, werden als *rhythmisch dissonant* [Yes76] oder *polymetrisch* [LK94] bezeichnet. Nicht alle auftretenden Noten, zum Beispiel vereinzelte Quintolen, Triller und Verzierungsnoten, treffen mit der Pulsserie der niedrigsten Hierarchie zusammen und werden daher als „außermetrisch“ (engl. *extrametrical*) bezeichnet [LJ83, Tem01]. Weiterhin kann ein gewisses Maß an „metrischer Unexaktheit“, beispielsweise durch Temposchwankungen, auftreten [LJ83].

Tatum Der Begriff *Tatum* wurde von Bilmes geprägt und bezeichnet den Puls auf der niedrigsten metrischen Ebene [Bil93]. Gouyon verwendet den Begriff *tick* [GHC02] und Yeston den Begriff *pulse level*, im Gegensatz zu höheren *interpretative levels* [Yes76]. Das Tatumraster ist von Bedeutung zur rhythmischen Synchronisation zu Musiksignalen und zur Ermittlung einer kompakten Darstellung des Musiksignals in Form einer Transkription, da es Aufschluss über die gespielten Notenwerte geben kann. Es entspricht oft einem idealen Raster bestehend aus den kleinsten in einem Musikstück vorkommenden

[2] Claves (Klanghölzer) sind Idiophone, die in der Regel aus 2 zylindrischen Holzstücken bestehen, die aneinander geschlagen werden und einen sehr durchsetzungskräftigen Klang hervorrufen.

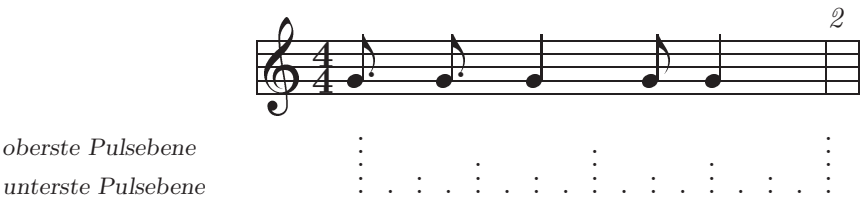


Abbildung 2.1: Rhythmusbeispiel *Son-Clave* und hierarchisch geschichtete Pulsserien. Die unterste Ebene entspricht einem Sechzehntelnoten-Raster, die oberste Ebene einem Raster aus ganzen Noten.

Noten. Allgemeingültiger ist die folgende Definition:

Das Tatumraster ist eine metrische Serie von Pulsen, die idealerweise mit den Einsätzen aller gespielter Noten übereinstimmt.

Das Tatummodell ist eng an das Pulsmodell gebunden. Es gibt jedoch eine Vielzahl von Musikstücken, bei denen die Anwendung des Pulsmodells kritisch zu betrachten ist. Dazu gehören Rhythmen mit *Swing* oder einer häufigen Variation der Pulsserie auf der niedrigsten hierarchischen Ebene, zum Beispiel dem Wechseln zwischen binärer und ternärer Mikrotime (siehe Seite 20).

Auch innerhalb eines Ensembles können unterschiedliche Tatumraster von den Musikern simultan wahrgenommen werden [Bil93].

Tempo Die herausragende Bedeutung des Tempos in der Musik betont W. A. Mozart:

Das Notwendigste, das Härteste und die Hauptsache in der Musik ist das Tempo.

Das musikalische Tempo bezeichnet die Rate aufeinander folgender Beats und bestimmt die absolute Dauer der einzelnen Notenwerte. Die am Anfang eines Musikstückes notierte Tempoangabe setzt sich aus dem Notenwert der Zählzeit und der Anzahl der in einer Minute auftretenden Zählzeiten zusammen. In Ausnahmefällen wird ein Tempo eines von der Zählzeit abweichenden Notenwertes angegeben, beispielsweise kann sich in einem $\frac{6}{8}$-Takt die Tempoangabe auf die punktierte Viertelnoten beziehen [Pus05].

Für Tempoangaben wird die Einheit *beats per minute* (bpm) oder *Metronom Mälzel*, nach dem Erfinder des Metronoms Johann Nepomuk Mälzel, verwendet. Im Tanzsport wird gelegentlich die Anzahl der Takte pro Minute als Tempo angegeben. In der klassischen Musik sind italienische Bezeichnungen für Tempo und Tempoänderung gebräuchlich. Französische und deutsche Komponisten benutzen auch Tempobezeichnungen in ihrer Muttersprache [Gra59, Hem97].

Das musikalische Tempo liegt in der Regel zwischen 40 bpm und 300 bpm [Moe02]. Für verschiedene folkloristische Musikstile werden auch höhere Werte berichtet, zum

Beispiel 330 bpm in [PAT04]. Das langsamste Tempo des in [GAD+06] beschriebenen *Audio Description Contest (ADC 2004)*[3] verwendeten Datensatzes beträgt 24 bpm. Die in Notenblättern angegebenen Tempi spiegeln nicht zwangsläufig das vom Zuhörer wahrgenommene musikalische Tempo wieder [Moe02].

Für das bevorzugte Tempo (auch *moderat*, *natürlich* oder *spontan* genannt) wurden verschiedene Werte zwischen 100 bpm und 130 bpm experimentell durch verschiedene Tapping-Experimente, die Auswertung von BPM-Listen von Tanzmusik, Hörtests bezüglich des Phänomens der subjektiven Rhythmisierung (siehe Abschnitt 2.3, Seite 34) und die Beobachtung von rhythmischem Applaus ermittelt [Fra82, Par94, NM99, Moe02].

Die Auswirkung der Bevorzugung von Tempobereichen gegenüber anderen bei der Identifikation der Hauptzählzeit wird unter anderem beim Spielen einfacher Polyrhythmen deutlich. In der Regel wird der Rhythmus als Zählzeit wahrgenommen, der dem moderaten Tempo näher ist. Abbildung 2.2 zeigt als Beispiel einen Polyrhythmus, der mit zwei Metronomen realisiert werden kann, deren Geschwindigkeiten im Verhältnis von 3:4 eingestellt ist. Für die Art der Notation wurde das *Time Unit Box System* nach Philip Harland verwendet, da es eine symbolische Darstellung ohne die Angabe einer Taktart erlaubt, die die Gleichberechtigung der zwei Metren einschränken würde.

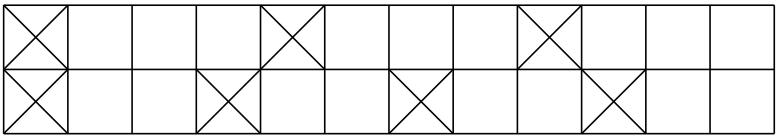


Abbildung 2.2: Beispiel eines Polyrhythmus, notiert im *Time Unit Box System* nach Philip Harland. Jede Box repräsentiert eine Zeiteinheit, ein Kreuz markiert einen Noteneinsatz.

Die A-priori-Wahrscheinlichkeiten $P(p_b)$ für das Auftreten von Beatperioden p_b wurden von Parncutt als logarithmisch normalverteilt ermittelt (siehe Gleichung 2.1).

$$P(p_b) = \exp(-0.5(\frac{1}{\sigma}\log_{10}(\frac{p_b}{\mu}))^2) \tag{2.1}$$

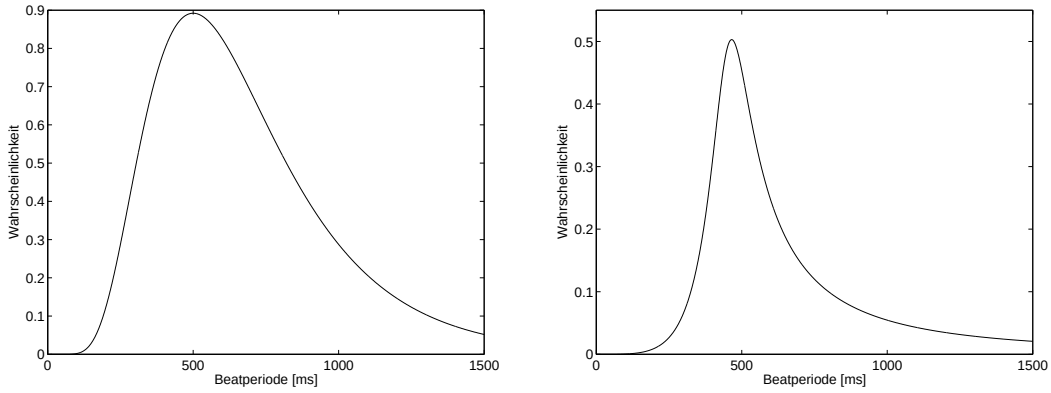
Für die moderate Pulsperiode wird $\mu = 600$ ms und für die Standardabweichung $\sigma = 0.2$ angegeben [Par94]. Van Noorden verwendet ein Resonanzmodell zur Beschreibung der Region des bevorzugten Tempos, dessen Gültigkeit anhand von Experimenten mit subjektiver Rhythmisierung, Tapping zu einfachen Tonsequenzen und Polyrhythmen sowie der Auswertung von BPM-Listen gezeigt wird. Die A-priori-Wahrscheinlichkeiten werden aus der effektiven Resonanzamplitude eines gedämpften harmonischen Oszilla-

[3] *ADC 2004* wurde im Rahmen der International Conference on Music Information Retrieval ISMIR 2004 veranstaltet, um Algorithmen der Musikanalyse anhand einheitlicher Datensätze zu evaluieren.

tors (siehe Gleichung 2.2) mit charakteristischer Frequenz f_0, Anregungsfrequenz f_{ext} und Dämpfung β ermittelt [NM99].

$$P(f_{ext}) = \frac{1}{\sqrt{(f_0^2 - f_{ext}^2)^2 + \beta f_{ext}^2}} - \frac{1}{\sqrt{f_0^4 + f_{ext}^4}} \tag{2.2}$$

In Abbildung 2.3 sind die A-priori-Wahrscheinlichkeit nach Parncutt und die effektive Resonanzkurve nach van Noorden mit Parametern $f_0 = 2.193$ Hz ($T_0 = 456$ ms), $\beta = 0.5$ und $p_b[\text{s}] = 60/f_{ext}[\text{bpm}]$ dargestellt.



(a) Parncutt (b) van Noorden

Abbildung 2.3: A-priori-Wahrscheinlichkeiten der Beatperiode nach Parncutt [Par94] (a) und nach van Noorden [NM99] (b).

Die Ergebnisse sind stark von der Methode abhängig. Drake *et. al.* ermittelten in Tapping-Experimenten mit 18 Musikern und 18 Nichtmusikern, bevorzugte Tempi von 58 beziehungsweise 70 bpm [DPB00]. In diesen Experimenten synchronisierten die Probanden ihre Bewegung zu klassischer Klaviermusik, wobei unterschiedliche metrische Ebenen als moderat ausgewählt wurden. Die Autoren schlussfolgern, dass Musiker metrische Ebenen mit langsameren Tempi bevorzugen als Nichtmusiker. Die Auswahl einer metrischen Ebene als Zählzeit kann weiterhin beeinflusst werden durch die Art der Aufführung, bei der ein Dirigent in Abhängigkeit vom Orchester die Ebene der Zählzeit auswählt [Pus05].

Takt und Taktart Der Takt (von *tactus*, das Berühren, der Schlag, der Gefühlssinn) ist die kleinste höhere Einheit, zu der mehrere Zählzeiten zusammengefasst werden. Er gliedert den Puls in gleichmäßig wiederkehrende Gruppen und schafft eine Gewichtung zwischen betonten und unbetonten Zählzeiten. Der erste Puls ist häufig betont und wird von einem oder mehreren unbetonten Pulsen gefolgt.

Die Taktart wird in Form eines mathematischen Bruches beschrieben. Der Taktnenner bezeichnet den Notenwert der Zählzeit. Der Taktzähler zeigt an, wie viele Zählzeiten zu einem Takt zusammengefasst sind. Es wird zwischen einfachen und zusammengesetzten Takten unterschieden. Einfache Taktarten werden in gerade beziehungsweise binäre und ungerade beziehungsweise ternäre Taktarten klassifiziert. Zusammengesetzte Taktarten werden aus Kombinationen von bis zu vier einfachen Takten gebildet, wobei Kombinationen gleicher Taktarten als *regelmäßig zusammengesetzt* und Kombinationen unterschiedlicher Taktarten als *unregelmäßig zusammengesetzt* bezeichnet werden [Hem97].

Innerhalb eines Stückes kann die Taktart häufig wechseln, wie zum Beispiel in Bernsteins Stück „I like to be in America“, welches zwar im Original im $\frac{6}{8}$-Takt geschrieben wurde, aber aus Gründen der Lesbarkeit häufig mit einem Wechsel zwischen $\frac{6}{8}$-Takt und $\frac{3}{4}$-Takt notiert wird.

Swing Für den mehrdeutigen Begriff *Swing* wird im Rahmen dieser Arbeit eine Definition ähnlich zu der in [Lar01] gewählt:

Swing ist eine leichte Verzögerung der unbetonten Pulse.

Die ungleiche Einteilung der Pulsserie fördert die Wahrnehmung der höheren metrischen Ebene, des Beats, durch die Gruppierung jeweils eines betonten und unbetonten Pulses[4].

Binäre und ternäre Mikrotime Als *Mikrotime* wird das im Allgemeinen ganzzahlige Verhältnis zwischen Periode der Zählzeit und Periode des Tatum bezeichnet [Mar91]. Ähnlich wie die Taktart ist es ein Merkmal der metrischen Struktur. Der Begriff Mikrotime ist jedoch weniger stark verbreitet und auch weniger eindeutig für viele Musikstücke.

Eine Mikrotime, die einem Vielfachen von zwei beziehungsweise drei entspricht, wird als *binär* beziehungsweise *ternär* bezeichnet.

Rhythmische Muster Als rhythmische Muster werden wiederkehrende Abfolgen von Noten und Pausen bezeichnet [LK94]. Sie entstehen durch Wiederholungen und durch Diskontinuitäten in der Musik [Meu03]. Das einfachste rhythmische Muster ist eine isochrone Abfolge von identischen Ereignissen [RWD02]. Large und Kolen verwenden die folgende Definition [LK94]:

When we speak of “a rhythm” or “a rhythmic pattern” we will mean the pattern of inter-onset durations associated with a music sequence. In music, a rhythm has an associated pattern of phenomenal accents, which is the physical patterning of events in the musical stream such that some seem to be stressed relative to others.

[4] Neben der hier dargestellten Bedeutung bezeichnet der Begriff *Swing* ein musikalisches Genre.

Für vom Schlagzeug gespielte rhythmische Muster hat sich der Begriff *Drumpattern* etabliert. Diese Muster vereinen Informationen über das zeitliche Auftreten von Akzenten mit Instrumenteninformationen.

Als *inhärent* werden Muster bezeichnet, die nicht direkt gespielt werden, deren Wahrnehmung jedoch durch komplexes Arrangement von einzelnen Stimmen hervorgerufen wird. Dieses Phänomen wurde besonders in afrikanischer Musik und Bach'scher Orgel- und Violinmusik mit schnellen Pulsserien beobachtet [CK99].

2.2 Psychoakustik

Psychoakustik ist als Teilgebiet der Psychophysik [Roe00] ein mehr als einhundert Jahre altes Forschungsgebiet über die auditive Wahrnehmung des Menschen. Verschiedene Erkenntnisse über Zusammenhänge zwischen akustischem Reiz und Hörwahrnehmung sind zumeist empirisch in subjektiven Hörtests ermittelt worden. Im Gegensatz zur klassischen Physik sind die von der Psychophysik getroffenen Aussagen zumeist statistischer Natur, ähnlich der Quantenphysik. Für die Analyse eines Musiksignals zur Beschreibung rhythmischer Eigenschaften sind Erkenntnisse bezüglich der Ausbildung von phänomenalen Akzenten, der zeitlichen Auflösung und der Gruppierung von Ereignissen von besonderem Interesse.

Das Gehör

Das Gehör ist das empfindlichste Sinnesorgan des Menschen [Zen94]. Ein erwachsener Mensch hört in der Regel Frequenzen zwischen 20 Hz und 15 bis 20 kHz und empfindet Lautstärkepegel zwischen 4 und 130 phon. Der Hörbereich ist nach unten von der Ruhehörschwelle und nach oben von der oberen Hörgrenze (Schmerzschwelle) begrenzt [Roe00].

Der äußere Gehörapparat des Menschen setzt sich aus Außen-, Mittel- und Innenohr zusammen (siehe Abbildung 2.4). Er hat die Aufgabe, einfallende Schallwellen aufzunehmen und so aufzubereiten, dass das Gehirn diese in einen Höreindruck umwandeln kann.

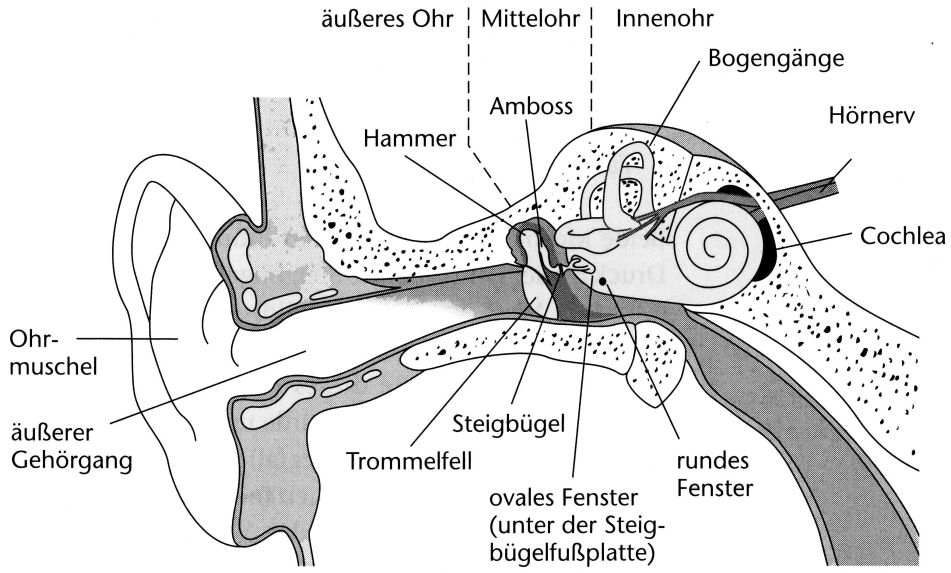


Abbildung 2.4: Schematische Darstellung des äußeren Gehörapparates nach [Gol02].

Das Außenohr Das Außenohr setzt sich aus der Ohrmuschel, der *Pinna*, und dem äußeren Gehörgang zusammen und leitet die Schallwellen zum Trommelfell weiter. Die Ohrmuschel stellt einen Schalltrichter dar, dessen Übertragungsfunktion von der Einfallsrichtung des Schalls abhängt. Die richtungsabhängige Betonung und Absenkung schmaler Frequenzbänder ermöglicht zusammen mit der Auswertung binauraler Intensitäts- und Phasenunterschiede die Lokalisation von Schallquellen.

Frequenzen zwischen 2 kHz und 5 kHz werden durch eine von der Form der Ohrmuschel hervorgerufene Resonanz, der *Cavumconchae-Resonanz*, verstärkt [Zen94]. Abschattungen und Reflexionen an Kopf und Schulterpartie verursachen weitere Modifikationen des Schallsignals [ZF99].

Der äußere Gehörgang kann näherungsweise als halboffenes Rohr mit 30 mm Länge und 8 bis 11 mm Breite betrachtet werden. Bedingt durch diese Geometrie werden Frequenzen um 3 kHz durch Resonanzvorgänge um bis zu 20 dB verstärkt[5].

Das Mittelohr Das Mittelohr wird durch die Paukenhöhle, einen kleinen luftgefüllten Raum, gebildet und ist mit dem Nasen-Rachen-Raum durch die Eustachische Röhre verbunden. Das Außenohr und die Paukenhöhle sind mit Luft gefüllt, das Innenohr mit Lymphflüssigkeit. Eine Impedanzanpassung verhindert Reflexionsverluste beim Übergang zwischen den zwei unterschiedlich schallleitenden Medien:

Die Mittelohrknöchelchen (Ossikel) Hammer (Malleus), Amboss (Incus) und Steigbügel (Stapes) leiten die Schwingungen des Trommelfells auf eine weitere Membran weiter, die eine Öffnung des Innenohrs, das ovale Fenster, abdeckt (siehe Abbildung 2.5). Da das Trommelfell eine größere Fläche als das ovale Fenster aufweist, wobei das Verhältnis der Flächen zueinander in etwa 17:1 beträgt, und die Gehörknöchelchen ein Hebelsystem bilden, wird eine Druckerhöhung erreicht. Im Zusammenspiel mit bisher weniger erforschten dynamischen Eigenschaften des Mittelohrs wird eine etwa 60-prozentige Absorption der Schallenergie erreicht. Zwei an Hammer und Steigbügel ansetzende Muskel erfüllen eine Schutzfunktion gegenüber großen Schalldrücken [Zen94].

Das Innenohr Das Innenohr enthält den für den Gleichgewichtssinn zuständigen Vestibularapparat und das eigentliche Hörorgan, die *Cochlea* (Hörschnecke). Die Cochlea ist äußerlich ein schneckenförmig gewundener Kanal mit zweieinhalb Windungen im menschlichen Schläfenknochen. Ihre Funktion ist die Umwandlung von Schall in Nervenimpulse.

Die kochleäre Trennwand unterteilt die Chochlea der Länge nach in zwei mit Perilymphe gefüllte Räume, die *Scala vestibuli* und die *Scala tympani*. Sie enthält die mit

[5] In der Literatur sind für die Resonanzfrequenz Werte zwischen 2500 Hz [Zen94] und 3400 Hz [Gol02] angegeben. In einem idealen halboffenen Rohr wird eine Frequenz, deren Wellenlänge dem Vierfachen der Länge des Rohrs entspricht, verstärkt. Für eine effektive Länge des Gehörgangs von 30 mm errechnet sich die Resonanzfrequenz bei Körpertemperatur zu 2950 Hz.

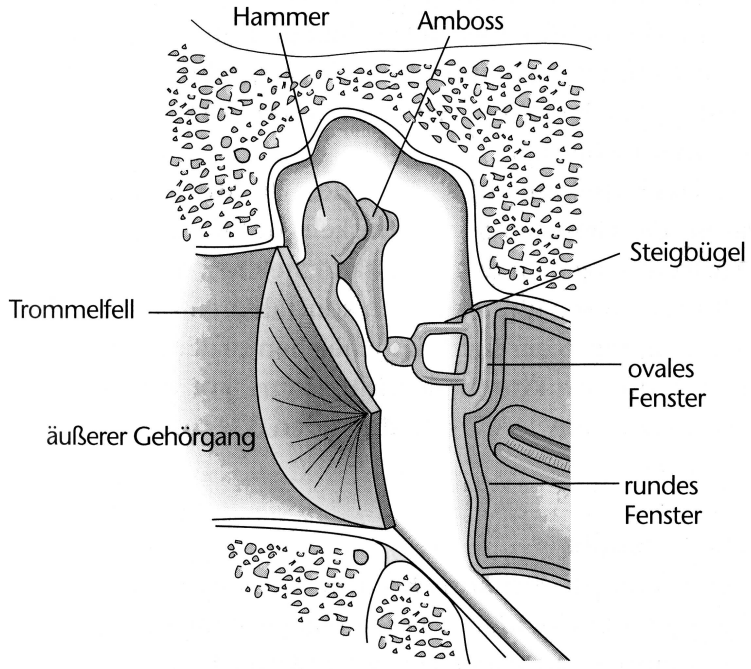


Abbildung 2.5: Schematische Darstellung des Mittelohres nach [Gol02].

Endolymphe gefüllte *Scala media*, die nach oben durch die Reissnersche Membran und nach unten durch die Basilarmembran (BM) begrenzt ist (siehe Abbildung 2.6).

Die Schallwellen treten am ovalen Fenster in die Scala vestibuli ein und lösen eine Wanderwelle entlang der BM aus. Die Amplitude der Welle nimmt bei ihrem Weg in die Schnecke zu, die Fortpflanzungsgeschwindigkeit jedoch ab, bis sie einen Resonanzpunkt erreicht. Die Lage des Resonanzpunktes ist frequenzabhängig: er liegt für tiefe Frequenzen weiter im Inneren der Schnecke als für hohe Frequenzen, so dass eine Frequenz-Orts-Analyse des Schalls stattfindet [Yat95, Gol02].

Das eigentliche Sinnesorgan des Gehörsinns ist das sich auf der BM befindende Cortische Organ, dessen wichtigste Bestandteile die Tektorialmembran und die Haarzellen (HZ) sind. Die Bewegungen der BM werden von einer inneren und drei äußeren Reihen von HZ aufgenommen. Die inneren HZ agieren als mechanische Dehnungsmesser, deren Auslenkung phasengekoppelte Ausschüttungen von chemischen Botenstoffen (Transmittern) an die mit den Haarzellen verbundenen Nervenfasern auslösen, die wiederum ein elektrophysiologisches Signal an das zentrale auditive System weiterleiten. Die inneren HZ werden bei leisen und mittleren Lautstärken nur angeregt, wenn die äußeren HZ die Resonanzschwingung verstärken.

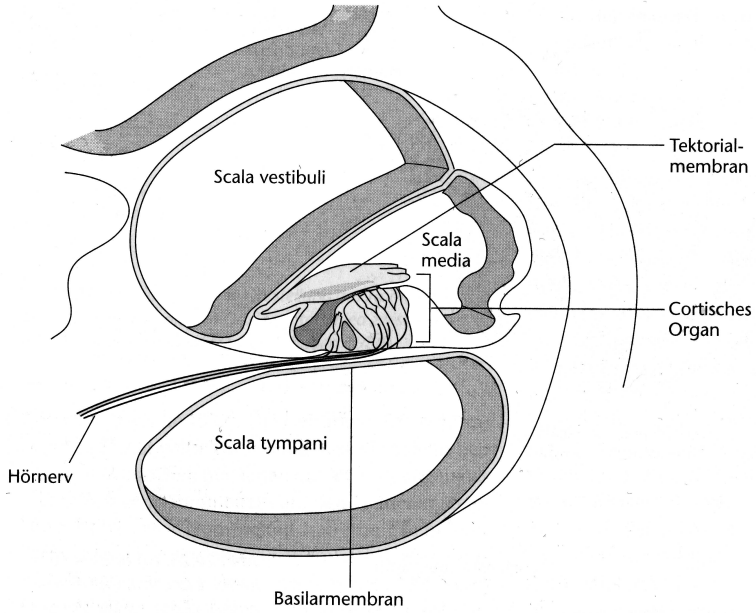


Abbildung 2.6: Querschnitt durch die Cochlea nach [Gol02].

Spektrale Auflösung des Gehörs Die in der Cochlea stattfindende Frequenz-Orts-Analyse des Schalls erfüllt die Funktion einer Filterbank. Die Übertragungsfunktion der Filter ist abhängig vom Eingangspegel. Bei moderaten Schallpegeln besitzen die Filter eine annähernd symmetrische und bei größeren Pegeln eine linksschiefe Übertragungsfunktion [Moo95].

Die Frequenz-Orts-Analyse scheint jedoch allein nicht ausreichend, um eine Frequenzanalyse komplexer Signale zu ermöglichen, da schon bei mittleren Schalldruckpegeln, besonders in Gegenwart von Hintergrundrauschen, nahezu alle Fasern des Hörnervs angeregt werden. Nach Klinke muss daher die Zusammensetzung eines Schallreizes aufgrund zusätzlicher Informationen vom Gehirn ermittelt werden [Zen94].

Der kleinste wahrnehmbare Unterschied (engl. *just noticeable difference, JND*) der Frequenz eines reinen Tones ist von Frequenz und Tondauer abhängig [Roe00] und beträgt 1 Hz für Frequenzen bis 500 Hz, darüber 0.2% der zu vergleichenden Frequenzen. Die kleinste wahrnehmbare Frequenzmodulation ist um Faktor 3 größer [ZF99].

Georg von Békésy ermittelte erstmals die Abhängigkeit des Ortes maximaler Erregung von der Frequenz eines sinusförmigen Erregers als logarithmischen Zusammenhang. Eine Verdopplung der Frequenz verschiebt den Resonanzbereich um 3.5 mm bis 4 mm [Roe00].

Zwei reine Töne werden nicht einzeln wahrgenommen, wenn die angeregten Resonanzbereiche der BM weniger als 1.2 mm voneinander entfernt sind [Roe00]. Die Frequenz-

differenz ist dann kleiner als die Bandbreite einer Frequenzgruppe (engl. *critical bandwidth*), und beide Frequenzen liegen innerhalb einer Frequenzgruppe (engl. *critical band*). Das Konzept der Frequenzgruppen wurde von Fletcher entwickelt, der den Schwellwert für die Detektion eines reinen Tones in Abhängigkeit von der Bandbreite eines bandbegrenzten Rauschsignals als Maskierer[6] ermittelte [ZF99, Moo95, Bra98].

In Fletchers Experimenten wurde die Mittenfrequenz des Maskierers identisch zur Frequenz des Testsinusoids gewählt. Bei konstanter Rauschenergiedichte wurde der Detektions-Schwellwert als Funktion der Bandbreite des Maskierers gemessen. Der Schwellwert steigt solange an, bis die Bandbreite des Maskierers der Bandbreite der Frequenzgruppe entspricht und bleibt danach konstant. Diese Erkenntnisse führten zur Entwicklung verschiedener gehörangepasster Frequenzskalen, wie zum Beispiel der Bark-, Mel- oder ERB (Equivalent Rectangular Bandwidth)-Skala. Da die Frequenzgruppen vom menschlichen Gehör dynamisch gebildet werden, ist die bei vielen Frequenzskalen implementierte feste Bandeinteilung für verschiedene Anwendungen nachteilig [Spo98].

Intensitätsempfinden Die Empfindung von Intensitäten und Intensitätsunterschieden spielt für die Ausbildung dynamischer Akzente und dadurch für die Wahrnehmung rhythmischer Strukturen eine bedeutende Rolle.

Eine Annahme der Psychophysik ist, dass die Wahrnehmung W eine Funktion des Reizes R ist. Nach Fechner ist der Zusammenhang zwischen Reiz- und Empfindungsstärke logarithmischer Natur und wird Weber-Fechnersches Gesetz genannt (siehe Gleichung 2.3, mit $k, d = \text{const}$).

$$W = k \cdot \ln R + d \tag{2.3}$$

Ein ähnlicher Zusammenhang ist als Stevens'sches Potenzgesetz bekannt (siehe Gleichung 2.4, mit $c, n = \text{const}$ und $n < 1$) [Ter98].

$$W = c \cdot R^n \tag{2.4}$$

Die Amplitude eines Schallreizes wird demzufolge im Gehör annähernd logarithmisch abgebildet [Püs88, ZF99].

Eine die Intensität eines auditiven Ereignisses beschreibende Kenngröße ist der von Barkhausen eingeführte Lautstärkepegel (engl. *loudness level*), der in der Maßeinheit *Phon* angegeben wird. Bei einer Schall-Frequenz von 1000 Hz stimmen Schalldruckpegel (engl. *sound pressure level*, SPL), gemessen in dB SPL, und Lautstärkepegel, gemessen in Phon, überein [ZF99].

Die subjektive Quantität einer auditiven Empfindung, welche am nahesten zu deren Intensität korrespondiert, ist die Lautheit (engl. *loudness*) [PC95, ZF99]. Sie

[6] Als Maskierer wird ein Signal bezeichnet, bei dessen simultaner Präsentation mit einem Testsignal, welches ohne Maskierer hörbar ist, das Testsignal nicht mehr wahrgenommen wird.

wird nach Stevens in der Einheit *Sone* gemessen, wobei ein Sone der Lautheit eines 1-kHz-Sinustones mit 40 dB SPL entspricht. Ein doppelt so laut empfundener Reiz entspricht einer Lautheit von zwei Sone. Das Lautheitsmodell nach Zwicker misst die spezifische Lautheit innerhalb einer Frequenzgruppe und bildet die Gesamtlautheit durch Integration über alle Frequenzgruppen.

Das Intensitätsempfinden wird beeinflusst von Frequenz, Bandbreite und Dauer des Schallreizes. Die Abhängigkeit von der Frequenz ist in den Kurven gleicher Lautstärke nach Fletcher und Munson, den Isophonen, veranschaulicht. Die Fähigkeit des Gehörs, Amplitudenunterschiede wahrzunehmen, wurde in verschiedenen Experimenten untersucht, wobei zwischen der kleinsten wahrnehmbaren Amplitudenvariation und Amplitudendifferenz unterschieden wird [PC95, ZF99]. Harris nennt diese zwei prinzipiellen Erscheinungen *loudness modulation* und *loudness memory* [Moo82].

Die JND ist abhängig von der Intensität, der Dauer und den klanglichen Eigenschaften des Stimulus. Die kleinste wahrnehmbare Intensitätsänderung ΔI verhält sich annähernd proportional zur Intensität I. Die Konstante $\Delta I/I$ wird als *Weber'scher Quotient* (WQ) und der grundsätzliche Zusammenhang als *Weber'sches Gesetz* bezeichnet. Amplitudenmodulationen in schmalbandigen Rauschsignalen werden ab $\Delta I/I = 0.25$ wahrgenommen. Dieser Wert entspricht einem logarithmischen Schwellwert-Faktor $\Delta L_s = 10log_{10}(1 + \Delta I/I)$ dB ≈ 1 dB [ZF99]. Für breitbandiges Rauschen sind in der Literatur Werte zwischen 0.5 dB und 1 dB angegeben, die innerhalb eines Dynamikbereiches von 20 dB bis 100 dB über der Ruhehörschwelle gelten [Moo82].

Leichte Abweichungen vom proportionalen Zusammenhang treten bei der Wahrnehmung reiner Töne auf: der Quotient $\Delta I/I$ sinkt bei steigender Intensität. Dieser Sachverhalt wird in der englischsprachigen Literatur als „near miss to Weber's Law“ bezeichnet [Moo82, PC95].

Die dargestellten Werte für den WQ wurden für Stimuli von Dauern über 500 ms ermittelt. Für kurze Reize bis zu einer kritischen Dauer ist jedoch die kleinste detektierbare Energie konstant, $I \cdot t = \text{const}$. Die kritische Dauer ist frequenzabhängig und sinkt nach Henning von 100 ms bei 250 Hz auf 10 ms bei 4 kHz. Florentine ermittelte für die kritische Dauer einen entgegengesetzten Zusammenhang mit Werten von 500 ms bei 250 Hz bis 2 s bei 8 kHz [PC95] an.

Zeitliche Zusammenfassung von Reizen Die Abhängigkeit der Lautheitsempfindung und des JND von der Reizdauer weist darauf hin, dass das Gehör Reize zeitlich integriert. Kürzere Stimuli erfordern größere Intensität als längere, um detektiert zu werden. Nach Munson geschieht die Integration gewichtet mit einer Exponentialfunktion mit einer Zeitkonstante von 200 ms [EG95].

Die Zeitkonstante wurde von Plomp und Boumann als frequenzabhängig ermittelt. Sie publizierten Werte von 375 ms bei 250 Hz und 150 ms bei 8000 Hz. Andere Arbeiten bestätigen die Frequenzabhängigkeit, jedoch nicht die konkreten Werte der Zeit-

konstante. Olsen und Carhart berichten identische Zeitkonstanten für Reize von 250, 1000 und 4000 Hz. Green experimentierte mit Reizen unterschiedlicher Länge und gleicher Energie. Die Detektierbarkeit dieser Stimuli war konstant für Dauern zwischen 15 und 150 ms. Dieses Plateau wurde auch von Sheeley und Bilger experimentell nachgewiesen, wobei für höhere Frequenzen kleinere Zeitkonstanten ermittelt wurden. Ein umfassender Überblick über die angeführten Arbeiten ist in [Moo82, EG95] zu finden.

Zeitliche Auflösung des Gehörs Erkenntnisse über die zeitliche Auflösung des Gehörs sind von Bedeutung für die automatische Extraktion von Noteneinsätzen sowie Tatumschätzung und -tracking[7]. In Untersuchungen zur Ermittlung der zeitlichen Auflösung wurden verschiedene Experimente mit zumeist synthetischen Stimuli durchgeführt.

In einem Experiment werden zwei Sinustöne dem Zuhörer zeitlich versetzt präsentiert, um den kleinsten Versatz zu ermitteln, der eine korrekte Bestimmung der Reihenfolge des Auftretens der Stimuli in 75% aller Fälle erlaubt [Hir59, EG95]. Für zwei Töne von 500 ms Dauer und unterschiedlicher Frequenz wurde unabhängig vom Frequenzunterschied ein Zeitintervall von 20 bis 30 ms ermittelt. Spätere Experimente zur Bestimmung der Reihenfolge zwischen zwei Lichtsignalen sowie einem Licht und einem Tonsignal ergaben eine identische Zeitspanne. Der notwendige Versatz sinkt für trainierte Zuhörer [HS61].

Diese Größe ist abhängig von der Dauer der Stimuli. Leshowitz benutzte in einem Experiment zwei Rechtecksignale von 10 μs Dauer zur Ermittlung des kleinsten Abstandes, der ein Hören von zwei getrennten Ereignissen erlaubt. Dies gelingt ab einem Intervall von 5 bis 10 μs. Leshowitz schlussfolgerte jedoch, dass die Wahrnehmung zweier getrennter Ereignisse auf Änderungen im Spektrum der Stimuli basiert und sein Experiment keine direkte Aussage über die Wahrnehmung zweier zeitlich getrennter Ereignisse liefert [EG95].

Verschiedene Experimente mit Stimuli mit identischen Energiespektren und unterschiedlicher Intensität deuten auf einen minimal notwendigen Versatz von 2 ms zur Ermittlung der Reihenfolge zweier in Folge präsentierter Ereignisse hin. Durch intensives Training kann dieser Wert auf 200 μs gesenkt werden [EG95].

In einem weiteren Experiment wird die binaurale Laterisation[8] von Sinusklängen mit Frequenzen kleiner als 1400 Hz zur Ermittlung der zeitlichen Auflösung des Gehörs verwendet. Wird ein Sinuston binaural identisch präsentiert, so wird der Ton inmitten des Kopfes geortet. Wird der Ton auf einer Seite um ein kleines Zeitintervall verzögert, so wird der Ton entfernt der Mitte geortet. Mit diesem Experiment wurde eine zeitliche

[7] *Tracking* bezeichnet in dieser Arbeit die Ermittlung der Positionen der Elemente eines metrischen Rasters.

[8] Laterisation bezeichnet im Kontext von Hörexperimenten den Umstand, wenn ein Proband einen Stimulus in einem Ohr, nicht in der Kopfmitte oder in beiden Ohren hört.

Auflösung von unter einer Millisekunde ermittelt [Pie99].

Zeitliche Auflösung in der musikalischen Praxis Inwieweit sind diese mit Laborsignalen ermittelten Erkenntnisse im Hinblick auf das Hören von Musik von Bedeutung? Während die zeitliche Auflösung des Ohres für verschiedene synthetische Teststimuli in etwa 2 ms beträgt, sind die perzeptuellen Einsatzzeitpunkte für musikalische Klänge nicht nur unterschiedlich zwischen Instrumenten, sondern auch zwischen verschiedenen Hörern und Observationen. Dabei können die wahrgenommenen Einsatzzeitpunkte für ein Instrument um bis zu 20 ms voneinander abweichen [Gor87]. Auftretende Reflexionen werden häufig erst ab einer Verzögerung von 60 bis 70 ms als separater Klang (Echo) wahrgenommen. Für kleinere Verzögerungszeiten erscheint das Geräusch von der Klangquelle ausgehend.

In der Aufführungspraxis kleiner Ensembles kann synchrones Spielen zwischen den Musikern in etwa um 30 bis 50 ms abweichen, mit einem Mittelwert von 36 ms, wie Untersuchungen anhand von drei klassischen Triobesetzungen zeigten [Ras79]. Bilmes ermittelte bei Auswertungen von Mehrspur-Aufnahmen eines kubanischen Perkussionsensembles, den *Los Muñequitos de Matanzas*, dass die *Quinto* und *Segundo*[9] im Mittel 20 bis 30 ms vor dem Metrum spielen [Bil93]. Abweichungen bis zu 4 ms zwischen den gespielten Noten und der Notation treten nach Sundberg häufig auf und werden vom Zuhörer oft nicht bemerkt [Dix99].

Die Toleranz gegenüber Abweichungen vom synchronen Spiel ist abhängig vom Klangcharakter (kleiner für transiente Klänge), Tempo (kleiner für schnelle Tempi) und von der Art der musikalischen Darbietung.

Subjektive Dauer von Klängen und Pausen Für die Messung der subjektiven Dauer wurde von Zwicker und Fastl die Einheit *Dura* vorgeschlagen. Ein 1-kHz-Ton mit einem SPL von 60 dB und einer Dauer von einer Sekunde wurde als Referenzwert für die subjektive Dauer von 1 Dura gewählt [ZF99].

Die subjektive und objektiven Dauer verhalten sich proportional für Stimuli, deren Dauer 100 ms überschreiten. Kürzere Stimuli werden verhältnismäßig länger wahrgenommen. Die subjektive Dauer einer Pause ist gleich der eines Klanges, wenn deren physikalische Dauern größer als 1 s sind. Ein 3200-Hz-Ton mit einer physikalischen Dauer von 100 ms verursacht die gleiche subjektive Dauer wie eine 400 ms lange Pause. Für einen 200-Hz-Ton oder weißes Rauschen ist dieser Effekt weniger stark ausgeprägt. Die subjektive Dauer eines 100 ms dauernden Stimulus entspricht der einer Pause von 200 ms.

Zwicker und Fastl erklären diese Wahrnehmung mit zeitlichen Verdeckungseffekten, wobei der Nachverdeckung eine größere Bedeutung beigemessen wird, und stellen wei-

[9] Quinto und Segundo sind lateinamerikanische Congas. Die Quinto ist von kleinerem Durchmesser als die Segundo, klingt höher und spielt in der Regel die führende Stimme.

terhin fest, dass der Ausarbeitung eines umfangreicheren Modells der geringe Umfang psychoakustischer Untersuchungen der subjektiven Dauer im Wege steht [ZF99].

In der musikalischen Praxis führt der Effekt der subjektiven Dauer dazu, dass schnelle Noten kürzer als die korrespondierenden Pausenwerte gespielt werden.

2.3 Kognitionswissenschaft

Ohne die strukturbildende Fähigkeit des menschlichen Geistes wäre Musik lediglich ein akustisches Signal [BG98]. Gegenstand dieses Abschnitts sind die kognitiven Aspekte des Musikhörens. Kognitionswissenschaft ist eine relativ junge Disziplin, die Erkenntnisse und Methoden der Psychologie, der Neurowissenschaften, der Sprachwissenschaft, der Philosophie und der Informatik anwendet. Ihr allgemeines Ziel ist Modellierung kognitiver Systeme [GK05].

Wahrnehmung und Kognition Wahrnehmung ist die gezielte Verarbeitung eintreffender Informationen. Die Gestaltpsychologie gibt als grundlegendes Rahmenprinzip zur Wahrnehmungsorganisation das Prägnanzprinzip an, welches Kofka folgendermaßen definiert:

„Wenn eine Reizkonfiguration mehrere alternative Gliederungen zulässt, wird sich von den mögliche Kombinationen stets jene durchsetzen, die die einfachste, einheitlichste oder auch beste Gestalt ergibt.“

Der Begriff „beste Gestalt“ wird durch die aus der Gestaltpsychologie stammenden Prinzipien konkretisiert, zu denen das Gesetz der Einfachheit, der Nähe, der Ähnlichkeit, der guten Fortsetzung, der Geschlossenheit und des gemeinsamen Schicksals gehören [CG91, Pos04]. Als weitere Theorien zur Wahrnehmungsorganisation sind die Schablonentheorie, die Theorie der Prototypen, die Theorie der Merkmalsanalyse und der Ansatz nach Marr, die auf Grund ihrer Ausrichtung auf die visuelle Wahrnehmung nicht näher betrachtet werden.

Der Begriff Kognition (vom lateinischen Wort *cognoscere* abstammend) charakterisiert Prozesse des Wahrnehmens, der Erkenntnis, des Vorstellens, des Wissens, des Denkens, der Kommunikation und der Handlungsplanung [Pos04].

Wahrnehmung von Rhythmus Zeitlich äquidistant auftretende Klänge rufen den Eindruck eines einheitlichen Rhythmus hervor, wenn die Stimuli von kurzer Dauer sind und ihre Hüllkurven steil ansteigen [ZF99]. Zeitspannen zwischen Noteneinsätzen im Bereich von 100 ms bis 5 s werden als rhythmisch im musikalischen Sinne beziehungsweise *organisiert rhythmisch* wahrgenommen [Kru00].

Die kognitive Leistung zeigt sich jedoch in der Verarbeitung von Stimuli mit expressivem Tempo[10], synkopierten Rhythmen, lückenhaften phänomenalen Akzenten und fehlerbehafteten Messdaten. Die erstere der genannten Schwierigkeiten ist nach Desain und Honing von größter Bedeutung [DH89].

[10] Die Variation des Tempos ist eine Gestaltungsmöglichkeit zum Erlangen von musikalischem Ausdruck.

Abweichungen einzelner Ereignisse von der erwarteten metrischen Position durch expressives Tempo führen zu mehrdeutigen Akzentinformationen. Tanguiane formuliert das Problem der Mehrdeutigkeit wie folgt [Tan93]:

„The difficulty of the problem is caused by the fact that the tempo is perceived with respect to repeating rhythmic patterns, whereas the rhythmic patterns are recognized as repeated with respect to a certain tempo. It implies the ambiguity in interpreting each duration, since a certain duration can be identified either as given or as another value distorted by a tempo change“

Ein Zuhörer kann nach einer kleinen Anzahl von Noten einen zugrunde liegenden Pulsschlag erkennen. Dieser Prozess wird in der Literatur als *bottom-up process* bezeichnet. Dieser Pulsschlag dient als Orientierung und zeitlichen Rahmen zur Verarbeitung der folgenden akustischen Signale, die dementsprechend als *top-down process* genannt wird [DH94].

Aus der Gestaltpsychologie ist bekannt, dass bestimmte Störungen einer bekannten Form ihrer Erkennung nicht im Weg stehen, da fehlende Details vom Beobachter rekonstruiert werden können. Diese Fähigkeit führt in der akustischen Wahrnehmung zum *Cocktailpartyeffekt* (siehe Abschnitt 3.6).

Als grundsätzliches Problem bei der Erforschung der kognitiven Vorgänge beim Musikhören nennt Tanguiane weiterhin das Fehlen von eindeutigen Definitionen der Objekte der Musikanalyse, zu denen beispielsweise Noten, Akkorde, Rhythmus und Tempo zählen, deren Definitionen zudem teilweise voneinander abhängen [Tan93].

Konnektionistischer Ansatz Wahrnehmung der zeitlichen Struktur von Musik ist eng verbunden mit der Wahrnehmung der metrischen Struktur, die einen zeitlichen Rahmen für die Erwartung von Ereignissen schafft. Diese Erwartung ist nach Meyer der Schlüssel zum Verständnis der intellektuellen und emotionalen Reaktion des Menschen auf Musik [LK94]. Der konnektionistische Ansatz zum Verständnis der musikalischen Wahrnehmung basiert auf rückgekoppelten Netzwerken zur Vorhersage zukünftiger Ereignisse auf Grundlage der Kenntnis vergangener Ereignisse.

Large und Kolens funktionaler Ansatz verwendet ein Netzwerk von *integrate-and-fire*-Oszillatoren [LK94] mit verschiedenen Frequenzbereichen. Die Art der Kopplung der einzelnen Oszillatoren wird jedoch als Gegenstand späterer Forschung nicht behandelt. Longuet-Higgins und Lees Modell basiert auf der Auswertung von Noteneinsätzen [LiL82].

Einen zu den auf Gestaltprinzipien basierenden Modellen alternativen Ansatz stellen die gedächtnisbasierten Modelle dar [ZBH05]. Gedächtnisbasierte Modelle leiten eine metrische Struktur auf der Grundlage von Wahrscheinlichkeiten ab, die aus dem Vergleich mit angelernten Beispielen ermittelt werden. Ein Überblick über weitere Ansätze enthält [DH94].

Lerdahl und Jackendoffs Regelsystem Lerdahl und Jackendoffs auf klassische westliche Musik beschränkte Theorie zur Wahrnehmung von rhythmischen Strukturen ist in zwei Regelsätzen formuliert. Die *well-formedness rules*[11] beschreiben mögliche metrische Struktur als aufgebaut aus Ebenen mit Perioden, die im Verhältnis 2:1 oder 3:1 stehen. Die *preference rules*[12] bestimmen, welche der möglichen metrischen Strukturen für ein vorliegendes rhythmisches Muster von einem erfahrenen Zuhörer wahrgenommen wird [LJ83].

Trotz der umfangreichen Forschungsaktivitäten bezüglich der Wahrnehmung von Rhythmus (siehe [ZF99, DH00, Tem01]) hat sich keine allgemein akzeptierte Theorie etabliert [Wey01, Fri04].

Soziale und kulturelle Aspekte Die Wahrnehmung und Kognition von Musik und Rhythmus ist beeinflusst von sensomotorischen Eigenschaften und von der sozialen und kulturellen Umgebung des Zuhörers [Che94, Hem97, Iye98, IPO04]. Es existieren universelle sensorische, perzeptuelle und kognitive Prozesse, unabhängig von sozialer Umgebung und musikalischer Kultur. Diese werden jedoch von kulturellen und sozialen Einwirkungen beeinflusst.

Das führt dazu, dass Zuhörern das Verstehen der rhythmischen Strukturen fremder Musikstile schwer fällt, wie am Beispiel von nordamerikanischen Zuhörern und klassischer indischer Musik gezeigt wurde [HT05]. Die prosodischen Besonderheiten der Muttersprache eines Komponisten können die Struktur seiner Werke beeinflussen, wie in Untersuchungen von klassischer Instrumentalmusik von britischen und französischen Komponisten des Neunzehnten und Zwanzigsten Jahrhunderts anhand von Vergleichen des *normalisierten paarweisen Variabilitätsindexes* empirisch gezeigt wurde [PD03].

Lehrdahl und Jackendoff gehen davon aus, dass „musikalische Kenntnis zu einem Teil durch das musikalische Umfeld erlernt oder angeeignet wird, zu einem anderen aber auch auf einer angeboren Prädisposition beruht“. Die Lernfunktion führt beispielsweise dazu, dass ein Zuhörer die Feinheiten eines komplexen Stückes beim ersten Hören weniger erfassen kann als nach einer Wiederholung [BG98]. Einen weiteren Hinweis auf die Konditionierung gibt die globale Aufzeichnung kortikaler Prozesse, die zeigt, dass die neuronale Aktivität beim Musikhören bei trainierten Musikern höher als bei Nicht-Musikern ist [Gol02].

Mechanismen der seriellen Gruppierung Die Gruppierung von auditiven Ereignissen erfolgt nach ähnlichen Prinzipien wie die visuelle Gruppierung: den Prinzipien der Gestaltpsychologie und der Gruppierung zu vertrauten und einfachen Konfigurationen [Bre90, Tan93, Deu99]. Nach Miller ist die Anzahl der in einer Einheit wahrgenom-

[11] In der deutschsprachigen Darstellung [BG98] von Lehrdahl und Jackendoffs Werk *Generative Theory of Tonal Music* (GTTM) wird dieser Begriff mit „formalen Regeln“ übersetzt.

[12] In [BG98] mit *Bevorzugungsregeln* übersetzt.

menen und im Kurzzeitgedächtnis gespeicherten Ereignisse auf fünf bis neun begrenzt [Wey01, Pos04].

Lerdahl und Jackendoff haben den Einfluss der Gestaltprinzipien anhand einfacher Tonfolgen untersucht. Nach dem Prinzip der Nähe werden in schneller Abfolge hintereinander erklingende Noten zu Gruppen zusammengefasst, die durch Pausen oder längere Noten begrenzt sind. Nach dem Prinzip der Ähnlichkeit werden Töne gleicher oder ähnlicher Tonhöhe, Dynamik oder Klangfarbe gruppiert. Große Intervalle, Pausen, Dynamik- und Klangfarbenunterschiede werden als Gruppierungsgrenzen aufgefasst.

Diese Prinzipien können in unterschiedlicher Intensität ausgeprägt sein, wobei der Grad der Ausprägung in Konfliktsituationen die Gruppierung leitet. Die Wahrnehmung einer Gruppierungsgrenze wird verstärkt, wenn zwei Prinzipien auf diese hinweisen, ebenso wie bei widersprechenden Prinzipien gleicher Intensität keine eindeutige Gruppierung wahrgenommen werden kann [LJ83]. Die genannten Prinzipien gelten für die Zusammenfassung von sequenziellen Ereignissen als auch für die Gruppierung von einzelnen Tönen zu Klanggemischen.

Die Gruppierung auf höherer Ebene wird weiterhin durch die Regeln der Intensivierung, Symmetrie und Parallelität geleitet. Nach der Intensivierungsregel müssen Gruppierungsgrenzen deutlicher ausgeprägt sein, um größere Gruppierungen zu bilden. Nach dem Prinzip der Symmetrie werden Ereignisse bevorzugt in ähnliche und gleichmäßige Gruppen eingeteilt. Die Regel der Parallelität besagt, dass die Unterteilung ähnlicher und größerer Gruppen parallel beziehungsweise ähnlich vollzogen wird.

Die Gruppierung unterschiedlicher Ereignisse wird auch *Objektive Rhythmisierung* genannt [Kru00]. In Experimenten von Royer und Garner wurden Testhörern sich wiederholende, aus zwei unterschiedlichen Ereignissen zusammengesetzte Sequenzen präsentiert. Die wahrgenommenen Muster begannen oder endeten mit einer Serie identischer Ereignisse [Moo82].

Wird einem Zuhörer eine Sequenz von unterschiedlichen Ereignissen einschließlich Pausen und Akzenten wiederholt vorgespielt, so wird diese Sequenz als begrenzt durch Pausen und Akzente wahrgenommen [Kru00, Fra82]. Pausen und Akzente haben stärkeren Einfluss auf die Gruppierung als die Musterstruktur [Moo82].

Wird eine Tonsequenz mit Tönen aus zwei unterschiedlichen Tonhöhenbereichen mit schnellem Tempo präsentiert, werden zwei unterschiedliche melodische Linien wahrgenommen [Bre90, Deu99]. Die Ereignisse werden dabei nach der Ähnlichkeit der Tonhöhe gruppiert. Dieser Effekt wird *stream segregation* genannt[13].

[13] Die Anwendung dieses Effektes in der Komposition wird Pseudopolyphonie genannt. Bei größeren Unterschieden in Tonhöhe oder Klangqualität muss der zeitliche Abstand erhöht werden, um den Eindruck einer verbundenen Serie zu erhalten. Unter bestimmten Umständen kann auch die Ähnlichkeit der Intensität der Ereignisse zur Gruppierung beitragen.

Subjektive Rhythmisierung Beim Hören einer Serie von äquidistanten und identischen Ereignissen mit moderatem Tempo wird ein Muster von Betonungen wahrgenommen, welches zu einer Gruppierung von jeweils zwei, drei oder vier Ereignissen führt. Bei moderatem Tempo liegen die Zeitintervalle zwischen den Ereignissen im Bereich des Echogedächtnisses, auch akustisches Ultrakurzzeitgedächtnis genannt. Ein verbreitetes Beispiel stellt die Wahrnehmung des gleichmäßigen Tickens einer Uhr als Abfolge zweier unterschiedlicher Ereignisse („Tick Tack“) dar. Dieser Effekt wird besonders bei anhaltendem Zuhören deutlich. Da keine objektive Ursache für die Gruppierung im Signal vorliegt, wird dieses Phänomen als *Subjektive Rhythmisierung* [Spi03, Kru00, NM99, Fra82] oder *beat grouping* [Bil93] bezeichnet.

3 Stand der Technik

Der in diesem Kapitel dargestellte Überblick über bestehende Verfahren zur Extraktion rhythmusbezogener Informationen aus Musiksignalen berücksichtigt in erster Linie die Verfahren, die Audiosignale auswerten. Die Analyse von symbolische Darstellungen, zum Beispiel MIDI-Noten (Musical Instrument Digital Interface) [The96], wird nur entfernt betrachtet.

Die im Abschnitt 3.1 betrachtete Detektion von Noteneinsätzen liefert wichtige Informationen für eine rhythmische Analyse. Verfahren zur Ermittlung des musikalischen Tempos und der Positionen der Zählzeiten werden in Abschnitt 3.2 dargestellt, die Analyse niedriger und höherer metrischer Ebenen in Abschitt 3.3 beziehungsweise 3.4.

Verfahren zur automatisierten Extraktion der Intensität und Komplexität von Rhythmen werden in Abschnitt 3.5 besprochen. Die Identifikation von perkussiven Instrumenten in polyphonen Audiosignalen ist Inhalt des Abschnittes 3.6. Weitere in der Literatur vorzufindende Merkmale sind in Abschnitt 3.7 dargestellt.

3.1 Detektion von Noteneinsätzen

Die Detektion von Noteneinsätzen spielt für die rhythmische Analyse von Audiosignalen eine zentrale Rolle. In verschiedenen Veröffentlichungen wird diese Funktionalität als *Segmentierung* bezeichnet [Sch85, GPD00, Hei05]. Falls nicht anders angegeben, bezeichnet der Begriff *Segmentierung* in dieser Arbeit die Unterteilung eines Musikstückes in charakteristische Abschnitte, die Strophe, Refrain oder ähnliches darstellen.

Für die in diesem Abschnitt betrachtete Problemstellung ist die Detektion von transienten Signalanteilen von Interesse, die in der Vergangenheit in großem Umfang untersucht wurde, zum Beispiel in [Hin90, RJ01, DMT01, KCZS03, VIS04]. Sie findet Anwendung in der Audiokodierung, der Spracherkennung, in Soundeffekten, bei der Modellierung und Resynthese von Signalen und in Überwachungssystemen.

In der Literatur wird zwischen perzeptuellem Noteneinsatzzeitpunkt (NEZ), das heißt dem Zeitpunkt der Wahrnehmung einer Note, und physischem NEZ, dem Zeitpunkt der Schallerzeugung, unterschieden [Sch85, Dix01c]. Die zeitliche Differenz zwischen perzeptuellem und physischem Noteneinsatz ist positiv und abhängig vom Anstieg der Hüllkurve des Signals.

Die durch die akustische Übertragungsstrecke bedingte Verzögerung zwischen physi-

schem und perzeptuellem Noteneinsatz[1] ist in der Regel klein und für alle Noten gleich. Da für die Ausprägung von Rhythmus die Intervalle zwischen den Noteneinsätzen von Bedeutung sind (für die hier der englischsprachige Begriff *inter onset intervals* (IOI) aufgrund seiner stärkeren Verbreitung in der Fachliteratur verwendet wird), wird diese Verzögerung vernachlässigt. Der durch das Hüllkurvenverhalten einer Note bedingte Unterschied zwischen physischen und perzeptuellem Noteneinsatz wird in dieser Arbeit nicht berücksichtigt, da in der Regel transiente Signale die Wahrnehmung von Rhythmus stärker prägen als nicht-transiente Signale.

Detektion im Hüllkurvensignal Ein von Schloss [Sch85] vorgeschlagenes Verfahren zur Transkription perkussiver Musik ermittelt Noteneinsätze aus Amplitudenhüllkurven des breitbandigen Audiosignals. Die Hüllkurven werden durch die Suche und das Verbinden der Maxima und Minima innerhalb kleiner, benachbarter Datenblöcke ermittelt. Durch lineare Regression von vier bis acht aufeinander folgenden Punkten werden Anstiege im Hüllkurvensignal gemessen. Noteneinsätze werden detektiert, wenn ein steiler Anstieg auftritt, und nach dem vorangegangenen detektierten Noteneinsatz eine festgelegte Zeitspanne vergangen ist. Das Verfahren arbeitet halbautomatisch, das heißt es können vom Nutzer Parameter verändert werden, bis ein zufriedenstellendes Ergebnis erreicht ist.

Energiebasierte Ansätze mit Frequenzbandseparation Die Analyse des breitbandigen Audiosignals ist jedoch nur Erfolg versprechend für die Detektion akzentuierter Noten und die Analyse monophoner Musik. Eine Verbesserung der Erkennungsleistung wird durch die Analyse von Teilbandsignalen erreicht. Die Frequenzbandseparation geschieht vorzugsweise mittels IIR-Filtern oder Diskreter Fouriertransformation (DFT).

Ein Verfahren von Klapuri [Kla99] separiert das Audiosignal in 21 Frequenzbänder zwischen 44 und 18000 Hz. Die Berechnung der Hüllkurve geschieht durch Einweggleichrichtung, Unterabtastung und Faltung mit einem halben Hanning-Fenster der Länge von 100 ms. Die relativen Differenzenfunktionen $D_i(t)$ der Hüllkurvensignale $E_i(t)$ mit Bandindex i werden nach Gleichung [3.1] ermittelt.

$$D_i(t) = \frac{\frac{d}{dt}E_i(t)}{E_i(t)} = \frac{d}{dt}\ln E_i(t) \tag{3.1}$$

Die Berechnung der relativen Differenzenfunktion ist motiviert durch das Weber'sche Gesetz. Noteneinsätze werden in $D_i(t)$ durch die Suche nach lokalen Maxima über einem Schwellwert detektiert und anschließend über alle Bänder kombiniert, so dass sich die aus den Teilbändern ermittelten Intensitäten der Note addieren.

[1] Die Ausbreitungsgeschwindigkeit von akustischem Schall in der Luft beträt ungefähr 343 m/s bei 20 Grad Celsius Lufttemperatur.

Problematisch bei dieser Vorgehensweise ist jedoch, dass die größten lokalen Maxima in $D_i(t)$ zu Zeitpunkten mit sehr kleinen $E_i(t)$ auftreten, bei denen das Weber'sche Gesetz nicht gilt.

Eine ähnliche Vorgehensweise beschreibt Seppänen. Anstelle der relativen Differenzenfunktionen $D_i(t)$ wird eine Detektionsfunktion nach Gleichung [3.2] berechnet [Sep01].

$$D_i(t) = \frac{E_i(t) - E_i(t-1)}{E_i(t) + E_i(t-1)} \tag{3.2}$$

In [MB96] wird der hochfrequente Energieanteil *HFE* durch lineare Wichtung der Frequenzbins $X(n,i)$ mit dem Binindex i zur Detektion transienter Anteile im Signal nach Gleichung 3.3 ausgewertet.

$$HFE(n) = \sum_{i=2}^{N/2+1} |X(n,i)|^2 \cdot i \tag{3.3}$$

Auswertung von Phaseninformation Die bisher vorgestellten Ansätze leisten jedoch keine robuste Detektion von unakzentuierten Noten in komplexen Klanggemischen. Methoden zur Detektion von Noteneinsätzen unter Berücksichtigung der Phaseninformation sind in verschiedenen Veröffentlichungen zu finden. Phasenspektrum $\phi(n,k)$ und Betragsspektrum $|X(n,k)|$ werden durch Umwandlung der komplexen Fourierkoeffizenten in Polarkoordinaten gewonnen. Wenn $x(n)$ ein stationäres Signal ist, sind die Phasendifferenzen zwischen benachbarten Blöcken konstant, und es gilt

$$\phi(n-1,k) - \phi(n-2,k) = \phi(n,k) - \phi(n-1,k) \tag{3.4}$$

Der princarg-Operator wandelt einen Winkel in das Intervall $[-\pi, \pi)$ um. Ist $x(n)$ kein stationäres Signal, kann eine Differenz der Phasenabweichungen zwischen benachbarten Blöcken

$$\Delta\phi(n,k) = \text{princarg}(\phi(n,k) - 2\phi(n-1,k) + \phi(n-2,k)) \tag{3.5}$$

berechnet werden. Diese wird in [DDS01] in Zusammenhang mit einer Multiskalenanalyse (engl. *multiresolution analysis*) mit frequenzabhängigen Blockgrößen zur Separation der transienten und sinusoidalen Anteile eines Signals ausgewertet. In [BS03] wird eine Detektionsfunktion aus der Kurtosis der Wahrscheinlichkeitsdichteverteilung der Phasenabweichung $\Delta\phi$ ermittelt. Auf einen Anstieg in der Phasenübereinstimmung zu Noteneinsätzen wird in [McD98a] hingewiesen.

Kombinierte Detektionsalgorithmen Ein kombinierter Ansatz, der den Verlauf der Phasenabweichung $\Delta\phi$ und Energiedifferenz ΔE zwischen benachbarten Blöcken in N

Frequenzbins auswertet, ist in [DBDS03a] beschrieben. Die Detektionsfunktion $f(n)$ wird nach Gleichung 3.6 ermittelt.

$$f(n) = \frac{1}{N-1}(\sum_{k=0}^{N-1} \Delta\phi_n \cdot \sum_{k=0}^{N-1} \Delta E_n) \tag{3.6}$$

Das in [DSD02] beschriebene Verfahren analysiert in den oberen Teilbänder ab 1200 Hz die Energiedifferenz und in den unteren Teilbänder die euklidische Distanz zwischen den komplexen Fourierkoeffizienten benachbarter Blöcke. Ein weiterer kombinierter Ansatz ist in [DBDS03b] veröffentlicht.

Auswahl des Schwellwertes Die bisher vorgestellten Verfahren basieren auf dem Vergleich von Detektionsfunktionen mit einem Schwellwert, geben jedoch wenig Einblick in dessen Auswahl. In [DBDS03a] und [DBDS03b] wird ein adaptiver Schwellwert aus dem Median eines gleitenden Fensters der Detektionsfunktion ermittelt. In [DSD02] wird ein globaler Schwellwert aus dem Histogramm der Detektionsfunktion abgeleitet. Klapuri verwendet einen festen Schwellwert, der durch Auswertung eines Trainingsdatensatzes ermittelt wurde [Kla99].

Weitere Methoden Die Anwendung Neuronaler Netze zur Detektion von Noteneinsätzen in Klaviermusik wird in [MKP02] vorgestellt. Das Audiosignal wird in zweiundzwanzig überlappende Bänder zerlegt, von denen für jedes Teilbandsignal aus den Differenzen zweier unterschiedlich stark geglätteter Hüllkurven die Merkmale zur Klassifikation berechnet werden.

Dixon extrahiert verschiedene Zeit- und Frequenzbereichsmerkmale zur Detektion von Noten in Klaviermusik. Zeitliche Abweichungen zwischen den lokalen Maxima der Merkmalsverläufe und manuell annotierten NEZ werden durch Korrelation berechnet und korrigiert. Ein genetischer Algorithmus ermittelt die Wichtungskoeffizienten der Merkmale und einen Detektionsschwellwert auf Grundlage von Trainingsdaten. Annähernd 90% der Noteneinsätze werden mit einer durchschnittlichen Abweichung von 10 ms detektiert [Dix01c].

In einem von Kapanci und Pfeffer vorgeschlagenen Ansatz werden benachbarte Signalausschnitte verglichen, wobei die Distanz zwischen den Ausschnitten bis zu einem Schwellwert schrittweise erhöht wird. Dies ist motiviert durch das Auftreten von Noten mit langer Attackphase und *legato* gespielten Noten. Dabei werden die Amplituden, die Grundfrequenzen und die Ausprägungen der ersten drei Harmonischen mittels *Support Vector Machines* ausgewertet. Die Evaluierung der Methode geschieht anhand einer Datenbasis von Exzerpten aus monophoner Vokalmusik [KP04].

Verschiedene Methoden zur Detektion von Zeitpunkten, in denen sich die harmonische Struktur eines Musiksignals ändert, sind in [HM03] präsentiert. Ein Verfahren von

Supper *et. al.* detektiert Noteneinsätze in Frequenzbändern, wenn die Fehlerfunktion eines Prädiktionsfilters größer als die verstärkte, tiefpassgefilterte Fehlerfunktion ist [SBR03]. In [McD98b] wird die spezielle Problematik der Detektion in Signalen mit verzerrten Gitarrenklängen behandelt.

Abdallah und Plumbley definieren einen Noteneinsatz als „Überraschungsmoment“, zu dessen Detektion ein „Überraschungsmaß“ aus aktuellem und vorhergehendem Signalblock unter Anwendung von *Independent Component Analysis* (ICA) ermittelt wird. Zur Auswertung und Detektion der Noten werden Hidden Markov-Modelle eingesetzt [AP03]

3.2 Temposchätzung und Beattracking

Das musikalische Tempo ist ein grundlegendes Merkmal von Musik. Temposchätzung heißt, den Verlauf des musikalischen Tempos aus dem Musiksignal zu ermitteln. Der Begriff *Beattracking* bezeichnet die Extraktion der zeitlichen Positionen der Zählzeiten.

Die Funktionalität des Beattracking ist von fundamentaler Bedeutung für das Hören von Musik und kann leicht von einem Zuhörer erfüllt werden [Dix97, Sch00]. Schwierigkeiten bei der Bewältigung dieser Aufgabe werden berichtet aus Beobachtungen mit polyrhythmischer Musik [Che94] und Musik mit expressivem, das heißt stark variierendem Tempo [DG02]. Untersuchungen von Drake *et. al.* zeigten, dass 89% von 18 Musikern und 75% von 18 Nichtmusikern die Fähigkeit zur spontanen Synchronisation innerhalb der ersten zehn Takte von verschiedenen klassischen Klavierstücken besaßen [DPB00].

Überblick Beattracking-Systeme (BTS) finden Anwendung in MIR-Systemen, der Mensch-Maschine-Kommunikation zur Musikproduktion, der automatischen Transkription, intelligenten Begleitautomaten, beim automatisierten Überblenden von Musik und bei der Synchronisation von multimedialen Inhalten zu Musik. Die große Anzahl potentieller Applikationen führte in der Vergangenheit zu umfangreichen Forschungsbemühungen, die in einer Vielzahl an Publikationen unterschiedlichster Verfahren mündeten. Ein Überblick ist in [Tem04, GD05] gegeben. Diese Verfahren werten entweder Audiosignale oder symbolische Repräsentationen, zum Beispiel MIDI-Daten [DH89, AD90, Ros92, Rap01, TNS03], aus. Erstere Verfahren können als merkmalsbasiert oder ereignisbasiert kategorisiert werden. Merkmalsbasierte Verfahren werten signalnahe Merkmale aus, ereignisbasierte Verfahren beginnen die Verarbeitung von Audiosignalen mit der Detektion von Klängen.

Der ISMIR Audio Description Contest als weltweit erster offener Vergleich bestehender Verfahren zur maschinellen Musikanalyse[2] zeigte, dass merkmalsbasierte Verfah-

[2] Weiterhin wurden in den Disziplinen Genre-Klassifikation, Künstler-Identifikation, Rhythmus-Klassifikation und Melodie-Erkennung verschiedene Systeme verglichen [CGG+06]. Dieser Wettbewerb ermöglicht den Vergleich der Verfahren anhand eines identischen Testapparates und bietet so größere Aussagekraft als von Autoren individuell präsentierte Ergebnisse.

ren im Mittel bessere Erkennungsleistungen erreichen und größere Robustheit gegenüber Signalmodifikationen, zum Beispiel Filtern und Überlagerung mit Rauschen oder künstlichem Hall) aufweisen als ereignisbasierte Verfahren [GAD+06].

Scheirer demonstrierte in einem psychoakustischen Experiment [Sch98], dass die aus Teilbändern extrahierten Pegelinformationen zur Detektion des Tempos und der Zählzeiten ausreichend sind. Dazu wurde ein Musiksignal in sechs Frequenzbänder zerlegt und die Amplitudenhüllkurven in den einzelnen Kanälen berechnet, differenziert und einweggleichgerichtet. Ein breitbandiges Rauschen wird ebenso in Frequenzbänder zerlegt, mit den aus dem Musiksignal extrahierten differenzierten Amplitudenhüllkurven moduliert und zusammengemischt[3]. Das resultierende Signal vermittelt den gleichen rhythmischen Eindruck wie das Musiksignal. Die Amplitudenhüllkurven der Teilbandsignale werden durch Zweiwegleichrichtung, Glättung mit einem FIR-Tiefpassfilter und Unterabtastung berechnet.

Zur automatisierten Analyse von Tempo und Beat werden aus den Amplitudenhüllkurven durch Differenzierung und Einweggleichrichtung *Akzentsignale* berechnet, welche lokale Maxima zu Zeitpunkten positiver Anstiege der Hüllkurve aufweisen, die mit phänomenalen Akzenten korrelieren. Die Akzentsignale der Teilbänder werden zu einem Akzentsignal aufsummiert und Periodizitäten im zeitlichen Verlauf des Akzentsignals mit Hilfe einer Resonanzfilterbank detektiert. Die Phasenlage des Beats wird aus dem Ausgangssignal des zum ermittelten Tempo korrespondierenden Resonators oder dem Inhalt seiner Verzögerungsschleife direkt ermittelt [Sch98].

Scheirers Ansatz wurde von verschiedenen Autoren weiterentwickelt. Die Modifikationen beinhalten die Untersuchung verschiedener Merkmale zur Generierung des Akzentsignals, alternative Verfahren der Periodizitätenschätzung und der Auswertung der Periodizitäten zur Schlussfolgerung auf das Tempo.

Akzentsignal Zur Zerlegung des Signals in Teilbänder werden FIR- und IIR-Filterbänke [Sch98, PK02] oder die DFT [Kla03] eingesetzt. Die grundsätzliche Äquivalenz von Teilbandzerlegungen und Transformationen ist in [Edl95] dargelegt. Weiterhin finden Diskrete Wavelettransformationen zur Repräsentation nichtstationärer Signale Anwendung [TEC01a, TEC01b]. Methoden zur Berechnung des Akzentsignals in Teilbändern sind die relative Differenzenfunktion [Kla99] (siehe Abschnitt 3.1, Gleichung 3.1), die modifizierte relative Differenzenfunkion nach Seppänen (siehe Gleichung 3.2), einfache Differenzierung, differenzierte Lautheit (Loudness) [KEA05] oder die Verwendung von IIR-Hochpassfiltern.

Eine Alternative zu bandweisen differenzierten Hüllkurvensignalen ist die Verwendung von Noteneinsätzen als Akzentsignal. In [Fri04] werden „gaussifizierte“ Noteneinsätze als Eingangssignal zur Tempo- und Taktschätzung benutzt. Gaussifizierung bezeichnet hier

[3] Diese Verarbeitung entspricht bis auf die Differenzierung der Hüllkurvensignale der des aus der Musikproduktion und Sprachsynthese bekannten *Vocoders*.

die Faltung einer Impulsfolge mit der Gauß'schen Funktion (siehe Gleichung 3.7).

$$g(t, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{3.7}$$

Die Gaussifizierung entspricht einer Tiefpassfilterung und ist für eine Vielzahl von Verfahren ein notwendiger Schritt zur Analyse auf Grundlage von Noteneinsätzen in Musik mit Temposchwankungen.

In [ABDR03, PK02] wird ein Sinusoidal-und-Noise-Modell [Ser97] vom Musiksignal erstellt. Zur rhythmischen Analyse wird nur der Rauschanteil des Modells verwendet, um transiente und nichtstationäre Signalanteile stärker zu gewichten.

Ein recheneffiziente Extraktion eines Akzentsignals ist der von Laroche vorgeschlagene spektrale Energiefluss (engl. *energy flux*) E_{flux}

$$\widehat{E}(n) = \sum_{f=f_{min}}^{f_{max}} G(|X(f,n)|) - G(|X(f,n-1)|) \tag{3.8}$$

$$E_{flux}(n) = \begin{cases} \widehat{E}(n) & | \quad \widehat{E}(n) > 0 \\ 0 & \text{sonst} \end{cases} \tag{3.9}$$

mit den Fourierkoeffizienten $X(f,n)$ und einer nichtlinearen Funktion $G(x)$ zur Kompression des Dynamikbereiches. Zur Kompression können logarithmische Funktionen, die Potenzfunktion $G(x) = x^n, 0 < n < 1$ oder die inverse hyperbolische Sinusfunktion $G(x) = \text{arcsinh}(x)$ eingesetzt werden [Lar01, Lar03]. In [KEA04] werden die Energiehüllkurven nach Gleichung 3.10 μ-law komprimiert.

$$y_b(k) = \frac{ln(1 + \mu x_b(k))}{ln(1+\mu)}, \qquad \mu = 100 \tag{3.10}$$

Goto und Muraoka publizierten verschiedene echtzeitfähige BTS. In [GM94] wird das Auftreten von Bassdrum und Snaredrum[4] detektiert und mit angelernten Drumpattern verglichen. Weitere Veröffentlichungen der Autoren adressieren die Analyse von Musiksignalen, in denen keine Schlagzeuginstrumente auftreten. In [GM97b, Got98] wird die Detektion von Akkordwechseln zur rhythmischen Analyse vorgeschlagen. Eine Kombination der Verfahren ist in [Got01] beschrieben. Ein in [GM96] dargestelltes Verfahren clustert detektierte Ereignisse in drei Gruppen, die Klänge mit unterschiedlichen spektralen Eigenschaften repräsentieren.

Die Verfahren von Goto und Muraoka setzen eine Reihe von Annahmen voraus. Die Taktart der Musikstücke wird als 4/4-Takt angenommen, das Tempo ist konstant und

[4] Die englischsprachigen Begriffe *Bassdrum* für Große Trommel und *Snaredrum* für Kleine Trommel werden hier auf Grund ihrer größeren Verbreitung im Vergleich zu den deutschsprachigen Namen verwendet.

liegt in einem festen Bereich, zum Beispiel zwischen 70 und 180 bpm [GM94], 65 und 185 bpm [GM95] oder 61 und 120 bpm [GM96].

Ausgehend von der Annahme, dass dynamische Akzente mit Transienten im Audiosignal zusammentreffen, untersuchen Jensen und Anderson neben dem Amplituden- und Energieverlauf weitere *Low-level*-Merkmale[5], zu denen der spektrale Zentroid (ein zur wahrgenommenen Helligkeit eines Klanges korrelierendes Maß), der spektrale Energiefluss (indiziert transiente Signalanteile) und die spektrale Irregularität (die ähnlich des Spektralen Flachheitsmaßes (SFM) die Rauschhaftigkeit beziehungsweise Tonalität des Signalausschnittes misst) gehören [JA03]. Die Anwendung der SFM und energiebasierten Merkmalen wurde weiterhin in [GH03b] untersucht.

Periodizitätenschätzung Die Verfahren von Scheirer und Klapuri verwenden zur Periodizitätenschätzung eine Bank von Resonanz-Kammfiltern [Sch98, Kla03]. Die Differenzengleichung des rekursiven Filters ist in Gleichung 3.11 angegeben.

$$y(t) = \alpha y(t-\tau) + (1-\alpha)x(t), \quad 0 < \alpha < 1 \tag{3.11}$$

Durch geeignete Wahl des Parameters τ wird jedes Filter auf ein Tempo eingestellt. Der Parameter α, mit $|\alpha| < 1$, wird so gewählt, dass alle eingeschwungenen Resonatoren die gleiche Leistung ausgeben.

Verschiedene Verfahren applizieren zur Periodizitätenschätzung die Autokorrelationsfunktion (AKF, siehe Gleichung 3.12), die von Akzentsignalen in einzelnen Frequenzbändern oder von über die Bänder aufsummierten Akzentsignale berechnet werden [TEC01a, DPG03, UH03, DP04, Fri04].

$$r_{xx}(\tau) = \frac{1}{N} \sum_{n=0}^{N-\tau} x(n)x(n+\tau) \tag{3.12}$$

Nach einer effizienteren Berechnungsvorschrift für Vektoren mit einer Länge $l > 64$ wird die AKF nach dem Wiener-Khinchin-Theorem durch DFT und Rücktransformation der quadrierten Beträge der Spektralkoeffizienten ermittelt (siehe Gleichung 3.13).

$$r_{xx}(\tau) = \text{IDFT}\{|\text{DFT}(x(n)\,|^2\} \tag{3.13}$$

Weiterhin ist die ursprünglich zur Tonhöhenermittlung vorgeschlagene *Average Magnitude Difference Function* (AMDF) [CK03] beziehungsweise *Squared Difference Function* (SDF) [MW05] verbreitet,

$$d(\tau) = \sum_{n=0}^{N-\tau} (x(n) - x(n+\tau))^2 \tag{3.14}$$

[5] *Low-level*-Merkmale sind signalnahe Merkmale, die direkt ohne interpretative Verarbeitung aus dem Signal gewonnen werden. Im Unterschied zu *High-level*-Merkmalen besitzen sie keine semantische Bedeutung und weisen eine höhere Datenrate auf.

sowie die *Cumulative Mean Normalized Difference Funktion* (CMNDF, siehe Gleichung 3.15) [CK03], die zum Beispiel in [PK02] zur rhythmischen Analyse eingesetzt wird.

$$d^{'}(\tau) = \begin{cases} 1 & | \quad \tau = 0 \\ d(\tau) \Big/ \frac{1}{\tau} \sum_{j=0}^{\tau} d(j) & \text{sonst} \end{cases} \tag{3.15}$$

Tzanetakis *et. al.* prägten den Begriff *Beathistogramm* zur Beschreibung der Zusammenfassung der n größten lokalen Maxima der AKF für überlappende Blöcke in einem Histogramm der Periodendauern [TEC01b, TEC02].

Bei der Verwendung von Noteneinsätzen als Akzentsignal kann eine Periodizitätenfunktion aus einem Histogramm der IOI ermittelt werden, wobei die Berechnung der IOI nicht auf direkt benachbarte Noteneinsätze beschränkt ist [DGW02, DPG03]. Das ermittelte IOI-Histogramm wird zur Weiterverarbeitung in der Regel mit einem Tiefpassfilter geglättet.

Sethares und Slaney verwenden eine in [SS99] entwickelte Periodizitätentransformation zur Ermittlung von Tempo und Taktart für Signale mit konstantem Tempo [SS01]. Foote und Uchihashi [FU01] und Pikrakis et. al. [PAT04] adaptieren die ursprünglich zur Audiosegmentierung vorgeschlagene Selbstähnlichkeitsmatrix [Foo00] zur rhythmischen Analyse.

Eine Zusammenstellung verschiedener Verfahren der Periodizitätenschätzung zur Anwendung für Temposchätzung enthält [ADR03], darunter die Autokovarianzfunktion (Gleichung 3.16), sowie Produkt und die Summe der von den Akzentsignalen je Teilband ermittelten Betragsspektren.

$$a(\tau) = \sum_{n=0}^{N} [x(n) - \overline{x}]\,[x(n+\tau) - \overline{x}] \tag{3.16}$$

Auswertung der Periodizitätenfunktion Scheirer berechnet das musikalische Tempo aus der Frequenz des Resonators mit maximaler Ausgangsenergie [Sch98]. Im Verfahren nach Dixon ermittelt ein Clustering-Algorithmus eine Liste von Tempo-Hypothesen aus IOI, die mit der Anzahl ihrer Clusterelemente bewertet werden [Dix01a].

Die Zuordnung des Maximums des zu musikalischen Tempi korrespondierenden Bereichs einer Periodizitätenfunktion zur Beatperiode führt jedoch bei stark synkopierten Rhythmen zu falschen Ergebnissen, da die stärkste Periodizität der Dauer der punktierten Note entsprechen kann.

Ein Beispiel eines synkopierten Rhythmus, der *Son-Clave* (siehe Abbildung 2.1 in Abschnitt 2.1), illustriert dieses Problem. Synkopierte Rhythmen dieser Art stellen eine besondere Schwierigkeit für BTS dar [CK02].

Die AKF eines Akzentsignals der *Son-Clave* (siehe Abbildung 3.1) zeigt ein der Beatperiode entsprechendes lokales Maximum bei einem Zeitabstand von 0.6 s. Das globale

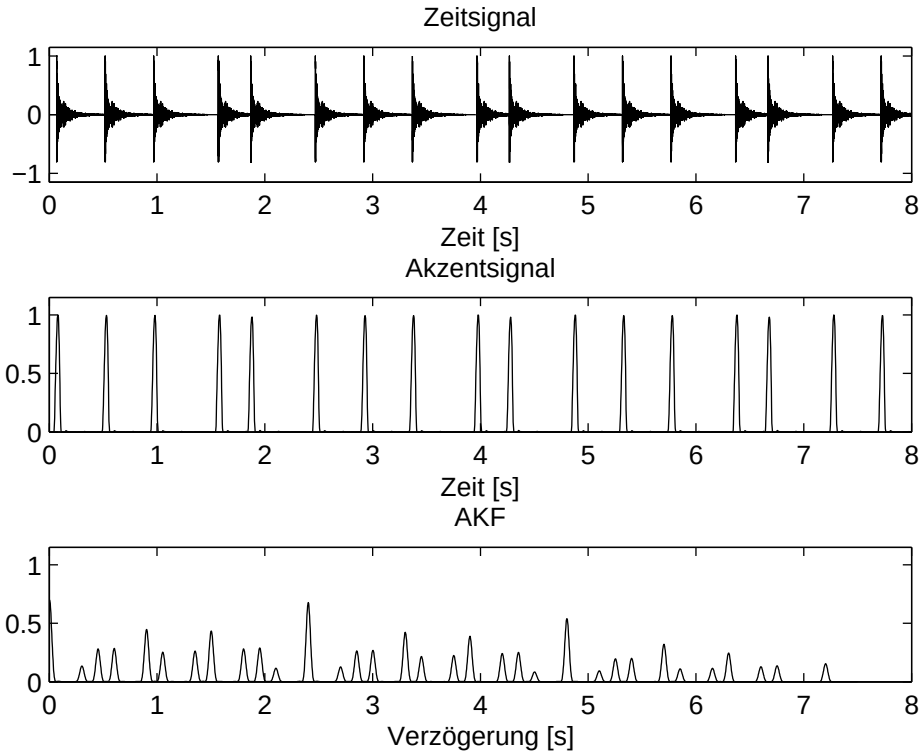


Abbildung 3.1: Zeitsignal, Akzentsignal und AKF der *Son-Clave* mit einem Tempo von 100 bpm.

Maximum entspricht der stärkeren Periodizität einer Taktlänge (2.4 s). Das zweitgrößte lokale Maximum verweist auf die Länge einer punktierten Viertelnote (0.9 s). Wenn das zur Taktlänge korrespondierende Tempo nicht im Suchbereich enthalten ist, wird das geschätzte Tempo 66% des Referenztempos betragen. Um eine robuste Temposchätzung für synkopierte Rhythmen zu erhalten, müssen neben der Intensität der Periodizität weitere Randbedingungen berücksichtigt werden, die aus dem grundsätzlichen Aufbau von metrischen Strukturen abgeleitet werden können.

Dazu wird eine kombinierte Schätzung mehrerer Pulsschichten angewendet, die die Tatsache ausnutzt, dass Beat-, Tatumperiode und Taktlänge in annähernd ganzzahligen Verhältnissen zueinander stehen. In [PK02, Kla03] wertet ein probabilistisches Modell die ermittelten Periodizitäten aus: Die A-priori-Wahrscheinlichkeit für relative Beatperioden, die dem zwei-, vier- oder sechsfachen der Tatumperiode entsprechen, ist größer als für das von fünf- oder siebenfache. A-priori-Wahrscheinlichkeiten für die absolute Beatperiode werden aus [Par94] und [NM99] (siehe Abschnitt 2.1) übernommen. In ei-

nem späteren Verfahren von Klapuri [KEA04] ist das probabilistische Modell durch ein Hidden-Markov-Modell ersetzt.

Das in [UH03] vorgeschlagene Verfahren schätzt die Tatumperiode aus einem Histogramm der IOI. Alle ganzzahligen Vielfachen (GZV) der Tatumperiode werden als Beatkandidaten und alle GZV der Beatkandidaten als Kandidaten für die Taktlänge ermittelt. Als Kriterium für die Entscheidung über die Taktlänge wird die AKF des Akzentsignals berechnet. Dieses Kriterium wird anschließend zur Entscheidung zwischen den zur ermittelten Taktlänge passenden Beatkandidaten ausgewertet.

Tracking In [Lar03] werden die Positionen der Zählzeiten durch Minimierung der euklidischen Distanz zwischen dem extrahierten und einem für binäre Rhythmen erwarteten Energiefluss ermittelt. Die Distanzberechnung geschieht für alle Tempi $v = 60...150$ bpm in 1-bpm-Schritten und alle Phasenlagen des mit einer Abtastfrequenz von 100 Hz vorliegenden Akzentsignals.

Goto und Muruoka setzen eine *multiple agent architecture* zu parallelen Verfolgung verschiedener Hypothesen ein [GM94, GM95, GM96, GM97b, Got98, Got01]. Ein *Agent* ist definiert als eine Softwarekomponente, die mit anderen interagiert, ihr Verhalten auf Grundlage des Eingangssignals evaluiert und anpasst und zu einer Beatperiode und -phase korrespondiert. Die Verfolgung verschiedener Hypothesen in Beattracking-Systemen ist weiterhin in den Systemen aus [AD90, RGM94, Dix00, Meu02b] angewendet.

Weitere Ansätze Pardo entwickelte ein BTS basierend auf einem Phase und Frequenz anpassenden Oszillator [Par04]. Cemgil *et. al.* verwenden ein Kalman-Filter-Modell zur Schätzung des Tempoverlaufs [CKDH01] und eine sequenzielle Monte Carlo-Methode zu Temposchätzung und Beattracking in symbolischen Darstellungen [CK02].

Ein von Gouyon vorgeschlagenes Verfahren ermittelt Beatperiode und -phase ausgehend von einem automatisiert extrahierten Tatumraster (siehe Abschnitt 3.3). Aus blockweise extrahierten Low-level-Merkmalen werden für jedes Tatumsegment die varianznormierten Mittelwerte berechnet, die zur AKF-basierten Ermittlung des Tempos und zur Ermittlung der Phase durch Vergleich mit einem Modellraster verwendet werden [GH03a].

In [SMS05] wird ein Akzentsignal aus Änderung des Energieverlauf, der Phaseninformation, des spektralen Schwerpunktes und der spektralen Spreizung extrahiert. Das Trackingverfahren basiert auf der Annahme, dass die Varianz des Akzentsignals zu Offbeat-Zeitpunkten kleiner als die Varianz des Akzentsignals zu Beat-Zeitpunkten ist.

Die Auswertung der Ergebnisse verschiedener Verfahren wird in [GAD$^+$06] unter dem Begriff „redundante Analyse“ diskutiert. Ein Abstimmungsverfahren dient zur Ermittlung des Tempos aus den Schätzergebnissen von fünf Verfahren, die aus einer Gesamtheit von elf Verfahren ausgewählt werden. Die durch umfassende Suche ermittelte beste

Kombination von Verfahren erzielte keine signifikante Verbesserung gegenüber dem besten Einzelverfahren. Bei Ausschluss des besten Verfahrens wurde eine signifikante Steigerung der Erkennungsleistung durch das Abstimmverfahren erreicht. Die Autoren vermuten eine weitere Verbesserung durch die Analyse der Stärken und Schwächen der einzelnen Verfahren und deren Berücksichtigung bei der Zusammenfassung der Ergebnisse.

3.3 Ermittlung des Tatumrasters

Gouyon adaptierte zur Ermittlung des Tatumrasters ein Verfahren zur Bestimmung der Grundfrequenz von Musik- und Sprachsignalen, die *Two-way mismatch error procedure* (TWME, [MB94]). Aus einem geglätteten Histogramm der IOI h wird das am häufigsten vorkommende IOI ioi_h zur Berechnung der Kandidaten für die Tatumperiode k_i (siehe Gleichung 3.17) ermittelt.

$$k_i = \frac{ioi_h}{d_i}, \quad \forall\, d_i \in \{1, 2, 3, 4, 6, 8, 9\} \tag{3.17}$$

Die Auswahl der Tatumperiode geschieht in Abhängigkeit von einem Fehlermaß e (siehe Gleichung 3.18).

$$e = \frac{e_{pm}}{N} + s \cdot \frac{e_{mp}}{K} \tag{3.18}$$

Der Faktor s dient der Gewichtung der beiden nachfolgend beschriebenen Teilfehlern. Der Teilfehler e_{pm} wird aus den Abweichungen Δf_n zwischen einer Folge $f_n = n \cdot p_t$ mit $n = 1...N$ und dem zu f_n nächstgelegenen lokalen Maximum des Histogramms h_n ermittelt, mit den empirisch ermittelten Parametern p, q, r und dem maximalen Histogrammeintrag $h_{max} = \max(h)$.

$$e_{pm} = \sum_{n=1}^{N} \Delta f_n \cdot f_n^{-p} + \frac{h_n}{h_{max}} \cdot (q \Delta f_n \cdot (f_n)^{-p} - r) \tag{3.19}$$

Der Teilfehler e_{mp} wird berechnet aus den zwischen allen lokalen Maxima des Histogramms h_k, $k = 1...K$, und den nächstgelegenen GZV der Tatumperiode f_k auftretenden Abweichungen Δf_k.

$$e_{mp} = \sum_{k=1}^{K} \Delta f_k \cdot f_k^{-p} + \frac{h_k}{h_{max}} \cdot (q \Delta f_k \cdot (f_k)^{-p} - r) \tag{3.20}$$

Das Tatumtracking geschieht durch umfassende Suche des kleinsten TWME aller Phasenlagen, wobei das Fehlermaß zwischen Rasterhypothese und NEZ berechnet wird. Dabei wird zusätzlich die Gridperiode leicht variiert, um eine genauere Schätzung der Periode zu erhalten [GHC02].

Ein Verfahren von Seppänen schätzt die Tatumperiode als größten gemeinsamen Teiler (GGT) aller IOI [Sep01]. Abweichungen der IOI von ganzzahligen Vielfachen der Tatumperiode p_t werden durch Berechnung einer Fehlerfunktion $e(p_t)$ nach Gleichung 3.21 berücksichtigt, mit IOI-Histogramm h, der Anzahl der Histogrammbins M und der Rundungsfunktion zur kleineren Ganzzahl $y = \lfloor x \rfloor$.

$$e(p_t) = \sum_{i=0}^{M-1} h(k) \left(\frac{h(k)}{p_t} - \left\lfloor \frac{h(k)}{p_t} + \frac{1}{2} \right\rfloor \right)^2 / \sum_{i=0}^{M-1} h(k) \tag{3.21}$$

Die Tatumphase wird durch Minimierung der Abweichung von Rasterelementen zu NEZ ermittelt.

Paulus [PK02] verwendet die gewichtet aufsummierten *CMNDF* von Hüllkurvensignalen in Teilbändern $x(\tau)$, mit zur Stärke der Periodizität im Teilband korrespondierenden Gewichten. Die Tatumperiode wird aus der Frequenz f berechnet, für die die Funtion $g(f) = \sqrt{f} \cdot |X(f)|$, mit $1.7Hz < f < 20Hz$ und der diskreten Fouriertransformierten $X(f)$ der *CMNDF* $x(\tau)$, ein globales Maximum annimmt.

3.4 Analyse höherer metrischer Ebenen

Palmer und Krumhansl [PK90] analysierten verschiedene Musikstücke westlicher tonaler Musik und schlussfolgerten, dass die Anzahl von Noten zum Beginn eines Taktes konzentriert ist. Das in [GM96] beschriebene Verfahren trifft die Annahme, dass Taktanfänge durch Akkordwechsel gekennzeichnet sind.

In [Bro98] wurde die AKF eines Vektors berechnet, der aus mit der Notenlänge gewichteten Amplitudenwerten der Noten einer monophonen melodischen Stimme zu NEZ gebildet wurde. Es wurde gezeigt, dass die resultierende AKF lokale Maxima korrespondierend zur Beatperiode, Länge eines Taktes und anderen metrischen Ebenen aufweist. In [Meu02a] wird eine hierarchische Struktur von akzentuierten Zählzeiten aus einer zeitlich quantisierten MIDI-Repräsentation ermittelt, aus der eine metrische Struktur durch eine AKF-basierte Methode ermittelt wird. Die Anwendung von Fuzzy-Logik zur Erkennung rhythmischer Gruppen in MIDI-Signalen ist in [DW03] beschrieben. Gouyon untersuchte den zeitlichen Verlauf verschiedener signalnaher Merkmale zur Unterscheidung zwischen binären und ternären Metren [GH03b].

Die in Abschnitt 3.2 genannten Verfahren aus [SS01, DPG03, Kla03, KEA04, UH03, PAT04, Fri04] ermitteln neben der Beatperiode die metrische Struktur auf verschiedenen Ebenen. Die Mehrzahl dieser Verfahren berücksichtigt den grundlegenden ganzzahligen Zusammenhang zwischen Periodendauern auf unterschiedlichen metrischen Ebenen.

3.5 Rhythmische Intensität und Komplexität

Intensität

Die automatische Extraktion eines die Intensität eines Musikstückes beschreibenden Maßes fand in jüngster Vergangenheit verstärkt Beachtung. Obwohl eine einheitliche Definition dieses Merkmals nicht vorliegt, und musikalische Intensität von Zuhörern unterschiedlich empfunden wird [ZP03], ist das Konzept „Intensität“ bei der umgangssprachlichen Beschreibung von Musik gebräuchlich und dadurch für MIR-Systeme von potentiellem Interesse.

In [ZP03] wird die Eignung verschiedener Low-level-Deskriptoren zur Extraktion der musikalischen Intensität untersucht. Als relevanteste Merkmale wurden spektrale Schiefe (spektrales Moment dritter Ordnung) und logarithmierte Varianz der Ableitung der Energie mit einem Korrelationskoeffizienten von 0.56 beziehungsweise 0.57 zwischen Merkmal und Trainingsdatensatz ermittelt. Eine Kombination von Merkmalen erreichte einen maximalen Korrelationskoeffizienten von 0.87. Weiterhin wurde das am *Sony Computer Science Laboratory Paris* entwickelte *Extractor Discovery System* (EDS) zur automatischen Ermittlung eines zur Intensität korrelierenden Merkmals eingesetzt. Das EDS basiert auf Genetischer Programmierung und wurde speziell für die Entwicklung musikalischer High-level-Deskriptoren entwickelt. Die mit EDS ermittelten Merkmale zeigen teilweise große Ähnlichkeit zu den relevanten Low-level-Merkmalen mit verbesserter Korrelation (bis 0.69) durch zusätzliche spezifische Vorverarbeitung. Die Kombination aller Merkmale erreichte eine Korrelation von 0.89.

Tzanetakis *et. al.* ermittelten verschiedene Deskriptoren der rhythmischen Intensität aus Beathistogrammen (siehe Abschnitt 3.2) [TEC01b]. In [TEC02] sind Untersuchungen zur menschlichen Wahrnehmung und automatisierten Berechnung der „Beatstärke“ (engl. *beat strength*, BS) präsentiert, wobei die BS von den Autoren definiert wird als eine rhythmische Charakteristik *“that could allow to discriminate between two pieces of music having the same tempo. Using this definition we might say that a peice of Hard Rock has a higher beat strength than a piece of Classical Music at the same tempo”*. In einem Hörtest stuften die Teilnehmer die BS von Musikstücken[6] auf einer kontinuierlichen Skale zwischen 1 und 5 ein. Die Ergebnisse des Hörtests zeigen eine Übereinstimmung zwischen den Aussagen der Testhörern mit einer mittleren Standardabweichung von 1.25 für alle Musikstücke. Zur automatischen Extraktion der BS werden zwei aus dem Beathistogramm extrahierte Maße untersucht: die Summe aller Histogrammbins und das Verhältnis von globalem Maximum und Mittelwert der Histogrammeinträge. Zur Evaluierung wurde ein Testdatensatz von 50 Stücken mit einer Länge von jeweils 15 s und von 32 Testhörern ermittelten Referenzwerten zwischen 1 und 5 verwendet. Die extrahierten Merkmale wiesen nach einer Umwandlung des Wertebereiches eine mittlere ab-

[6] Den Teilnehmern des Hörtests wurde keine Definition der BS vorgegeben.

solute Differenz von 1.12 beziehungsweise 1.08 auf.

In [BL03] werden weitere aus einem Beathistogramm ermittelte Merkmale zur Beschreibung der BS und der rhythmischen Gleichmäßigkeit im Kontext der Klassifikation des musikalischen Genres untersucht. Zur Bestimmung der BS wurden statistische Momente des Beathistogramms und seiner Ableitung ermittelt. Rhythmische Gleichmäßigkeit liegt vor, wenn die lokalen Maxima des Beathistogramms in regelmäßigen Abständen auftreten. Zur Berechnung wird die Auswertung der AKF des Beathistogramms vorgeschlagen. Die Größe der lokalen Maxima der AKF korreliert mit der Gleichmäßigkeit.

Komplexität

Komplexität ist, ebenso wie Rhythmus, ein je nach Forschungsgebiet unterschiedlich definierter Begriff und wird häufig als das Gegenteil von Einfachheit aufgefasst. In der Informationstheorie wird die Komplexität von Daten nach Shannon durch ihren Informationsgehalt bestimmt. Die Komplexität eines Problems ist durch den Ressourcenverbrauch eines optimalen Algorithmus zur Lösung des Problems definiert und ist unter dem Begriff *Kolmogorov*-Komplexität bekannt.

Der Kognitionswissenschaftler Pressing ergänzt diese Darstellung um zwei weitere Aspekte von Komplexität, die nicht notwendigerweise voneinander unabhängig sind [Pre04]. Die *hierarchische* Komplexität verweist auf die Existenz von Struktur auf hierarchischen Ebenen. Die *dynamische* beziehungsweise *adaptive* Komplexität definiert Pressing wie folgt:

> *"Systems that show a rich range of behaviours over time, or adapt to unpredictable conditions, or monitor their own results in relation to a reference source, or can anticipate changes in self or environment, we may take to be complex."* [Pre04]

Goto verweist auf die Bedeutung der Berechnung eines Maßes der „rhythmischen Schwierigkeit“ zur Evaluierung von BTS [GM97a]. Dieses Merkmal ist mehrdimensional und wird nach Goto neben der Komplexität der rhythmischen Muster, Taktart, Instrumentierung und dem Auftreten von Tempoänderungen von der Stärke und Häufigkeit von Synkopationen bestimmt. Zur quantitativen Beschreibung von Synkopationen eines Musikstückes werden die lokalen Maxima der Energie zu Zählzeiten L_n^b und Zwischenzählzeiten L_n^o nach Gleichung 3.22 und 3.23 mit n-ter Zählzeit C_n, n-tem Beatintervall I_n, Energie des Audiosignals $a(t)$ und einer Zeitspanne $\epsilon = 23\text{ms}$ ermittelt.

$$L_n^b = \max_{C_n - \epsilon \leq t < C_n + I_n/4 - \epsilon} (a(t)) \tag{3.22}$$

$$L_n^o = \max_{C_n + I_n/4 - \epsilon \leq t < C_{n+1} - \epsilon} (a(t)) \tag{3.23}$$

Weiterhin wird ein „Energie-Differenz-Maßes“ EDM nach Gleichung 3.24 mit Anzahl der Zählzeiten N und der normalisierten Energiedifferenz d_n nach Gleichung 3.25 berechnet.

$$EDM = \frac{1}{N-1} \sum_{n=1}^{N-1} d_n \tag{3.24}$$

$$d_n = 0.5 \frac{L_n^o - L_n^b}{\max(L_n^o, L_n^b)} + 0.5 \tag{3.25}$$

Eine quantitative Bewertung der Übereinstimmung zwischen „Energie-Differenz-Maßes“ EDM und subjektiven Einschätzungen der „rhythmischen Schwierigkeit“ wurde nicht gegeben.

Zur Beschreibung der Komplexität von quantisierten Rhythmen, deren symbolische Repräsentation vorliegen, wurden Berechnungen von Maßzahlen von Lempel und Ziv [LZ76], Tanguiane [Tan93] sowie Shmulevich und Povel [SP98a] vorgeschlagen. Eine vergleichende Darstellung dieser Maße ist in [SP98a, SP98b] gegeben.

Das Kompexitätsmaß nach Lempel und Ziv, welches als im Vorfeld des von den Autoren entwickelten Datenkompressionsverfahren für Zeichenketten vorgestellt wurde [LZ77], berechnet sich aus der Anzahl von unterschiedlichen Teilfolgen, aus denen die endliche Sequenz zusammengesetzt ist und die beim Lesen der Sequenz von links nach rechts auftreten. Tanguianes Komplexitätsmaß wird als maximale Anzahl von „root pattern“ berechnet, in die das betrachtete rhythmische Muster zerlegt werden kann [Tan93]. Das Komplexitätsmaß nach Shmulevich und Povel baut auf dem von Povel und Essens vorgeschlagenem Maß der „induction strength of the best clock“ auf, welches als gewichtete Summe der zu Zählzeiten auftretenden unakzentuierten Noten und Pausen berechnet wird. Dieses Maß ist klein, wenn die Zählzeit („best clock“) durch viele Akzente deutlich wahrnehmbar ist und wird als *C-Score* bezeichnet. Der *C-Score* wird einerseits ergänzt durch Komplexitätswerte, die für von den Autoren festgelegte Segmenttypen (siehe Tabelle 3.1) vordefiniert sind, und andererseits durch Werte ergänzt, die die Ähnlichkeit von aufeinander folgenden Segmenten bewerten. Eine Abfolge unterschiedlicher Segmenttypen wird als komplexer angenommen. Die Autoren geben keine Komplexitätswerte für die vier Segmenttypen an.

Pressings Maß der Komplexität von rhythmischen Mustern evaluiert die Synkopation auf unterschiedlichen Pulsebenen [Pre04], welches Tonhöhen- oder Klangfarbenunterschiede[7] unberücksichtigt lässt. Die Bewertung geschieht anhand von definierten Gewichten für unterschiedliche Arten der Synkopierung in einer Abfolge von vier Noten

[7] Es wird hier vereinfacht angenommen, das Klangfarben und Tonhöhen eine ähnliche Bedeutung für die Wahrnehmung von Rhythmus besitzen.

o . . .	Leeres Segment
o . o .	Gleichmäßig unterteiltes Segment
o . . o	Ungleichmäßig unterteiltes Segment
. o . .	Mit Stille beginnendes Segment

Tabelle 3.1: Die vier Segmenttypen zur Berechnung der Komplexität von quantisierten Rhythmen nach Shmulevich und Povel [SP98a]: Noten sind durch o, Pausen durch . dargestellt.

oder Pausen, wobei nach Pressing sechs Synkopationstypen auftreten können. Die Synkopation wird auf unterschiedlichen metrischen Ebenen ausgewertet und zu einem Gesamtergebnis addiert.

Toussaint analysierte eine Reihe von afro-kubanischen Clavemustern und ermittelt ein Komplexitätsmaß basierend auf den metrischen Positionen, auf denen Noten auftreten [Tou02]. Noten, die mit Pulsen auf höheren metrischen Ebenen zusammentreffen, tragen weniger zur Komplexität bei. Die von Toussaint definierte Komplexität korreliert daher mit dem Grad der Synkopation des Rhythmus.

Die Komplexitätsmaße von Povel und Shmulevich, Pressing, Tanguiane und Toussaint sind auf die Untersuchung von Abfolgen von identischen Noten beschränkt und deshalb für die Betrachtung realer Stimuli weniger geeignet. Die ersten beiden Methoden sind nur für binäre Rhythmen geeignet. Die in [SP98b, Tou02] dargestellten Ergebnisse zeigen, dass das Komplexitätsmaß nach Lempel und Ziv für die Analyse sehr kurzer Segmente nicht geeignet ist.

3.6 Identifikation perkussiver Instrumente ohne Tonhöheninformation

Rhythmus wird häufig durch das Auftreten von perkussiven Instrumenten ohne Tonhöheninformation[8] manifestiert. Die Identifikation dieser Instrumente liefert wertvolle Informationen für eine rhythmische Analyse.

Die Ansätze zur Identifikation von perkussiven Instrumenten werden unterschieden in

[8] Als „perkussive Instrumente ohne Tonhöheninformation“ werden die zu den Membranophonen (Fellklinger) und Idiophonen (Selbstklinger) gehörenden Perkussionsinstrumente bezeichnet, die keinen oder nur einen schwachen Eindruck von Tonhöhe vermitteln, und deren Transkription keine explizite Tonhöhenangabe, sondern nur die Art des Instruments und gegebenenfalls die Spielweise erfordert.

Verfahren mit und ohne Quellenseparation (engl. *source separation*). Die Dissertation von FitzGerald [Fit04] enthält eine aktuelle und umfassende Darstellung des Standes der Technik. Die in diesem Abschnitt enthaltene Vorstellung der wichtigsten Verfahren ist unterteilt in Verfahren mit und ohne Quellenseparation.

Quellenseparation (engl. *Blind Source Separation*) dient der Entmischung der einzelnen auditiven Ströme (engl. *auditory streams*). Das menschliche Gehör ist zu sehr leistungsstarken Quellenseparationen in der Lage und erlangt zu einem wesentlichen Anteil durch die Fähigkeit der Quellenseparation einen Vorteil gegenüber maschinellen Systemen bei einer Reihe von Anwendungen. Das Phänomen der separaten Wahrnehmung einzelner Quellen wird als *Cocktailparty-Effekt* bezeichnet, da die Quellenseparation es dem Zuhörer ermöglicht, sich in Umgebungen mit vielen akustischen Quellen auf einzelne Szenen zu konzentrieren, beispielsweise auf einer Cocktailparty.

Zu den Anwendungen, die von Fortschritten auf dem Gebiet der Quellenseparation profitieren werden, gehören neben der Musikanalyse zum Beispiel die Klanglokalisation und die Spracherkennung in gestörten Umgebungen.

Instrumentenerkennung ohne Quellenseparation

Ein frühes Verfahren von Schloss extrahiert Instrumenteninformationen aus perkussiver Musik [Sch85]. Dabei wird neben den Instrumenten die Schlagtechnik bei Handtrommeln klassifiziert. Die Klassifikation geschieht hierarchisch beginnend mit einer Unterscheidung zwischen gedämpften und ungedämpften Schlägen auf Grundlage von Hüllkurveninformationen. Spektrale Merkmale werden anschließend zur Klassifikation von Schlagtechnik und Instrument analysiert, wobei vier Klassen für Schlagtechniken und zwei Instrumentenklassen unterschieden werden.

In [GH01] werden verschiedene Signalverarbeitungs- und Klassifikationstechniken zur Instrumentenerkennung in perkussiver Musik verglichen. Ausgehend von einer auf dem Tatumraster basierenden Segmentierung werden verschiedene Low-level-Merkmale extrahiert. Zur Klassifikation werden Clustering-Techniken, Lineare Diskriminanzanalyse (LDA) und Entscheidungsbäume untersucht. Eine Übersicht über Merkmalsselektions- und Klassifikationstechniken zur Identifikation einzelner Instrumentensamples enthalten [HYG02, HDG03].

Die Eignung verschiedener signalnaher Merkmale zur Instrumentenerkennung in polyphoner Musik wurde in [GPD00] untersucht. Die auftretenden Noten werden ausgehend von einer Hüllkurvenextraktion in Anschlags- und Ausklingphase segmentiert. Die zur Klassifikation verwendeten Merkmale werden aus Zeit- (zum Beispiel Anschlags- und Ausklingzeit und Nulldurchgangsrate) und Frequenzbereichsdarstellung (zum Beispiel die Anzahl der Sinusoide in einem Modell nach Prony) extrahiert.

Das in [McD98a] publizierte Verfahren verwendet eine waveletbasierte Cochlea-Filterbank und ein Modell der inneren Haarzellen nach Meddis. Zur Identifikation der

Quellen werden die Ausgangssignale der Filterbank als Merkmale und ein geometrischer Klassifikator verwendet. Ein weiterer merkmalsbasierter Ansatz ist in [Sil00] beschrieben. In [ZPDG02] ist ein Analyse-Synthese-Verfahren zur Instrumentenerkennung vorgestellt.

Goto entwickelte ein Verfahren zur Detektion von Bass- und Snaredrum zum Einsatz in BTS. Das Auftreten der Bassdrum wird detektiert durch die Auswertung der Frequenzgewichte, die zu detektierten Noteneinsätzen gemessen werden. Die charakteristischen Frequenzen des Perkussionsinstrumentes sind nicht a-priori festgelegt. Die tiefste zu einem lokalen Maximum einer Histogrammdarstellung der zu Noteneinsätzen gemessenen Frequenzgewichte korrespondierende Frequenz wird als Detektionkriterium verwendet. Zur Detektion von Snaredrums wird ein spektrales Maß, welches zur Flachheit des Spektrums korreliert, ermittelt und setzt somit die Annahme der Rauschhaftigkeit des Klanges voraus.

Yoshii *et. al.* entwickelten ein auf Template-Adaption und -Abgleich basierendes Verfahren und erreichten bei der Detektion von Bass- und Snaredrum eine Erkennungsrate von 90% [YGO04].

Das in [SKSV00] präsentierte Verfahren verwendet zur Vorverarbeitung ein Sinusoidal-und-Noise-Modell, wobei die Annahme getroffen wird, dass nur geringe Signalanteile der Schlagzeuginstrumente im Sinusoidalsignal auftreten. Zur Detektion von Trommelklängen wird der residuale Signalanteil mit einem Mustererkennungsverfahren analysiert, wobei als Merkmale die Energieanteile in einzelnen Barkbändern verwendet werden. Die untersuchten Signalausschnitte sind so ausgewählt, dass ihr Beginn mit metrischen Pulsen zusammentrifft. Eine weitere Besonderheit des Verfahrens ist die Ergänzung des datenbasierten Ansatzes durch die Einbeziehung von rudimentären Vorwissen und einfachen Regeln, indem A-priori-Wahrscheinlichkeiten für die Instrumentenklassen festgelegt und in Abhängigkeit von vorangegangenen Detektionen modifiziert werden. Die sinusoidale Komponente ist jedoch für Klänge perkussiver Instrumenten ohne Tonhöheninformation von großer Bedeutung für die Klangcharakteristik [Ros00].

Paulus und Klapuri veröffentlichten ein Verfahren zu Klassifikation von drei abstrakten Instrumentenklassen [PK03b]. In vielen populären Musikgenres wird die perkussive Instrumentierung von den Instrumenten Bassdrum, Snaredrum und Hi-hat oder ähnlich klingenden Instrumenten dominiert. Die Autoren definieren drei Instrumentenklassen in Abhängigkeit von ihrer Funktion in einem Musikstück. Das Verfahren ist zur Anwendung in einem *Query-by-rhythm*-System konzipiert, welches eine Nutzeranfrage, die mit beliebigen Klängen produziert wird, verarbeitet. Die Ereignisse im Audiosignal werden auf Basis von Low-level-Merkmalen geclustert und ihre metrischen Positionen in einem 4/4-Takt ermittelt. Zur Klassifikation werden A-priori-Wahrscheinlichkeiten des Auftretens der Klassen zu bestimmten metrischen Positionen verwendet, die aus einem Trainingsdatensatz ermittelt wurden.

Ein weiteres Verfahren der Autoren klassifiziert perkussive Instrumente in sieben Klassen zuzüglich einer Rückweisungsklasse unter Verwendung von *N-grams* zur Auswer-

tung der zeitlichen Abfolge der Observationen [PK03a]. *N-grams* werden in der Spracherkennung zur Ermittlung einer Wahrscheinlichkeit des Auftretens eines Ereignisses in Abhängigkeit vorangegangener Ereignisse eingesetzt.

Instrumentenerkennung mit Quellenseparation

Da in einem musikalischen Kontext häufig Ereignisse aus mehreren Instrumentenklängen zusammengesetzt sind, stellt Quellenseparation eine geeignete Vorverarbeitungsstufe für die Merkmalsextraktion dar.

Verschiedene Verfahren mit Quellenseparation basieren auf einer Zerlegung des Audiosignals durch die von Hyvärinen und Hoyer eingeführte Independent Subspace Analysis (ISA) [HH99]. In [CW00] wurden die Grundlagen des Verfahrens zur Separation von mehreren Quellen aus einkanaligen Mischungen von Audiosignalen gelegt. Die Erhöhung der Dimensionalität der zu separierenden Beobachtung ist eine Voraussetzung für die Anwendung verschiedene Ansätze zur Quellenseparation. Hier sei im Speziellen die Independent Component Analysis [Com94] genannt, die bedingt, dass die Anzahl der Beobachtungen gleich der oder größer als die größeren Anzahl der zu separierenden Quellen ist. Diese Vorausetzung erfüllt eine Spektrogrammdarstellung $S(f, t)$ des Audiosignals. Mittels Singular Value Decomposition und Auswahl der Komponenten wird die Dimensionalität dieser Beobachtungsmatrix wiederum aus Gründen der Recheneffizienz und zur Beschränkung der Anzahl der separierten Quellen verkleinert. Eine Herausforderung liegt hier in der Auswahl der Anzahl der Komponenten.

Die in [Ori01] publizierte Arbeit verwendet Caseys Verfahren zur rhythmischen Analyse mit Quellenseparation, gibt aber keine Informationen über Details der Evaluierung.

Caseys Verfahren beinhaltet jedoch keine Klassifikation der Quellen. Das in [UDS03] vorgestellte Verfahren erweitert Caseys Ansatz um eine nachträgliche Klassifikation der separierten Quellen in perkussive Klänge ohne Tonhöheninformation und tonale Klänge [UDS03]. Dazu wurden Merkmale aus den spektralen Profilen und den Amplitudenhüllkurven der Quellen extrahiert, die die Tonalität auswerten und transiente Zeitverläufe detektieren.

Eine Variante der ISA stellt die von FitzGerald vorgeschlagene getrennte Analyse eines unteren und oberen Teilbandes dar, motiviert durch die Annahme, dass Klänge von Membranophonen im unteren Frequenzband ($< 1000Hz$) und von Idiophonen im oberen Teilband ($> 2000Hz$) den größten Energieanteil aufweisen. FitzGerald veröffentlichte Ergebnisse der Detektion von drei Instrumentenklassen (Bassdrum, Snaredrum und Hi-hat)

Ein alternativer Ansatz zur ISA ist die von Fitzgerald entwickelte *Prior Subspace Analysis* dar. Hier werden die Frequenzgewichte der zu detektierenden Instrumente vorgegeben [FLC03a, FLC03b].

3.7 Alternative Merkmale zur Beschreibung rhythmischer Eigenschaften

Zur Ermittlung des Swing in Audiosignalen mit konstantem Tempo und $\frac{4}{4}$-Takt schlägt Laroche die Auswertung eines Histogramms vor, in denen die zeitlichen Abstände von Transienten im Hüllkurvensignal zu den vorherigen Zählzeiten eingetragen werden [Lar01].

Herrera *et. al.* schlugen verschiedene neue rhythmische Merkmale zur Anwendung in MIR-Systemen vor [HSG04]. Der *Perkussions-Index* wird berechnet als das Verhältnis der Anzahl von Noten von perkussiven Instrumenten zur Anzahl aller Noten. Verschiedene Low-level-Merkmale dienen zur Klassifikation der Noten in perkussiv und nicht-perkussiv. Das *Perkussions-Profile* ist ein Merkmal, welches einen Überblick über die im Signal enthaltenen Perkussionsinstrumente gibt. Das verwendete Verfahren zur Instrumentenerkennung ist nur ansatzweise skizziert. Ein hierarchischer Klassifikator ohne Quellenseparation klassifiziert zu Noteneinsätzen extrahierte Segmente des Audiosignals in perkussiv vs. harmonisch, Membranophone vs. Idiophone, einzelne Instrumente vs. Kombinationen aus Instrumenten und in die Instrumentenklassen. Die Anzahl der detektierten Instrumentenklassen und die Menge der Kombinationen aus unterschiedlichen Instrumenten ist mit drei Instrumenten und zwei Kombinationen gering.

Als weiteres abstraktes Merkmal (*Kick-Snare Crossings*) wird die mittlere Anzahl der auf eine Note der Bassdrum folgende Note der Snaredrum und der auf eine Note der Snaredrum folgende Note der Bassdrum untersucht, wobei jedoch laut den Autoren kein für MIR-Anwendungen einsetzbares Merkmal ermittelt wurde.

Das Merkmal *Perkussivität* ist ein Maß für die Steilheit des Anschlags einer Note und basiert auf dem in [UDS03] publiziertem Perkussivitätsmerkmal zur Klassifikation separierter Quellsignale.

4 Extraktion der rhythmischen Eigenschaften eines Musiksignals

4.1 Überblick

Zur Beschreibung der rhythmischen Eigenschaften eines Musiksignals wurden die folgenden Merkmale ausgewählt:

- das musikalisches Tempo, die Taktart und die Mikrotime
- die gegebenenfalls auftretenden Drumpattern
- die rhythmische Eingängigkeit und Intensität

Tempo, Taktart und Mikrotime beschreiben die metrische Struktur in kompakter Weise. Sie zählen zu den grundlegenden Merkmalen von musikalischem Rhythmus.

Vorangegangene Arbeiten zeigten, dass aus Drumpattern abgeleitete Merkmale zur Klassifikation des musikalischen Genres erfolgreich angewendet werden können [UD04a]. Das Auftreten von Drumpattern ist ebenso wie das der metrischen Merkmale nicht immer gegeben und genreabhängig. Es wird jedoch in populärer Musik und deshalb gemessen am Umsatz sehr häufig beobachtet. Eine spezielle Anwendung der automatisierten Extraktion der Drumpattern ist *Query-by-beatboxing* (QBB) [KBT04]. QBB bezeichnet ein Verfahren zur Suche von Musik anhand einer Vorgabe, die mit Vokalklängen das Drumpattern imitiert.

Intensität beschreibt Eigenschaften, die von oben genannten, in einer Transkription beschriebenen Merkmalen nicht erfasst werden und die Aufführung beschreiben. Eingängigkeit wird hier als komplementäres Merkmal zur Komplexität ermittelt. Beide Merkmale werden mit dem Ziel extrahiert, alternative und intuitive Suchanfragen ohne ausgeprägte musikalische Kenntnisse zu ermöglichen.

Rhythmische Merkmale aus Audiosignalen zu extrahieren ist ein perzeptueller Prozess, bei dem, wie bei allen perzeptuellen Prozessen, Information auf verschiedenen Abstraktionsebenen ausgewertet wird. Da auf jeder Ebene Fehler oder Mehrdeutigkeiten auftreten können, sollten diese Prozesse nicht in einer festgelegten Reihenfolge, zum Beispiel beginnend auf der Ebene der größten Abstraktion, ablaufen. Vorzugsweise sollte die die zuverlässigste Information liefernde Ebene ausgewertet werden.

Die Auswertung verschiedener Informationen zum Erreichen einer Erkenntnis motiviert Bregman [Bre98] wie folgt:

„It is fortunate that many cues and heuristics of ASA (Auditory Scene Analysis, der Autor) have been identified, because none of them can be trusted all the way, due to the fact that any particular cue may be blurred or absent in a particular environment.“

Die rhythmische Analyse erfolgt auf zwei Ebenen, die hier *Low-level-* und *High-level-*Analyse benannt sind. Die überwiegende Mehrheit von Verfahren zur Extraktion rhythmischer Merkmale basieren auf einer Analyse von signalnahen Merkmalen. Diese Verfahren arbeiten robust für eine Vielzahl von Musikstücken bei der Schätzung des Tempos, der Taktart oder der Klassifikation in Rhythmus-Kategorien [DPG03, GD04]. Diese Analyse ist anfällig gegen verschiedene Fehler, zu denen Tempo-Oktav-Fehler (das heißt die Ermittlung des doppelten oder halben Tempos), die Ermittlung des synkopierten Tempos und Phasenfehler beim Tracking (zum Beispiel die Ermittlung des Zählzeitenraster zwischen den Zählzeiten bedingt durch eine Betonung des Off-Beats) zählen. Musikalisches Vorwissen kann nur begrenzt in die Entscheidungsfindung einbezogen werden, da dieses Wissen nicht die Interpretation der Ausprägung von Low-level-Merkmalen erlaubt.

Zur Nutzung von musikalischem Vorwissen bei der rhythmischen Analyse und zur Extraktion weiterer semantisch hochwertiger Merkmale werden in dieser Arbeit Instrumenteninformationen ausgewertet. Die Instrumenteninformationen können in begrenzten Maß direkt aus dem Signal mit modernen Mustererkennungsmethoden extrahiert werden, wie in [DU04] für perkussive Instrumente ohne Tonhöheninformation gezeigt wurde. Diese Methoden setzen jedoch das Auftreten perkussiver Instrumente voraus und Versagen bei Nichtauftreten der antrainierten Instrumente, wenn die getroffenen Annahmen nicht zutreffen. Die Auswertung der Ergebnisse muss deshalb Konfidenzmaße der einzelnen Verfahren berücksichtigen.

4.2 Ermittlung der metrischen Struktur aus Low-level-Signal-Deskriptoren

Im Rahmen der Low-level-Analyse wird das Signal zunächst so vorverarbeitet, dass ein einkanaliges Audiosignal mit einer ausreichenden Aussteuerung vorliegt. Da die Charakteristiken eines Musiksignal sich über der Zeit ändern, werden die Signale in charakteristische Abschnitte segmentiert (siehe Abschnitt 4.2.1).

In Abschnitt 4.2.2 wird die Berechnung des Akzentsignals aus Hüllkurveninformationen beschrieben. Die Detektion von Noteneinsatzzeitpunkten (NEZ) ist in Abschnitt 4.2.3 dargestellt. Das Akzentsignal und die NEZ dienen als Grundlage für die in Abschnitt 4.2.4 beschriebene Extraktion der Merkmale Tempo, Taktart und Mikrotime.

Weiterhin wird das Tatumraster ermittelt, welches in der High-level-Analyse zur Quan-

tisierung der detektierten Noten der perkussiven Instrumente Verwendung findet. Das Blockschaltbild der Low-level-Analyse zeigt Abbildung 4.1.

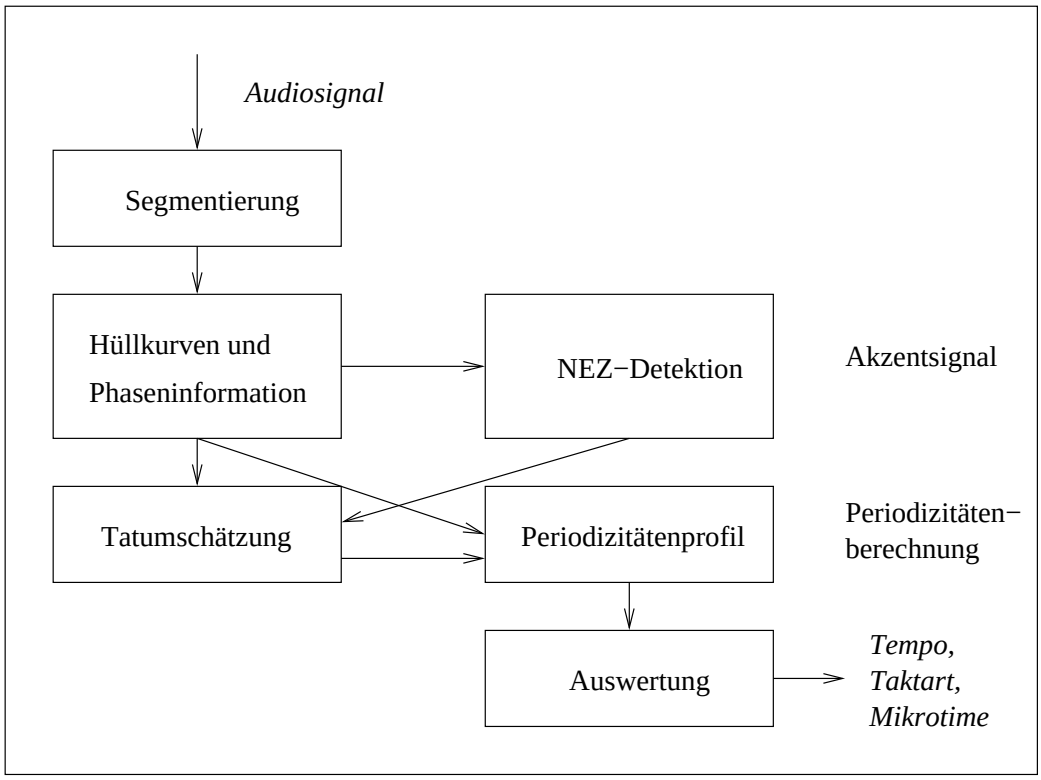


Abbildung 4.1: Blockdiagramm der Low-level Analyse.

4.2.1 Segmentierung in charakteristische Abschnitte

Die Segmentierung des Musiksignals in charakteristische Abschnitte ist durch die Beobachtung motiviert, dass die zu extrahierenden Merkmale innerhalb eines Musiksignals unterschiedlich sein können, jedoch innerhalb charakteristischer Segmente identisch sind. Solche Segmente sind zum Beispiel Einleitung, Strophe und Refrain. Die Methode zur Detektion von Segmentgrenzen wird im folgenden Abschnitt beschrieben.

Das Verfahren basiert auf der Selbstähnlichkeitsanalyse nach Foote [FC03]. Zur Bestimmung der Selbstähnlichkeit werden die Ausprägungen verschiedener Low-level-Merkmale untersucht. Eine Bibliotheksfunktion zur Segmentierung wurde im Rahmen einer Diplomarbeit am Fraunhofer IDMT erstellt [Pin04], welche auf Erkenntnissen einer vorangegangenen studentischen Arbeit [Vei03] aufbaut.

Folgende Verarbeitungsschritte werden zur Bestimmung der Segmentgrenzen durchgeführt:

- Extraktion von signalnahen Merkmalen,
- Erstellung einer Selbstähnlichkeitsmatrix,
- Berechnung eines Neuheitsmaßes und Bestimmung der Segmentgrenzen

Extraktion der Low-level-Merkmale Aus den Frequenzspektren nichtüberlappender Blöcke des Audiosignals mit einer Länge von 30 ms werden die Merkmale *Audio Spectrum Envelope* (ASE), *Spectral Flatness Measure* (SFM) und *Mel-Frequency Cepstral Coefficients* (MFCC) berechnet.

Die Merkmale ASE und SFM sind im MPEG7-Standard [MPE01] spezifiziert, wobei SFM dort als *Audio Spectrum Flatness* bezeichnet wird. Die Merkmale werden in 16 Frequenzbändern gemessen, die logarithmisch zwischen 250 Hz und 4 kHz verteilt sind. Die ASE beschreibt den Energiegehalt in jedem Frequenzband, der aus dem Leistungsdichtespektrum (LDS) ermittelt wird. Zusätzlich zu den 16 Frequenzbändern werden zwei weitere Bänder hinzugefügt, die den Energiegehalt unter 250 Hz und über 4 kHz beschreiben. Das SFM beschreibt die Flachheit eines LDS beziehungsweise eines Ausschnittes eines LDS. Es ist definiert als Quotient aus geometrischen und arithmetischen Mittelwert der LDS-Koeffizienten. Das Merkmal dient der Beschreibung der Rauschhaftigkeit eines Klanges.

MFCC wurden in der Vergangenheit in der Sprach- und Sprechererkenung eingesetz, finden aber auch in der automatisierten Musikanalyse breite Anwendung [Log00]. Die Koeffizienten werden aus einem Amplitudenspektrum berechnet, wobei die logarithmierten Koeffizienten mit einer Mel-Filterbank gewichtet und anschließend mit einer Diskreten Kosinustransformation (DCT) transformiert werden. Zur Segmentierung werden die ersten 12 Koeffizienten verwendet.

Zur Extraktion der Low-level-Merkmale wird das am Fraunhofer IIS und Fraunhofer IDMT entwickelte Softwareframework *XProEx* eingesetzt. Eine Anzahl von benachbarten Merkmalsvektoren wird gruppiert und durch den arithmetischen Mittelwert repräsentiert. Die Gruppierung ermöglicht eine effizientere Berechnung der Selbstähnlichkeitsmatrix und erfüllt eine Glättungsfunktion.

Erstellung der Selbstähnlichkeitsmatrix Die Selbstähnlichkeitsmatrix S wird durch paarweisen Vergleich aller durch Low-level-Merkmale repräsentierten Abschnitte des Audiosignals v_i berechnet.

$$S(i,j) = d(v_i, v_j) \qquad i,j = 1,...,N \tag{4.1}$$

mit einem Distanzmaß $d(v_i, v_j)$. Als Distanzmaß wird die Kosinusdistanz nach Gleichung 4.2 verwendet.

$$d_{cos}(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i| \cdot |v_j|} \tag{4.2}$$

Die Wahl der Kosinusdistanz ist motiviert durch die Unabhängigkeit der ermittelten Distanz vom Pegel der Merkmale. Bei n Merkmalsvektoren sind n^2 Distanzberechnungen notwendig, die durch Gruppierung von g Merkmalsvektoren auf n^2/g^2 reduziert werden.

Berechnung des Neuheitsmaßes und der Segmentgrenzen Aus der Selbstähnlichkeitsmatrix $S(i, j)$ wird ein Neuheitsmaß berechnet, welches lokale Maxima an den Segmentgrenzen aufweist.

Zur Berechnung des Neuheitsmaßes wird $S(i, j)$ mit einer schachbrettartigen Kernelmatrix K entlang der Hauptdiagonalen korreliert. Die Matrix K ist in ihrer einfachsten Form eine 2×2 Matrix:

$$K = \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \tag{4.3}$$

Durch Bildung des Kronecker-Produkts aus K und einer $n \times n-$Matrix M, $m_{i,j} = 1, \quad i, j = 1, .., n,$ wird eine vergrößerte Kernelmatrix K' der Größe $2n \times 2n$ gebildet. Dies ist in Gleichung 4.4 für den Fall $n = 2$ dargestellt.

$$\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} +1 & +1 & -1 & -1 \\ +1 & +1 & -1 & -1 \\ -1 & -1 & +1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix} \tag{4.4}$$

Die resultierende Kernelmatrix K' wird zur Abflachung ihrer Kanten gaußförmig gewichtet. Die Größe von K' hat Einfluss auf die Glattheit des Neuheitsmaßes und muss auf die Größe der zu detektierenden Segmente abgestimmt sein. Zusätzlich wird das Neuheitsmaß mit einem FIR-Tiefpassfilter vierzigster Ordnung geglättet. Segmentgrenzen werden beim Auftreten lokaler Maxima des Neuheitsmaßes detektiert, wenn eine Anzahl vorangehender Werte monoton steigend und eine Anzahl folgender Werte monoton fallend sind.

4.2.2 Ermittlung eines Akzentsignals

Da die Lautheit die Quantität ist, die am besten zur Intensität einer auditiven Empfindung korrespondiert, ist die Berechnung eines Akzentsignals basierend auf der spezifischen Lautheit nahe liegend [ZF99]. Zur vereinfachten Berechnung der spezifischen Lautheit wird das Audiosignal in Teilbandsignale zerlegt, die entsprechend den Kurven gleicher Lautstärke gewichtet werden.

Durch die starke Kompression des Dynamikbereiches bei der Lautheitsberechnung sind Akzente und NEZ jedoch schwieriger zu detektieren als im Verlauf der Amplitude eines Teilbandsignales.

Zur Berechnung des Akzentsignals wird das Audiosignal in 24 Frequenzbänder zwischen 0 Hz und 15500 Hz zerlegt. Die Breite der Frequenzbänder ist der Bark-Skala entlehnt. Die Verwendung der Mel- und Bark-Skala führt in der Spracherkennung zu ähnlich guten Ergebnissen [SP03].

Die Frequenzbandsignale werden gewonnen durch DFT und Zusammenfassen der Bins, die in einem Band liegen, wobei die Bins an den Bandgrenzen anteilig gewichtet werden. Für die Berechnung der DFT werden Blöcke mit einer Länge von 23 ms (1024 Samples bei 44100 Hz Abtastrate) mit 87.5% Überlappung verarbeitet, die mit einem Hann-Fenster gewichtet werden. Die relativ große Überlappung zwischen den Blöcken gewährleistet eine ausreichend hohe Auflösung der im weiteren Verlauf berechneten Hüllkurvensignale, die mit einer Abtastrate von 344 Hz vorliegen.

Die Hüllkurvensignale $E_{i,j}$, mit Teilbandindex i und Index des Abtastwertes j werden berechnet durch Glättung mit einem FIR-Tiefpass, dessen Koeffizienten hier ein 100 ms langes Blackman-Fenster (Gleichung 4.5) bilden.

$$h(n) = 0.42 - 0.5 \cdot \cos(2\pi n/M) + 0.08 \cdot \cos(4\pi n/M) \tag{4.5}$$

Von den Hüllkurvensignalen werden modifizierte relative Differenzenfunktionen dritter Ordnung $D'_{i,j}$ berechnet (Gleichung 4.6) und halbweggleichgerichtet (Gleichung 4.7).

$$D'_{i,j} = (\max(E_{i,k}) - E_{i,j})/\text{mean}(E_{i,k}) \quad , \quad k = j...j+3 \tag{4.6}$$

$$D_{i,j}(n) = \begin{cases} D'_{i,j}(n) & \text{falls} \quad D'_{i,j}(n) > 0 \\ 0 & \text{sonst} \end{cases} \tag{4.7}$$

Weiterhin wird der Dynamikbereich der resultierenden Akzentsignale komprimiert (Gleichung 4.8).

$$D_{i,j}(n) = \log(1 + \mu \cdot D_{i,j}(n))/\log(1 + \mu) \tag{4.8}$$

Die Akzentsignale der Teilbänder werden für die in Abschnitt 4.2.4 beschriebene Schätzung der metrischen Merkmale zu einem Akzentsignal aufsummiert. Die Analyse der Teilbandakzentsignale stellt eine alternative Möglichkeit der rhythmischen Analyse dar.

Entsprechend der Annahme, dass auditive Ströme in unterschiedlichem Maß die Wahrnehmung von Rhythmus beeinflussen, werden die Teilbänder entsprechend ihrer Periodizität nach Gleichung 4.9 gewichtet.

$$\widehat{D}_{i,j} = (r_{xx,1}/r_{xx,0})^c \cdot D_{i,j} \tag{4.9}$$

Die Gewichte w_i werden aus der AKF r_{xx} der Akzentsignale $D_{i,j}$ berechnet, wobei $r_{xx,0}$ das globale Maximum der AKF und $r_{xx,1}$ das globale Maximum eines bei einer Verzögerung von $\tau = 150$ ms beginnenden Ausschnittes der AKF bezeichnet und die Konstante c den Grad der Wichtung beeinflusst. Teilbandsignale mit einem unperiodischen Verlauf werden so stärker gedämpft und verlieren in weiteren Analyseschritten an Einfluss. Die Berechnung der Wichtungsfaktoren wurde heuristisch ermittelt.

4.2.3 Detektion von Noteneinsätzen

Die Detektion von Noteneinsätzen dient zur Ermittlung eines alternativen Akzentsignals für die metrische Analyse und zur Ermittlung des Tatumrasters. Diese Anwendung stellt gemäßigtere Anforderungen an die Leistungsfähigkeit des Verfahrens im Gegensatz zur automatisierten Transkription. Während für eine automatisierte Transkription idealerweise eine fehlerfreie Erkennung erwünscht ist, ist die Bestimmung der metrischen Struktur robust gegenüber Einzelfehlern bei der Notendetektion.

Die hier verwendete Methode stellt eine Kombination aus amplituden- und phasenbasiertem Ansatz dar. Wie Abbildung 4.2 exemplarisch zeigt, kann zu Noteneinsätzen ein Anstieg der Amplitude und der Phasenkongruenz beobachtet werden.

Die Amplituden E_i und relativen Differenzenfunktionen der Amplituden D_i der $i-ten$ Teilbänder werden mit einem einfachen Regelwerk ausgewertet. Noteneinsätze werden in Teilbandsignalen zum Zeitpunkt j detektiert, wenn

- $D_{i,j} > D_{i,j-1} \quad \& \quad D_{i,j} > D_{i,j+1}$
- $D_{i,j} > d_s, \quad$ mit festem Schwellwert d_s
- $D_{i,j} > d_a, \quad$ mit adaptiven Schwellwert d_a
- $E_{i,j} > e_s, \quad$ mit festem Schwellwert e_s

Der adaptive Schwellwert d_a wird bei Auftreten einer Note zum Zeitpunkt j auf den Wert $d_a = D_{i,j} - \epsilon$ gesetzt und verringert sich nach 66 ms um die Hälfte. Weitere Bedingungen zur Detektion einer Note sind der Abfall von E_i und D_i zwischen zwei detektierten Noten unter einen adaptiven Schwellwert und ein Mindestabstand zwischen zwei

detektierten Noten. Die Schwellwerte und Zeitkonstanten wurden durch sorgfältige Beobachtung eines Trainingsdatensatzes ermittelt.

Die in Teilbändern detektierten Noten werden über alle Bänder kombiniert und die Intensitäten aus den Teilbändern aufakkumuliert. Zur Detektion von Noteneinsätzen mit geringer Energie und Verbesserung der Genauigkeit werden die Phaseninformationen ausgewertet.

Die Phaseninformation wird durch Berechnung der Phasensumme über die DFT-Bins nach Gleichung 4.10 mit Frameindex i und Binindex k des Phasenwinkels der komplexen Fourierkoeffizienten $\phi(k, i)$ ausgewertet.

$$\widehat{\phi}(i) = \sum_{k=0}^{K} \phi(k, i) \tag{4.10}$$

Die Phasenwerte werden „entrollt“, das heißt, dass ganzzahlige Vielfache von π addiert werden, um einen kontinuierlichen Verlauf des Phasenwinkels zu erhalten.

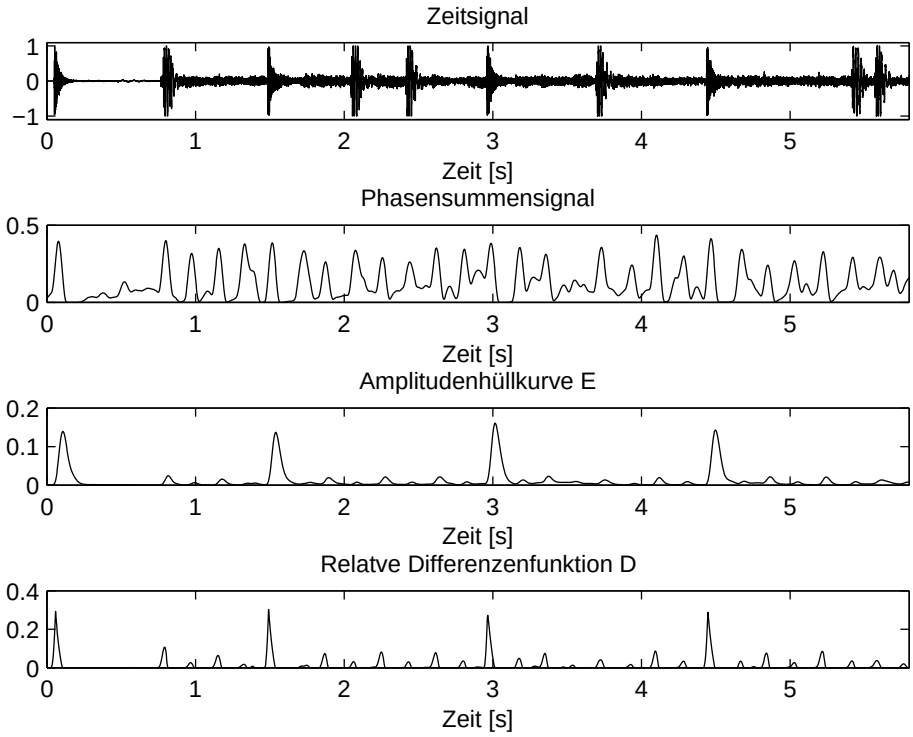


Abbildung 4.2: Detektionsfunktionen für Noteneinsätze: Zeitsignal, Phasensummensignal des Audiosignals und Amplitudenhüllkurve E und Relative Differenzenfunktion D eines Teilbandsignals.

4.2.4 Extraktion der metrischen Merkmale

Die Ermittlung der metrischen Struktur eines Musiksignals beinhaltet die Schätzung der Perioden und Phasenlagen von Pulsserien auf den hierarchischen Ebenen des Tatums, der Zählzeit und des Taktes.

Tatumschätzung Die Schätzung des Tatums erfolgt durch eine Auswertung der Schätzungen vier verschiedener Methoden (siehe Abschnitt 3.3) in einem Abstimmungsverfahren. Die erste Methode (TWME-NEZ) wertet die NEZ mit dem TWME-Verfahren nach Gouyon [GHC02] aus. Als weiteres Verfahren (GGT-NEZ) findet die von Sepännen vorgeschlagene Suche des GGT Anwendung.

Zur Behandlung von Teststücken mit geringen Auftreten von Noten beziehungsweise unzuverlässig extrahierten NEZ werten zwei weitere Verfahren die AKF des über die Bänder aufsummierten Akzentsignals aus. Dabei wird eine modifizierte Version der

TWME-Analyse (TWME-AKF) und des Verfahrens nach Paulus (DFT-AKF) [PK02] eingesetzt.

Jede der verwendeten Methoden analysiert das Akzentsignal blockweise mit einer Blockgröße in der Größenordnung von 2.5 s, um auf Temposchwankungen reagieren zu können.

Die Ergebnisse aller Berechnungsmethoden werden mit dem *Sequential Leader* Algorithmus [Har75] geclustert. Als Ergebnis der Tatumschätzung wird der Mittelwert des Clusters gewählt, welches die größte Zuverlässigkeit besitzt. Diese Zuverlässigkeit wird ermittelt aus der Anzahl der zu einem Cluster gehörenden Punkte und einem für jedes Verfahren ermittelten Konfidenzwert k_t, der nach Gleichung 4.11 aus der Anzahl der korrekt geschätzen Werte N_c und der Gesamtanzahl aller Schätzungen N bei der Klassifikation eines Trainingsdatensatzes (siehe Abschnitt 5.2) ermittelt wurde. Die Konstante d dient der Wichtung des Einflusses der Konfidenzwerte der Verfahren.

$$k_t = (\frac{N_c}{N})^d \tag{4.11}$$

Periodizitätenberechnung und -profil Eine geeignete Berechnungsvorschrift zur Ermittlung der in einem Signal auftretenden Periodizitäten ist die AKF, bei der die Ähnlichkeit zwischen zeitlich verschobenen Vektorelementen durch Produktbildung bestimmt wird. Bei der Periodizitätenberechnung eines mehrkanaligen Akzentsignal $A(n, b)$ (wobei ein Kanalsignal z.B. das Akzentsignal eines Teilbandes ist) mit Stützstellenindex $n = 1...N$ und Kanalindex $b = 1...B$ wird ein Akzent zum Zeitpunkt n durch einen Vektor repräsentiert und ein geeignetes Ähnlichkeitsmaß kann durch Berechnung des Skalarproduktes ermittelt werden (siehe Gleichung 4.12) beziehungsweise durch die rechnerisch äquivalente Summation der AKF der einzelnen Teilbänder.

$$\widehat{r}_{xx}(\tau) = \frac{1}{N} \sum_{n=1}^{N-\tau} \sum_{b=1}^{B} A(n, b) \cdot A(n + \tau, b) \tag{4.12}$$

Um Überbewertungen kleiner Verzögerungen zu vermeiden, wird die in Gleichung 4.13 für den einkanaligen Fall dargestellte entzerrte AKF verwendet.

$$r_{xx}(\tau) = \frac{1}{N - \tau} \sum_{n=1}^{N-\tau} a(n) \cdot a(n + \tau) \tag{4.13}$$

Aus dem Akzentsignal wird eine kompakte Darstellung der Intensitäten von Periodizitäten abgeleitet, für die hier der Begriff Periodizitätenprofil eingeführt wird.

Ein Periodizitätenprofil V stellt die Intensitäten von Periodizitäten dar, die zu Verzögerungen auftreten, die ganzzahligen Vielfachen der Tatumperiode entsprechen.

Diese Darstellung ähnelt einer unterabgetasteten Periodizitätenfunktion, mit einer der Tatumperiode entsprechenden Unterabtastung.

Die Berechnung geschieht jedoch nicht durch eine präzise Unterabtastung, sondern durch Zuordnung der lokalen Maxima zu den nächstgelegenen Vielfachen, da Rundungsfehler und ein eventuelles nicht exakt ganzzahliges Verhältnis von Tatumperiode und höheren metrischen Ebenen zu falschen Ergebnissen führen würden. Dieses Vorgehen ist in Abbildung 4.3 anhand eines Beispieles illustriert.

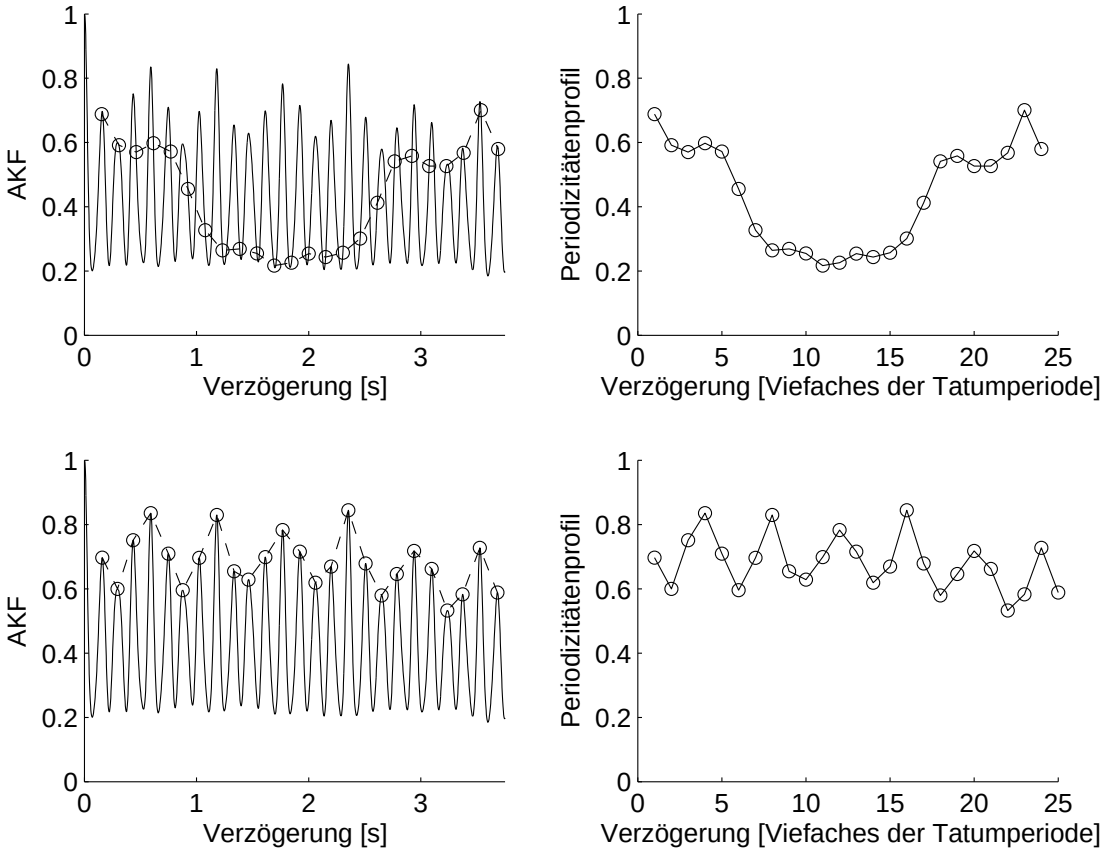


Abbildung 4.3: Periodizitätenfunktion (AKF, links) und Periodizitätenprofil V (rechts) mit äquidistanter Unterabtastung (oben) und mit Zuordnung der lokalen Maxima (unten).

Metrische Templates Die Ermittlung der metrischen Struktur geschieht durch einen Vergleich des Periodizitätenprofils V mit einer Anzahl von metrischen Templates R_i. In R_i ist in Form eines Periodizitätenprofils spezifiziert, wie stark die Periodizitäten für Verschiebungen von ganzzahligen Vielfachen der Tatumperiode ausgeprägt sind, welche Taktlänge als Vielfaches der Tatumperiode vorliegt, und welche Taktarten und welche

metrischen Ebenen als Zählzeit (beziehungsweise welche Mikrotimes $m_{R,i}$) möglich sind.

Zusätzlich wird eine A-priori-Wahrscheinlichkeit $P_{R,i}$ für das Auftreten von R_i angegeben. Tabelle 4.1 zeigt exemplarisch die Struktur der R_i. Die verwendeten R_i wurden durch Beobachtung des Datensatzes heuristisch ermittelt und die korrespondierenden Periodizitätenprofile V sind in Anhang A aufgelistet.

Taktlänge $[p_t]$	Mikrotime	Taktarten	P_R	V
16	2, 4	$\frac{4}{4}$	0.05	[0 0.2 0 0.3 0 0 0 0.4 0 ...]

Tabelle 4.1: Beschreibung der Struktur der metrischen Templates R_i anhand des Beispiels eines binären Rhythmus.

Die Übereinstimmung zwischen gemessenen V und R_i wird aus dem Korrelationskoeffizienten (nach Gleichung 4.14), durch Bildung des Skalarproduktes oder eines beliebigen Ähnlichkeits- oder Distanzmaß berechnet, wobei x und y die zu vergleichenden Vektoren sind.

$$X(n,k) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.14}$$

Alternativ kann ein Distanzmaß zum Vergleich verwendet werden. Die Länge des aktuellen V wird vor der Ähnlichkeitsberechnung der des R_i angepasst. Die Taktlänge wird aus dem ähnlichsten R_i ermittelt.

Bestimmung von Mikrotime, Beatperiode und Taktart Nach der Klassifikation der metrischen Struktur und Ermittlung der Taktlänge als Vielfaches der Tatumperioden p_t ist die Auswahl möglicher Mikrotimes, Beatperioden und Taktarten eingeschränkt. Mit einem probabilistischen Ansatz werden Mikrotime m und Beatperiode p_b nach Gleichung 4.15 und 4.16 ermittelt.

$$m = m_i \quad | \quad V(m_i) \cdot P(m_i) = \max(V(m_i) \cdot P(m_i)) \tag{4.15}$$

$$p_B = m \cdot p_T \tag{4.16}$$

mit der A-priori-Wahrscheinlichkeit P des Auftretens einer Beatperiode $p_B = m_i p_T$ nach der in Gleichung 2.1 beschriebenen Verteilung nach Parncutt [Par94] mit $\mu = 500$ und $\sigma = 0.2$.

Abweichungen der berechneten von der tatsächlichen Tatumperiode führen zu starken Ungenauigkeiten bei der Ermittlung der Beatperiode. Die Verbesserung der Genauigkeit der Temposchätzung wird durch die Auswertung des Beathistogramms erreicht.

4.2.5 Beathistogramm

Motivation Die zusätzliche Auswertung des Beathistogramms ist motiviert durch die Beobachtung, dass die auf dem Tatum basierende Analyse in verschiedenen musikalischen Situationen anfällig gegenüber Fehlern ist. Obwohl sie im Mittel eine erfolgreichere Analyse gestattet als Verfahren, welche nur Periodizitäten auf einer metrischen Ebene auswerten, versagt sie bei verschiedenen Stücken durch zum Beispiel Fehler in der Tatumschätzung oder der Berechnung und Auswertung der Periodizitätenprofile.

Ein weiteres Analyseverfahren ermöglicht in diesen Fällen, die durch niedrige Konfidenzwerte der Schätzung identifiziert werden sollen, eine robustere Schätzung des Tempos. Beathistogramme liefern eine von Tempooktavfehlern abgesehen präzise Schätzung in Musikstücken mit geraden Rhythmen und deutlichen Akzenten. Ein Beispiel zeigen die Abbildungen 4.4 und 4.5. Für zwei aufeinanderfolgende Segmente eines Musiksignals sind jeweils das Zeitsignal, das Akzentsignal und das Beathistogramm dargestellt.

Beobachtungen eines Testdatensatzes zeigen, dass in Fällen einer fehlerhaften Tatumanalyse häufig eine prominentere Periodizität auf der Beatebene auftritt. Des Weiteren ist in synkopierten Rhythmen die Ausprägung des Tatums deutlicher, um trotz der Betonungen zwischen den Zählzeiten die rhythmische Struktur zu erfassen.

Beathistogramme bieten weiterhin die Möglichkeit, die bei der in Abschnitt 4.2.4 und 4.3.5 beschriebenen Temposchätzungen auftretenden Rundungsfehler zu vermindern.

Berechnung des Beathistogramms Die hier verwendete Berechnung des Beathistogramms detektiert eine vorgegebene Anzahl der größten lokalen Maxima der Periodizitätenfunktion innerhalb des Tempobereiches und akkumuliert diese in einem Histogramm. Leichte Variationen des Tempos und Messfehler werden durch Faltung mit einem Hann-Fenster mit einer Länge von 100 ms ausgeglichen, so dass benachbarte Einträge zu einem mittleren Tempo beitragen. Das resultierende Beathistogramm wird mit der A-priori-Wahrscheinlichkeit $P(p_B)$ des Auftretens einer Beatperiode p_B nach Parncutt [Par94] (siehe Gleichung 2.1) gewichtet.

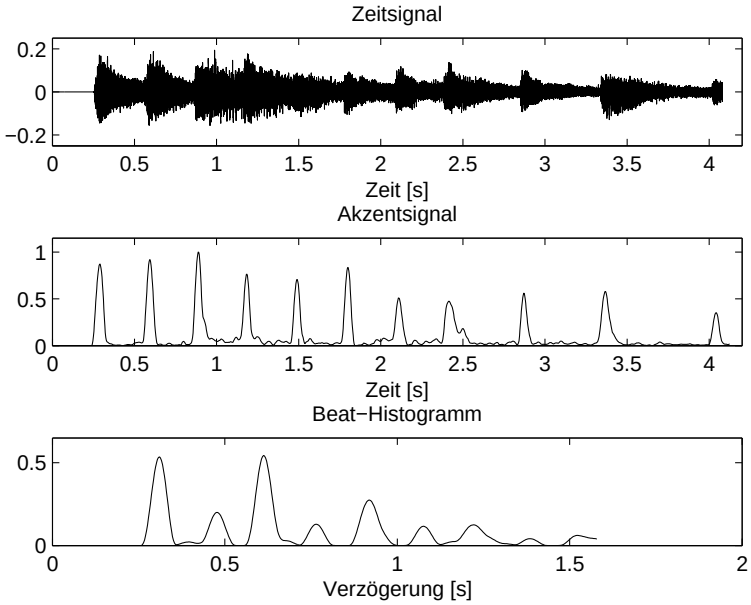


Abbildung 4.4: Zeitsignal, Akzentsignal und Beathistogramm eines Segmentes eines Musikstückes. Das globale Maximum des Beathistogramms bei 0.6 s weist auf das korrekte Tempo von 100 bpm hin.

4.2.6 Tatumtracking

Zur Ermittlung des Tatumrasters wird die in Abschnitt 3.3 beschriebene TWME-Methode verwendet, um ein Tatumraster mit in der Low-level-Analyse ermittelten Noteneinsätzen zu synchronisieren. Die Verarbeitung geschieht blockweise mit einer Blocklänge von ca. 2.5 s, die jedoch in Abhängigkeit der detektierten Noteneinsätze angepasst wird. Beim Auftreten von Ausschnitten ohne Noteneinsätze mit einer über einem Schwellwert liegenden Intensität wird ein Tatumraster bis zur letzten Note ermittelt und ab der nächsten Note fortgesetzt. Für den Bereich ohne für ein Tatumtracking genügende Noteninformation wird ein Tatumraster auf Grundlage des Akzentsignals ermittelt. Jedes ermittelte Tatumelement wird für die weitere Analyse mit einem Konfidenzmaß versehen.

4.2.7 Beattracking

Die Phase und Periode höherer metrischer Ebenen wird mit einem probabilistischen Ansatz anhand der Energie des Akzentsignals A_i in einer Umgebung der Tatumelemente i ermittelt. Ausgehend von der niedrigsten metrischen Ebene wird die nächsthöhere Ebene bestimmt, wobei die Wahrscheinlichkeit, dass Mikrotime m und Phase ϕ vorliegen

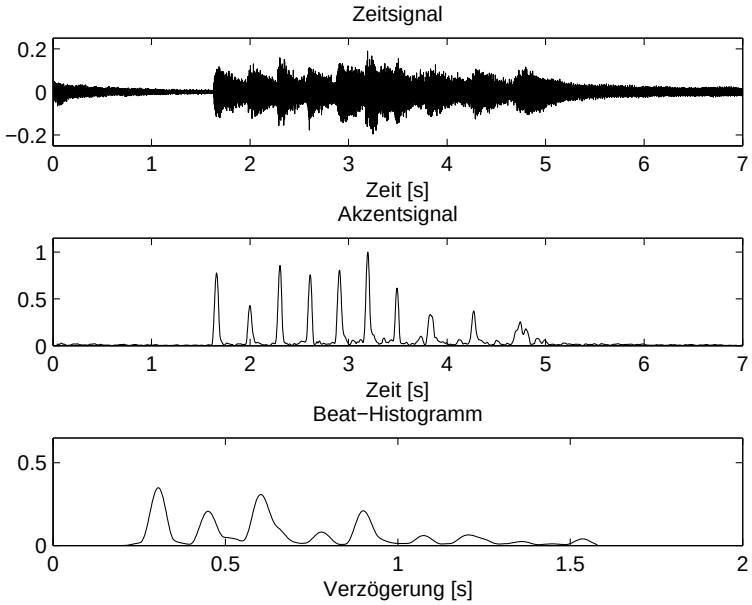


Abbildung 4.5: Zeitsignal, Akzentsignal und Beathistogramm eines Segmentes eines Musikstückes. Das globale Maximum des Beathistogramms bei 0.3 s korrespondiert zum Doppelten des Referenzwertes des Tempos.

nach Gleichung 4.17 berechnet wird.

$$P(m, \phi \mid A) = \frac{m}{N} \sum_{i}^{\frac{N}{m}} A_{im+\phi} \tag{4.17}$$

4.3 Extraktion und Auswertung von Instrumenteninformationen

Die High-level-Analyse schätzt die metrischen Merkmale unter Berückichtigung von Instrumenteninformationen. Die Instrumenteninformationen sind auf perkussive Instrumente ohne Tonhöheninformation beschränkt. Drumpattern werden als weiteres rhythmisches Merkmal detektiert. In Abbildung 4.6 ist das Blockschaltbild der High-level-Analyse illustriert.

In Abschnitt 4.3.1 wird das Verfahren zur Detektion der perkussiven Instrumente vorgestellt. Abschnitt 4.3.2 beschreibt die Quantisierung der Noteneinsätze. Die auf die resultierende Repräsentation angewandten Berechnungen von Periodizitäten sind in Abschnitt 4.3.3 erläutert. Abschnitt 4.3.4 legt das Verfahren zur Ermittlung wiederkehrender Muster der perkussiven Instrumente dar. Die Ermittlung der metrischen Merkmale ist in Abschnitt 4.3.5 beschrieben.

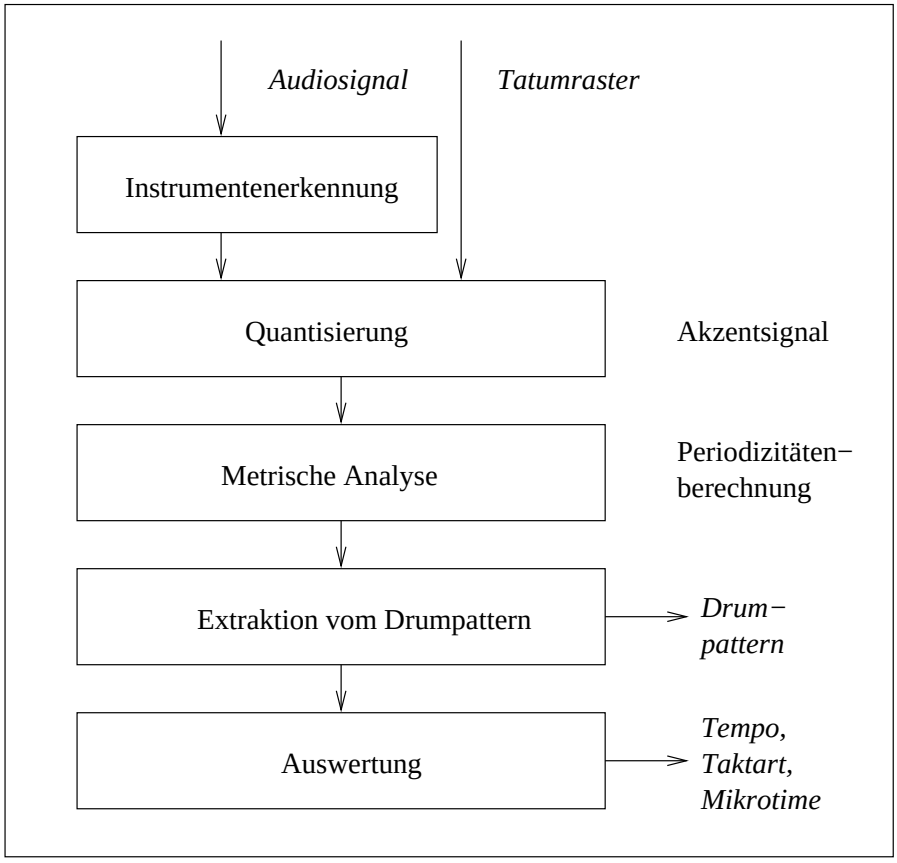


Abbildung 4.6: Blockdiagramm der High-level Analyse.

4.3.1 Detektion perkussiver Instrumente ohne Tonhöheninformation

Perkussive Instrumente ohne Tonhöheninformation spielen eine große Rolle für die Ausprägung der rhythmischen Struktur für eine Vielzahl von Musikstücken, insbesondere in den verschiedenen Kategorien populärer Musik. Deshalb ist die Kenntnis über das Auftreten von Klängen dieser Instrumente eine wertvolle Grundlage zur rhythmischen Analyse.

Die Detektion der perkussiven Instrumente ohne Tonhöheninformation geschieht in einer mit STFT berechneten Zeit-Frequenzdarstellung X des Audiosignals. Ein nicht-negatives Differenzspektrogramm $\widehat{X}$ wird durch Differenzierung und Einweggleichrichtung aus dem Spektrogramm X ermittelt. Ein Akzentsignal wird aus der Summe von $\widehat{X}$ über alle Bins ermittelt und dient zur Detektion von Noteneinsätzen. Zu allen Zeitpunkten eines Noteneinsatzes wird je ein Differenzenspektrum zur weiteren Analyse in einer Matrix $\widehat{X_t}$ akkumuliert.

Mit einer Hauptachsentransformation werden d dekorrelierte und varianznormierte

Komponenten $\widetilde{X}$ aus $\widehat{X_t}$ ermittelt (siehe Gleichung 4.18).

$$\widetilde{X} = \widehat{X_t} \cdot W \tag{4.18}$$

Die Transformationsmatrix W bewirkt eine Dimensionsreduktion, Dekorrelation und Varianznormierung.

Die Komponenten $\widetilde{X}$ werden mittels Non-Negative Independent Component Analysis (NNICA) [Plu94] entmischt, um die spektralen Profile der als unabhängig angenommenen Quellen zu ermitteln (siehe Gleichung 4.19)

$$F = A \cdot \widetilde{X} \tag{4.19}$$

Die Entmischungsmatrix A wird durch Minimierung des Auftretens der negativen Elemente in F durch die NNICA ermittelt. Die unabhängigen spektralen Profile F charakterisieren die im Signal auftretenden Klangquellen (siehe Abbildung 4.7) und werden im folgenden Berechnungsschritt zur Berechnung der Amplitudenbasisfunktionen (ABF) E verwendet (siehe Gleichung 4.20).

$$E = F \cdot X \tag{4.20}$$

Dieses Verfahren besitzt deutliche Parallelen zur Prior Subspace Analysis (PSA) [FLC03b]. Die Weiterentwicklung des hier anwandten Verfahrens gegenüber PSA liegt in der Extraktion der spektralen Profile aus dem Audiosignal anstelle der Verwendung a-priori angenommener Profile. Als weiterer Unterschied wird hier auf eine Entmischung der ABF mittels Independent Component Analysis (ICA) verzichtet, da die Voraussetzung der Unabhängigkeit der ABF in einem häufig von Synchronität geprägten musikalischen Kontext nicht gegeben ist. In Abbildung 4.8 sind Beispiele für ABF dargestellt.

Die extrahierten Quellen werden anhand von zeitlichen und spektralen Merkmalen in siebzehn Instrumentenklassen klassifiziert. Eine Vorklassifikation detektiert nichtperkussive Quellen und schließt diese von der weiteren Verarbeitung aus.

Die Vorklassifikation setzt die verbreitete Annahme voraus, dass perkussive Instrumente ohne Tonhöheninformation durch transiente Hüllkurvensignale und nichtharmonische Spektren gekennzeichnet sind. Ein von Transienten geprägter Verlauf der ABF wird mit dem in [UDS03] vorgestellten Perkussivitätsmerkmal detektiert. Das in [Set93] vorgestellte Dissonanzmaß wird in abgewandelter Form zur Auswertung der spektralen Profile berechnet. Die Zuordnung zu den Instrumentenklassen geschieht durch einen *Nearest-Neighbor*-Klassifikator durch Vergleich der spektralen Profile mit einem Trainingsdatensatz, mit einem auf dem Korrelationskoeffizienten basierenden Distanz- beziehungsweise Ähnlichkeitsmaß. In Fällen kleiner Konfidenzmaße werden zusätzliche Merkmale ausgewertet, die aus den spektralen Profilen berechnet werden. Dazu gehören spektrale Momente (Zentroid, Spreizung und Schiefe) und die Lage und Intensität von Partialtönen. Ein detaillierte Beschreibung des Verfahrens ist in [DU04] veröffentlicht.

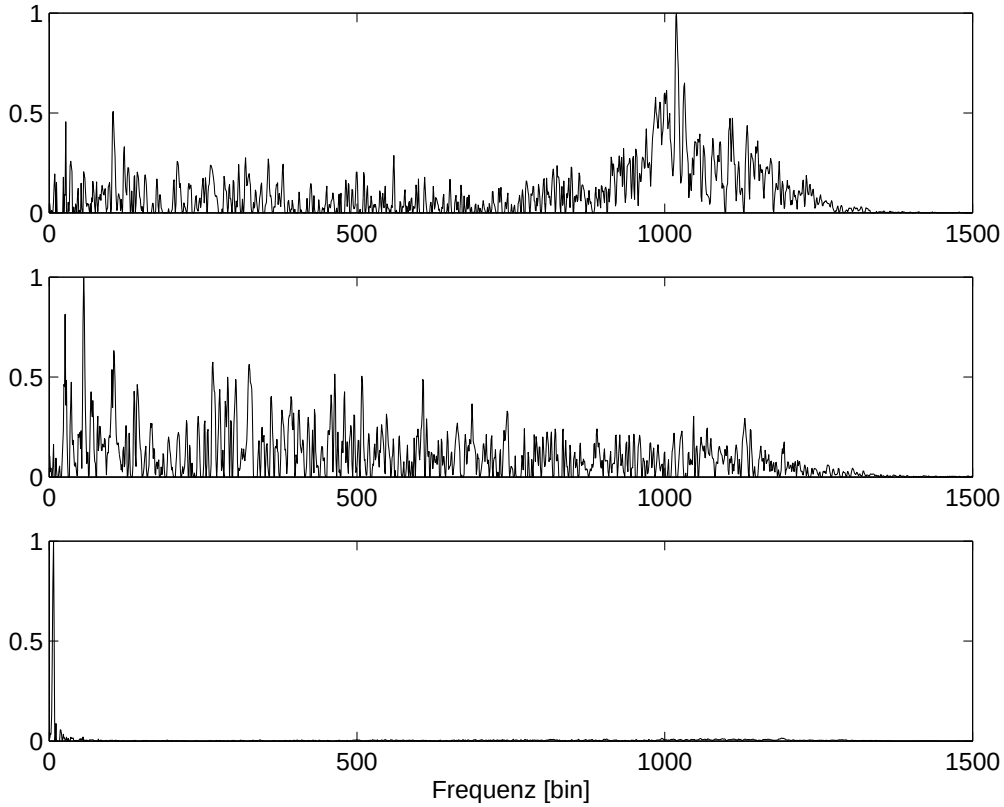


Abbildung 4.7: Spektrale Profile von separierten Quellsignalen: Hi-hat (oben), Snaredrum (mitte), Bassdrum (unten).

4.3.2 Quantisierung der Noteneinsätze

Zur Quantisierung der Noteneinsätze wird ein auf Basis von Low-level-Merkmalen geschätztes Tatumraster verwendet, dessen Extraktion in Abschnitt 4.2.6 beschrieben ist. Dieses Vorgehen ist äquivalent zu der in modernen MIDI-Sequenzern implementierten Quantisierungsfunktion für MIDI-Noten.

Das Ergebnis des Quantisierungsprozesses ist eine Matrixdarstellung $Q(i, j)$, $i = 1...n$ und $j = 1...m$, mit Anzahl der Tatumelemente n und Anzahl der Instrumente m. Diese Matrixdarstellung stellt das Äquivalent zum Akzentsignal für die Periodizitätenberechnung und -auswertung dar.

4.3.3 Periodizitätenberechnung in symbolischen Darstellungen

Die Länge eines Drumpatterns entspricht in der Regel einem GZV der Taktlänge [PK02]. Um eine größere Robustheit der Schätzung gegenüber Fehlern in den vorangegangenen metrischen Analysen zu erreichen, wird die Patternlänge auf Grundlage der Instrumen-

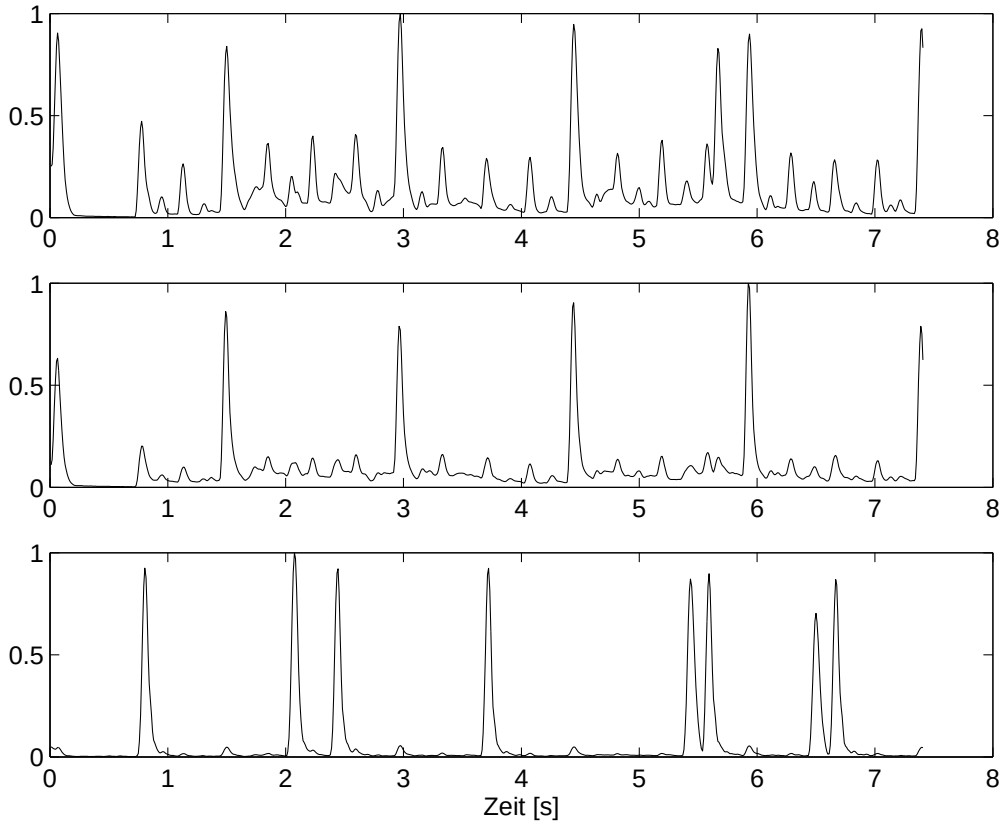


Abbildung 4.8: ABF von separierten Quellsignalen: Hi-hat (oben), Snaredrum (mitte), Bassdrum (unten).

teninformationen geschätzt.

Dazu wird eine Periodizitätenfunktion berechnet, die die Übereinstimmung zwischen dem Pattern und einer zeitlich verschobenen Kopie in Abhängigkeit von der Verschiebung darstellt.

Für die Ermittlung der Übereinstimmung wurden verschiedene Ähnlichkeits- und Distanzmaße untersucht. Die Hamming-Distanz ist ein aus der Informationswissenschaft bekanntes Distanzmaß für Boolsche Vektoren und wird aus der Anzahl unterschiedlicher Bits in den zu vergleichenden Vektoren b_1 und b_2 der Länge n berechnet (siehe Gleichung 4.21).

$$d_h = \sum_{i=1}^{n} b_1(i) \veebar b_2(i) \tag{4.21}$$

Eine geeignete Erweiterung stellt die unterschiedliche Gewichtung von gleichzeitig auftretenden Noten und Pausen dar, mit den Gewichten für Noten α, für Pausen β und für

unterschiedliche Belegung γ.

$$d_{h,w} = \sum_{i=1}^{n} \alpha \cdot [b_1(i) \wedge b_2(i)] + \beta \cdot [\neg b_1(i) \wedge \neg b_2(i)] + \gamma \cdot [b_1(i) \veebar b_2(i)], \quad \text{mit} \quad a, b < 0 \tag{4.22}$$

Es wurden weiterhin die L1- und L2-Norm als Distanzmaße sowie die AKF und SDF (siehe Gleichung 3.14) zur Berechnung der Periodizitäten untersucht. Eine weitere Option liegt in der Gewichtung in Abhängigkeit von Instrumenteninformationen und den Intensitäten beziehungsweise Lautstärkeinformationen. Vorrangegangene Arbeiten zeigten eine Verschlechterung der Erkennung unter Einbeziehung der Intensitäten in die Distanzberechnung [UD04b]. Die vom Detektionsverfahren extrahierten Intensitätswerte spiegeln auch nur begrenzt die perzeptuell wahrgenommene Lautstärke wieder. Des Weiteren wurde auch nicht untersucht, inwieweit Lautstärkeinformationen einzelner Ereignisse in komplexer Musik die Wahrnehmung von Rhythmus beeinflussen.

4.3.4 Identifikation charakteristischer Drumpattern

Nach der Ermittlung der Länge eines Drumpatterns aus einer Periodizitätendarstellung wird ein Histogramm berechnet, welches die Häufigkeit des Auftretens der einzelnen Instrumente an den metrischen Positionen darstellt.

Das nach Gleichung 4.23 aus der quantisierten Darstellung der Intrumenteneinsätze $Q(i,j)$, mit $i = 1...N$, ermittelte Patternhistogramm H_P dient der Entscheidung, ob ein Ereignis Bestandteil des Patterns der Länge L ist, oder eine Variation im Spiel darstellt.

$$H_P(i,j) = \sum_{k=0}^{R} Q(i + kL, j), \quad \text{mit} \quad R = \lfloor N/L \rfloor \tag{4.23}$$

Dabei ist R die Anzahl der Pattern zur Ermittlung des Patternhistogramms. Zur Ermittlung des Drumpatterns $M(i,j)$ wird das in Gleichung 4.24 dargestellte Schwellwertverfahren verwendet.

$$M(i,j) = \begin{cases} H_P(i,j) & | \quad H_P(i,j) > h \\ \\ 0 & \text{sonst} \end{cases} \tag{4.24}$$

Der Schwellwert h ist heuristisch zu ermitteln.

4.3.5 Ermittlung von Tempo, Taktart und Mikrotime

Die Schätzung der metrischen Merkmale beruht auf dem Abschnitt 4.2.4 beschriebenen Verfahren des Vergleichs einer Periodizitätendarstellung mit Metrischen Templates. Die robuste Ermittlung von Tempo und Taktart setzt die Anwendung von musikalischen

Regeln voraus [Pus05]. Häufig sind verschiedene Tempo- und Taktangaben für ein Musikstück möglich.

In [UD04b] ist deshalb vorgeschlagen worden, verschiedene markante Spielweisen, die aus Drumpattern ermittelt werden, in die Entscheidungsfindung einzubeziehen. Das äquidistante Auftreten von Snaredrum oder Handclaps wird in vielen populären Musikstilen, zum Beispiel in Pop, Rock und Swing auf der zweiten und vierten Zählzeit eines $\frac{4}{4}$-Taktes eingesetzt. In vielen Tanzmusikstilen ist die Platzierung der Bassdrum auf den Zählzeiten gebräuchlich. *Clave*-Pattern und Pattern mit Offbeat-Betonungen lassen Rückschlüsse auf die korrekte Tempooktave zu.

4.4 Ansätze zur redundanten Analyse

Ein einfaches Abstimmverfahren ermöglicht die Auswertung unterschiedlicher Verfahren zur Ermittlung der rhythmischen Merkmale. Dazu wird ein Histogramm $H(i)$ berechnet, in dem die die Konfidenzwerte der n Einzelschätzungen k_e aller m Segmente gewichtet mit einem Konfidenzmaß des Segments k_s akkumuliert werden (siehe Gleichung 4.25).

$$H(i) = \sum_{j=1}^{n} \sum_{l=1}^{m} q_{j,l}, \quad \text{mit} \quad q_{j,l} = \begin{cases} k_e(j) \cdot k_s(l) & | \quad i = e_{j,l} \\ 0 & \text{sonst} \end{cases} \tag{4.25}$$

Das Konfidenzmaß k_s des Segments q wird nach Gleichung 4.26 proportional zu seiner Länge und zum Mittelwert des Akzentsignals $A_q(i)$ des Segmentes der Länge n ermittelt.

$$k_s(q) = \sum_{i=1}^{n} A_q(i) \tag{4.26}$$

Das Konfidenzmaß der Low-level-Analyse $k_{e,LL}$ wird in Abhängigkeit der aus den vier Tatumschätzverfahren und den zur Schätzung analysierten Teilsegmenten ermittelten Tatumwerte berechnet, da eine zuverlässige Tatumschätzung die Voraussetzung für eine erfolgreiche metrische Analyse darstellt.

Die in Gleichung 4.27 notierte Berechnung wertet die relative Anzahl der innerhalb einer Toleranz korrekt ermittelten Werte n_c und der doppelt und halb ermittelten Werte n_d beziehungsweise n_h, die mit den Faktoren g_d beziehungsweise g_h gewichtet werden, aus.

$$k_{e,LL} = n_c + g_d \cdot n_d + g_h \cdot n_h \tag{4.27}$$

Das Konfidenzmaß der High-level-Analyse $k_{e,HL}$ wird nach Gleichung 4.28 aus Patternhistogramm H_P und Drumpattern M mit Patternlänge n und Anzahl der Instrumente m ermittelt, so dass das Konfidenzmaß sich proportional zu den im Pattern auf-

tretenden Noten verhält. Dadurch ist gewährleistet, dass kleine Konfidenzwerte ermittelt werden, wenn die Patternlänge falsch ermittelt wurde oder keine Pattern auftreten.

$$k_{e,HL} = \sum_{i=1}^{n}\sum_{j=1}^{m} M(i,j) / \sum_{i=1}^{n}\sum_{j=1}^{m} H_P(i,j) \tag{4.28}$$

Die Histogrammberechnung geschieht getrennt für die Auswertung der Tempo-, Mikrotime- und Taktartschätzung. Das Histogramm zur Temposchätzung wird mit einem Hanning-Fenster mit einer Länge von 12 bpm gefaltet, so dass benachbarte Tempohypothesen zu einem Wert beitragen.

4.5 Rhythmische Eingängigkeit und Intensität

Neben den bisher beschriebenen Merkmalen unterscheiden sich musikalische Rhythmen durch ihre Intensität und Komplexität beziehungsweise Eingängigkeit. Diese Merkmale sind in noch viel geringerem Maße durch eine einheitlich verbreitete Definition gekennzeichnet als die bisher beschriebenen Merkmale, wie zum Beispiel Metrum und Drumpattern.

Die Beschreibung von Intensität und Eingängigkeit ist ebenso wie die Beschreibung von Klang problematisch, da es sich um mehrdimensionale Größen handelt. In diese Arbeit wird jeweils eine Dimension der jeweiligen Merkmalsräume untersucht, ohne den gesamten Merkmalsraum zu spezifizieren. Zur Ermittlung der Eingängigkeit wird die Auswertung der aus der High-level-Analyse erhaltenen symbolischen Darstellung analysiert, die Intensität wird aus signalnahen Merkmalen ermittelt.

4.5.1 Eingängigkeit

Eine Reihe bekannter Ansätze ermittelt Komplexität in Abhängigkeit der metrischen Positionen, auf denen Noten auftreten [GM97a, SP98a, Pre04, Tou02]. In dieser Arbeit werden die Patternhistogramme H_P zur Ermittlung der Eingängigkeit ausgewertet. Diese ist auf Musikstücken mit binären Metren beschränkt, für die fünf metrische Ebenen angenommen werden (siehe Seite 16). Das in Abbildung 2.1 dargestellte Notenbild illustriert ein Beispiel für eine derartige metrische Struktur, deren metrische Ebenen durch die geschichteten Punke gekennzeichnet sind.

Es werden zwei Teilmaße E_m und E_p für die Eingängigkeit extrahiert, die anschließend gewichtet zusammengefasst werden. Die Eingängigkeit E_m wird aus den metrischen Positionen der Akzente so ermittelt, dass ein Akzent mehr zur Eingängigkeit beiträgt, je höher das metrische Niveau seines Auftretens ist. Dazu werden jede der fünf metrischen Ebenen heuristisch ermittelte Werte a_m verwendet, die zur Gewichtung der Noten des

Patternhistogramms H_P nach Gleichung 4.29 dienen.

$$E_m = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{m_i} \cdot H_P(i,j) \tag{4.29}$$

Dabei entsprechen n und k der Länge des Patternhistogramms beziehungsweise die Anzahl der Instrumente. Die Variable a_{m_i} bezeichnet das für die i-te Position gültige metrische Gewichte. Die Zugehörigkeit der metrische Gewichte zur Position des Patternhistogramms sind in Tabelle 4.2 dargestellt

p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	a_1	a_5	a_4	a_5	a_3	a_5	a_4	a_5	a_2	a_5	a_4	a_5	a_3	a_5	a_4	a_5

Tabelle 4.2: Metrische Gewichte a in Abhängigkeit von der Position p im Patternhistogramm in einem binaeren Rhythmus mit fünf Ebenen.

Die Bestimmung des Taktanfangs und die Zuordnung der Positionen im Patternhistogramm zu den metrischen Positionen geschieht durch Auswertung des Patternhistogramms H_P. Hierbei werden die Annahmen verfolgt, dass die erste Zählzeit einen großen Histogrammwert für das Instrument Bassdrum hat, einen großen Histogrammwert für alle Instrumente aufweist und die um ein ganzzahliges Vielfaches einer Viertelnote entfernten Zählzeiten große Histogrammwerte für das Instrument Snaredrum aufweist.

In Fällen, in denen keine Bassdrum beziehungsweise Snaredrum detektiert wurde, werden alternativ andere Instrumente ausgewertet, die entweder eine ähnliche Funktion in der Instrumentierung spielen, oder häufig im Falle einer Fehlklassifikation bei der Instrumentenerkennung mit Bassdrum beziehungsweise Snaredrum verwechselt werden, zum Beispiel Toms beziehungsweise Handclaps.

Die Berücksichtigung des Informationsgehaltes geschieht durch Auswertung der Ähnlichkeit zwischen Teilpattern der Längen $2^n, n = 1...3$, aus der das zweite Teilmaß E_p ermittelt wird. Die Ähnlichkeit wird durch Bildung des Skalarprodukts ermittelt. Diesem Vorgehen liegt die Annahme zugrund, dass die Eingängigkeit mit Ähnlichkeit

4.5.2 Intensität

Rhythmische Intensität beschreibt hier den Grad an perkussiven Klängen gegenüber Klängen mit wenig ausgeprägter Anschlagphase. Musik, bei der Rhythmus eine bedeutende Rolle spielt, ist häufig mit perkussiven Klängen instrumentiert sind. Diese perkussiven Klänge können tonalen Instrumenten entstammen, wie zum Beispiel bei *staccatos* und sind durch den Verlauf der Hüllkurve charakterisiert.

Zur automatisierten Ermittlung wird der Mittelwert des Akzentsignals berechnet. Informative Untersuchungen bezüglich der Auswertung der Instrumenteninformationen aus Abschnitt 4.3.1 verdeutlichen, dass diese Information nur für einen begrenzten Umfang der Musikstücke geeignet ist. Das Akzentsignal korreliert mit der Wahrnehmung von dynamischen Akzenten und kann aus dem Audiosignal robust extrahiert werden.

Die Wichtung der Akzentsignale der Teilbandsignale proportional zu ihrer Periodizität nach Gleichung 4.9 ist begründet durch die Aussage aus Abschnitt 2.1, Seite 16, dass Rhythmus intensiver empfunden wird, je gleichmäßiger er auftritt.

4.6 Verwendung der Merkmale in einer MIR-Anwendung

Die in dieser Arbeit betrachteten Eigenschaften eines Musiksignals können, eine zuverlässige Extraktion vorausgesetzt, für verschiedene MIR-Anwendungen von Bedeutung sein. Ein Beispiel für eine Anwendung ist die Klassifikation des musikalischen Genres [UD04a, GDPW04].

In vielen populären Musikstilen werden Schlagzeuginstrumente eingesetzt. Hier treten häufig Drumpattern auf, die charakteristisch für das musikalische Genre sind. Auf Grundlage dieser Drumpattern wird eine Klassifikation des musikalischen Genres durchgeführt, die auf Grund der Unvollständigkeit der ausgewerteten Informationen in ihrer Leistungsfähigkeit beschränkt ist, jedoch im Zusammenhang mit der Auswertung von melodischen, harmonische und klanglichen Informationen leistungsfähig sein kann.

Die automatisiert extrahierten Drumpattern werden bezüglich verschiedener Spielcharakteristiken (SC) analysiert, die zusammen mit den Instrumenten und metrischen Informationen die Merkmale für die Klassifikation liefern. Diese Charakteristiken sind auf Grund von Erfahrungswerten nach ihrer Diskriminanzfähigkeit für das Klassifikationsproblem ausgewählt und in Tabelle 4.3 dargestellt. Die Namensgebung orientiert sich zum Teil an gebräuchlichen Anglizismen.

Die Instrumenteninformationen werden zu einem weiteren Merkmal zusammengefasst, welches fünf Zustände annehmen kann, indem die auftretenden Instrumente genretypischen Kombinationen zugeordnet werden, die hier Drumsettypen (DST) genannt werden und in Tabelle 4.4 aufgeführt sind.

Diese DST werden nicht nach den klanglichen Eigenschaften der Instrumente zugeordnet, sondern allein nach dem Auftreten der Instrumente. Wichtige Informationen, die ein Zuhörer auswerten kann, werden dabei vernachlässigt, da der Klang eines bestimmten Instrumententypes in der Regel genrespezifisch ausgeprägt sein kann.

Zur Klassifikation wird ein regelbasierter Klassifikator eingesetzt, der durch Erfahrungswerte und die Auswertung repräsentativer Beispiele trainiert wurde. Die antrainierten Klassen stellen keine vollständige Taxonomie dar, da nur Genres berücksichtigt werden, in denen der Einsatz von Drumpattern gebräuchlich ist. Eine Übersicht der Klassen und der Merkmalsausprägung ist in Tabelle 4.5 dargestellt.

Charakteristik	Beschreibung	Instrument
a) Four-on-the-floor 1	Ereignisse treten ausschließlich zu jeder Hauptzählzeit auf	Bassdrum
b) Four-on-the-floor 2	Ereignisse treten zu jeder Hauptzählzeit auf	Bassdrum
c) Backbeat 1	Ereignisse treten ausschließlich zu jeder zweiten und vierten Hauptzählzeit eines $\frac{4}{4}$-Taktes auf	Snaredrum, Handclaps
d) Backbeat 2	Ereignisse treten zu jeder zweiten und vierten Hauptzählzeit eines $\frac{4}{4}$-Taktes auf	Snaredrum, Handclaps
e) Offbeat 1	Ereignisse treten zu jedem Offbeat auf	Idiophone
f) Offbeat 2	Ereignisse treten zu jedem Offbeat auf	Snaredrum
g) Synkopation der Idiophone	Verhältnis der Anzahl der zu Zwischenzählzeiten und Zählzeiten auftretenden Ereignisse	Idiophone
h) Synkopation der Membranophone	Verhältnis der Anzahl der zu Zwischenzählzeiten und Zählzeiten auftretenden Ereignisse	Membranophone
i) Doublebass	Ereignisse treten an aufeinander folgenden Tatumpositionen auf	Bassdrum
j) Clavepattern	Mehr als ein IOI entspricht drei Tatumperioden	Idiophone
k) Swingpattern	Ereignisse treten auf der ersten und dritten Note einer Achteltriole auf	Idiophone
l) Bassdrumvariation	Anzahl unterschiedlicher IOI	Bassdrum

Tabelle 4.3: Charakteristische Merkmale *SC* von Drumpattern.

Rock	Jazz	Latin	Disco	Techno
Bassdrum, Snaredrum, Hi-hat, Becken	Bassdrum, Snaredrum, Becken	Bongo, Conga, Claves, Shaker	Bassdrum, Snaredrum, Hi-hat, Conga, Handclap	Bassdrum, Hi-hat

Tabelle 4.4: Taxonomie und Beschreibung von Drumsettypen *DST*.

Genre	DST	Tempo [bpm]	Taktart	Mikro-time	SC
Disco / House (DH)	Disco	115-132	$\frac{4}{4}$	2	a,c,e
HipHop / R'nB (HR)	Rock	70-120	$\frac{4}{4}$	2	c,d,l
Soul / Funk (SF)	Disco	85-130	$\frac{4}{4}$	2	c,d,f,g
Drum'nBass (DB)		160-200	$\frac{4}{4}$	2	c,d,f,l
Jazz / Swing (JS)	Jazz	60-300	$\frac{4}{4}$	3	c,k,l
Rock / Pop (RP)	Rock	80-180	$\frac{4}{4}$	2	a,b,c,d
Heavy Metal (HM)	Rock	140-260	$\frac{4}{4}$	2,3	i
Latin (LA)	Latin	70-170	$\frac{4}{4}$	2	f,g,j
Walzer (WA)		40-200	$\frac{3}{4}$	2	
Polka / Punk (PP)	Rock	140-240	$\frac{4}{4}$	2	a,b,h
Techno (TE)	Techno	132-165	$\frac{4}{4}$	2	a,f

Tabelle 4.5: Klassen zur Genreerkennung und Merkmalsausprägung.

5 Evaluierung der Verfahren

Nach Temperley existieren die vier folgenden Anforderungen an ein System zur Evaluierung von Verfahren zur Extraktion von Informationen [Tem04]:

- das Vorliegen einer vereinbarten Repräsentation der zu extrahierenden Informationen zum Vergleich der Ergebnisse
- die Auswertung eines bezüglich Umfang und Charakteristik repräsentativen Datensatzes
- eine korrekte Analyse des Testdatensatzes zur Gewinnung der Referenzdaten
- eine vereinbarte Methode zum Vergleich von Referenz und Analyseergebnis sowie zur Ermittlung einer Bewertung

Ein grundsätzliches Problem bei der Evaluierung von Analysealgorithmen ist, dass die in Veröffentlichungen berichteten Ergebnisse mit wenigen Ausnahmen auf der Auswertung von unterschiedlichen Testdatensätzen beruhen. Neben Unklarheiten bezüglich urheberrechtlicher Rahmenbedingungen ist eine weitere Ursache für die fehlende Vereinheitlichung der Testdaten und Evaluierungsmethoden die Unterschiedlichkeit der anvisierten Anwendungen [Dix01b]. Die von der MIR-Gemeinschaft organisierten offenen Vergleiche *ADC 2004* und *MIREX 2005*[1] setzten diesbezüglich positive Akzente für die Durchsetzung einheitlicher und repräsentativer Vergleichsmethoden [Dow05].

Im Abschnitt 5.1 wird das Verfahren zur Detektion von NEZ evaluiert. Abschnitt 5.2 stellt die Ergebnisse der metrischen Analyse dar. Die Bewertung der automatischen Extraktion von Drumpattern ist in Abschnitt 5.3 beschrieben. Die Evaluierung der Ermittlung von Eingängigkeit und Intensität ist in Abschnitt 5.4 dargestellt. Die Schätzung des musikalischen Genres als beispielhafte Anwendung wird in Abschnitt 5.5 bewertet. Dieses Kapitel schließt mit Abschnitt 5.6, in dem der Vergleich mit anderen Verfahren zur Temposchätzung und zum Beattracking anhand der Ergebnisse des *ADC 2004* und des *MIREX 2005* beschrieben ist.

[1] *Music Information Retrieval Evaluation Exchange* (MIREX 2005) ist der anlässlich der ISMIR 2005 veranstaltete offene Vergleich von Verfahren zur automatisierten Musikanalyse.

5.1 Noteneinsatzdetektion

Vorüberlegungen zur Evaluierung der NEZ-Detektion Die Evaluierung automatisierter Detektionsverfahren erfordert einen Testdatensatz mit manuell annotierten NEZ. Die Referenzwerte sollten aufgrund der Subjektivität der Aufgabe von mehreren Testhörern, unterstützt durch visuelle Beobachtung des Zeitsignals und eines Spektrogramms, ermittelt werden. In Untersuchungen zur paarweisen Abweichung zwischen den von Testhörern annotierten NEZ mit drei Teilnehmern und einem Datensatz mit 750 Noten wurde eine mittlere Differenz von 10.5 ms zwischen den von den Testhörern annotierten NEZ identifiziert [LDR04].

Alternativ können Testdaten aus MIDI-Dateien erstellt und die NEZ direkt ausgelesen werden, wobei jedoch sichergestellt werden muss, dass die Noten für einen Zuhörer detektierbar sind.

Unkontrollierte und nicht beabsichtigte Klänge wie zum Beispiel Atem-, Klappen- und Greifgeräusche können bei der Klangproduktion auftreten [Fle03]. Da diese Geräusche in der Regel keine musikalische Bedeutung besitzen, sollten sie idealerweise nicht detektiert [LDR04], jedoch eventuell zu Evaluierungszwecken in den Testdaten erfasst werden.

Beschreibung der Datensätze Zur Evaluierung der Notendetektion stehen 130 Musikstücke mit insgesamt 1824 manuell annotierten NEZ zur Verfügung. Die mittlere Länge eines Musikstückes beträgt 2,4 s. Die Ermittlung der NEZ wurde von einem Testhörer unterstützt durch visuelle Beobachtung des Zeitsignals und des Spektrogramms durchgeführt. In unsicheren Fällen wurde ein zweiter Testhörer (Experte) zu Rate gezogen. Die Teststücke sind unterteilt nach polyphoner Musik (99 Stücke mit 739 Noten und einer mittleren Länge von 1,61 s) und perkussiver Musik (31 Stücke mit 1085 Noten und einer mittleren Länge von 4,95 s).

Ergebnisse Für den Vergleich zwischen manuell annotierten und detektierten NEZ wurde eine Toleranz von 20 ms gewählt. Die Ergebnisse sind in Tabelle 5.1 dargestellt. Fehldetektionen sind in die Kategorien *falsch positive* Erkennung f_p (Detektion einer nicht vorhandenen Note) und *falsch negative* Erkennung f_n (keine Detektion einer vorhandenen Note) unterteilt. Zur Bewertung des Klassifikationsergebnisses sind *Precision* P, *Recall* R und *F-Score* F (siehe Gleichung 5.1 bis 5.3, mit Richtigklassifikation t_p) angegeben.

$$P = t_p/(t_p + f_p) \tag{5.1}$$

$$R = t_p/(t_p + f_n) \tag{5.2}$$

$$F = 2 \cdot P \cdot R/(P + R) \tag{5.3}$$

	f_p	f_n	P	R	F
polyphon	12%	7%	0.88	0.93	0.90
perkussiv	0.6%	15%	0.99	0.87	0.93
gesamt	5.3%	12%	0.95	0.89	0.92

Tabelle 5.1: Resultate der NEZ-Detektion für den gesamten Datensatz und getrennt für polyphone und perkussive Musik. Die Anzahl der Fehlklassifikationen ist relativ zur Anzahl der manuell detektierten Noten angegeben, weiterhin sind Precision P, Recall R und F-Score F aufgelistet.

Die Ergebnisse der Notendetektion entsprechen dem Stand der Technik, wie der Vergleich mit den Ergebnissen des *MIREX 2005* der Disziplin *Audio Onset Detection* zeigt [MIR05].

Zu den verschiedenen zu Fehldetektionen führenden Ursachen gehören die Wahl fester Parameter und Schwellwerte, deren Auswirkung beim Vergleich der Ergebnisse für polyphone und perkussive Musik deutlich wird. Obwohl in der Regel in perkussiver Musik die Noteneinsätze leichter zu detektieren sind, sind die Unterschiede in der Erkennungsleistung zwischen perkussiver und polyphoner Musik nur gering, da die Parameterauswahl für polyphone Musik getroffen wurde.

Bei perkussiver Musik ist die Rate von NEZ deutlich größer als in polyphoner Musik. In polyphoner Musik treten verstärkt falsch positive Detektionen bedingt durch längere Ausklingphasen und Geräusche auf, die nicht als Noten manuell annotiert wurden. Ein anschauliches Beispiel dafür ist ein ausgehaltener synthetischer Sound mit Modulationen in Frequenz oder Amplitude.

Der Notenbegriff ist nicht für alle Klangerzeuger vollständig definiert und nicht allein durch Energieanstieg und Phasenkongruenz gekennzeichnet. Die Detektion von Noten scheint beim Menschen durch Erfahrungswerte geleitet zu sein und ist durch das Wiedererkennen einer Klangquelle und musikalische Kenntnisse beeinflusst. Da die Detektion der NEZ als Vorverarbeitung zur Analyse des Rhythmus und nicht einer Transkription des Musikstückes dient, sind die Fehldetektionen jedoch in Maßen tolerierbar.

5.2 Metrische Analyse

Schätzung der Tatumperiode

Zur Evaluierung der Schätzung der Tatumperiode wird von 83 Musikstücken ein Ausschnitt von 10 s analysiert. Die Referenzwerte für die Tatumperioden wurden von zwei Testhörern unter zusätzlicher Verwendung des automatischen Schätzverfahrens ermittelt. In Fällen von korrektem Tatumtracking wurde das maschinell ermittelte Ergebnis als Referenzwert gesetzt.

In Tabelle 5.2 sind die Ergebnisse der einzelnen Schätzverfahren und das Gesamtergebnis zusammengestellt. Das Verfahren des GGT der IOI zeigte die besten Ergebnisse, die Mehrzahl der Verfahren tendierte zu langsameren Tempos. Die Ergebnisse des erfolgreichsten Verfahrens werden durch das Abstimmungsverfahren verbessert.

	TWME-NEZ	GGT-NEZ	TWME-AKF	DFT-AKF	VOTING
richtig	57.83	67.47	55.42	55.42	77.11
doppelt	9.64	1.2	3.61	1.2	1.2
halb	9.64	14.46	20.48	10.84	10.84
falsch	22.89	16.87	20.48	32.53	10.84

Tabelle 5.2: Resultate der Tatumperioden-Schätzung für die einzelnen Schätzverfahren (TWME-NEZ, TWME-GCD, TWME-AKF, DFT-AKF) und das durch eine Abstimmungsmethode ermittelte Gesamtergebnis (VOTING) in Prozent.

Temposchätzung

Beschreibung der Referenzdaten Obwohl für die Mehrzahl der Musikstücke einer Teilmenge in Test- und Trainingsdatenbank übereinstimmende Werte für Tempo und Taktart von unterschiedlichen Zuhörern ermittelt wurden, so gibt es eine Reihe von Gegenbeispielen, deren Tempo unterschiedlich empfunden wird. Mit einer Auswahl an Teststücken wurde ein informeller Hörtest bezüglich der Ermittlung des Tempos mit acht Versuchspersonen (VPn) durchgeführt. Die Ergebnisse der Tempoermittlung sind in Tabelle B.1 im Anhang B dargestellt. Bei der Auswahl der zum Hörtest verwendeten Musikstücke wurde gezielt nach Stücken gesucht, für welche auf Grund der metrischen

Struktur unterschiedliche Temposchätzungen der Zuhörer erwartet wurden.

Die Stücke 1 bis 3 sind Swingstücke und werden aufgrund ihres schnellen Tempos und der damit verbundenen Abweichung zum moderaten Tempo häufig in der langsameren Tempooktave eingestuft. VP D tendierte auch bei moderaten Tempos zur langsameren Tempooktave (siehe Stück 4 bis 6), VP G zur schnelleren Tempooktave (7). Bei nur einem Stück (8) stimmten alle Zuhörer überein.

Neben der Wahl unterschiedlicher Tempooktaven führt auch die Auswahl verschiedener metrischer Ebenen in ternären Metren als tempogebend zu unterschiedlichen Angaben. Zwei Ebenen kamen bei Stück 9 bis 14 als Zählzeit in Betracht, jeweils 3 Ebenen für die restlichen Teststücke.

Die Ergebnisse zeigen, dass Referenzwerte nicht in allen Fällen eindeutig bestimmbar sind. Die in Notenblättern dargestellten Tempoangaben sind Spielanleitungen, nicht zwingend das Tempo, welches Musiker und Zuhörer empfinden, da der betreffende Notenwert nicht unbedingt mit dem gefühlten Puls übereinstimmt [Pus05].

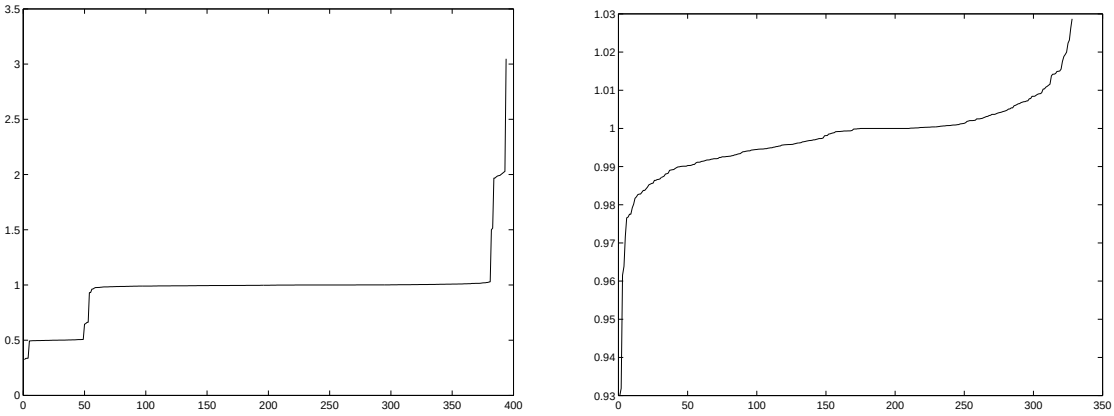
Als Testdaten zur Evaluierung stehen unter anderem die Datensätze des *ADC 2004* zur Verfügung. Auch hier verdeutlicht der Vergleich der von unterschiedlichen Testhörern ermittelten Tempowerte die Unschärfe, die bei der Ermittlung der Referenzdaten auftritt. Die Tempowerte einer Teilmenge des Datensatzes[2] mit 450 Stücken wurden erneut vom Autor manuell ermittelt und mit den ursprünglichen Referenzdaten verglichen. Die Verhältnisse r_{1i}/r_{2i} aus ursprünglichem Referenzwert vom Testhörer 1 r_{1i} und nachträglich annotiertem Referenzwert vom Testhörer 2 r_{2i} des $i-ten$ Stückes sind in Abbildung 5.1 größensortiert dargestellt.

Für $83,25\%$ aller Stücke wurde die identische metrische Ebene als Zählzeit ermittelt, Testhörer 2 ermittelte häufiger die höhere Tempooktave ($11,42\%$) als die tiefere Tempooktave ($2,54\%$) verglichen mit Testhörer 1. In Stücke mit ternären Metren wurde ein gerundetes Verhältnis von 3 beziehungsweise $\frac{1}{3}$ für $0,25\%$ beziehungsweise $1,01\%$ der Stücke ermittelt, welches der nächsthöheren oder -tieferen metrischen Ebene entspricht. Weitere auftretende Kategorien für unterschiedliche Ergebnisse sind durch ein Verhältnis von $\frac{3}{2}$ ($0,51\%$) und $\frac{2}{3}$ ($1,01\%$) gekennzeichnet.

Eine genauere Inspektion der Stücke zeigte eine deutliche Abhängigkeit der Unterschiede vom musikalischen Genre. Testhörer 1 ordnete Swing- und Drum&Bass[3] nahezu vollständig und lateinamerikanische Musik häufig eine Tempooktave niedriger als Testhörer 2 ein. Die Tempoangaben für identische metrische Ebenen weichen bis zu 7%

[2] Der Datensatz beinhaltet Musikstücke der Kategorien „Tanzmusik“, „Songs“ und „Loops“. Die Kategorie „Tanzmusik“ beinhaltet Stücke mit sehr ähnlicher Instrumentierung und Rhythmik und scheint deshalb für die Evaluierung nicht repräsentativ. Die Kategorie „Loops“ umfasst kurze Musikstücke mit konstantem Tempo, die in der Musikproduktion als Baustein verwendet werden und in der Regel nicht als Musikstück wahrgenommen werden. Auch diese Daten sind auf Grund ihrer Ausrichtung auf wenige musikalische Genres und ihrer Länge nicht repräsentativ. Lediglich die Musikstücke aus der Kategorie „Songs“ stellen einen gültigen Datensatz dar.

[3] „Drum&Bass“ bezeichnet ein Genre elektronischer Tanzmusik.



(a) (b)

Abbildung 5.1: Darstellung der größensortierten Verhältnisse zweier ermittelter Referenzwerte eines Testdatensatzes mit 394 Musikstücken. Die Werte für den gesamten Datensatz sind in Abbildung a) dargestellt, Abbildung b) zeigt die Teilmenge an, für die beide Referenzwerte identische metrische Ebenen beschreiben.

voneinander ab.

Wie sollte eine Temposchätzung bewertet werden? Sind verschiedene Tempi in mehrdeutigen Fällen richtig?

Die Antwort auf diese Frage wird sicher durch die Anwendung gegeben. Für MIR-Applikationen wie zum Beispiel der Suche nach ähnlichen Musikstücken ist die Kenntnis aller metrischen Ebenen von Vorteil. Die Vorschläge zur Evaluierungsprozedur zum *MIREX 2005* tendieren zur Erfassung von Referenzwerten von mehreren Zuhörern, aus deren Verteilung die zwei wahrscheinlichsten Tempoangaben zur Evaluierung herangezogen werden.

Experimente zur Parametereinstellung Zur Auswahl der Parameter der Temposchätzung wurde ein Trainingsdatensatz mit 490 Musikstücken mit einer Länge von jeweils 25 s verwendet. Die Referenzwerte wurden durch Messung des Tempos des Mitklopfens ermittelt. In Zweifelsfällen wurden weitere Testhörer zu Rate gezogen.

Die Erkennungsraten sind in Abhängigkeit der Parameterwahl im Anhang C angegeben. Zur Bewertung wurden zur Übersicht die Anzahl der richtig und falsch ermittelten Tempi E_r beziehungsweise E_f und das Verhältnis der Tempooktavfehler E_d/E_h dargestellt, mit Anzahl der Ermittlung des doppelten und halben Tempos E_d beziehungsweise E_h.

Zu den untersuchten und hier dokumentierten Parametern gehören

- die Konstante c zur nichtlinearen Beeinflussung der Wichtung der Teilbänder bei der Bildung des Akzentsignals in Abhängigkeit von ihrer Periodizität (siehe Gleichung 4.9)
- der Einfluss der Konstante d bei der Wichtung der Tatumschätzungen zur Ermittlung der Tatumperiode aus mehreren Schätzverfahren (siehe Gleichung 4.11)
- der Einfluss der A-priori-Wahrscheinlichkeit für das Auftreten der Beatperiode nach Parncutt [Par94] (siehe Seite 70 und Gleichung 2.1)
- der Einfluss des Ähnlichkeits- beziehungsweise Distanzmaßes zur Berechnung der Periodizität in symbolischen Darstellungen (siehe Seite 78)

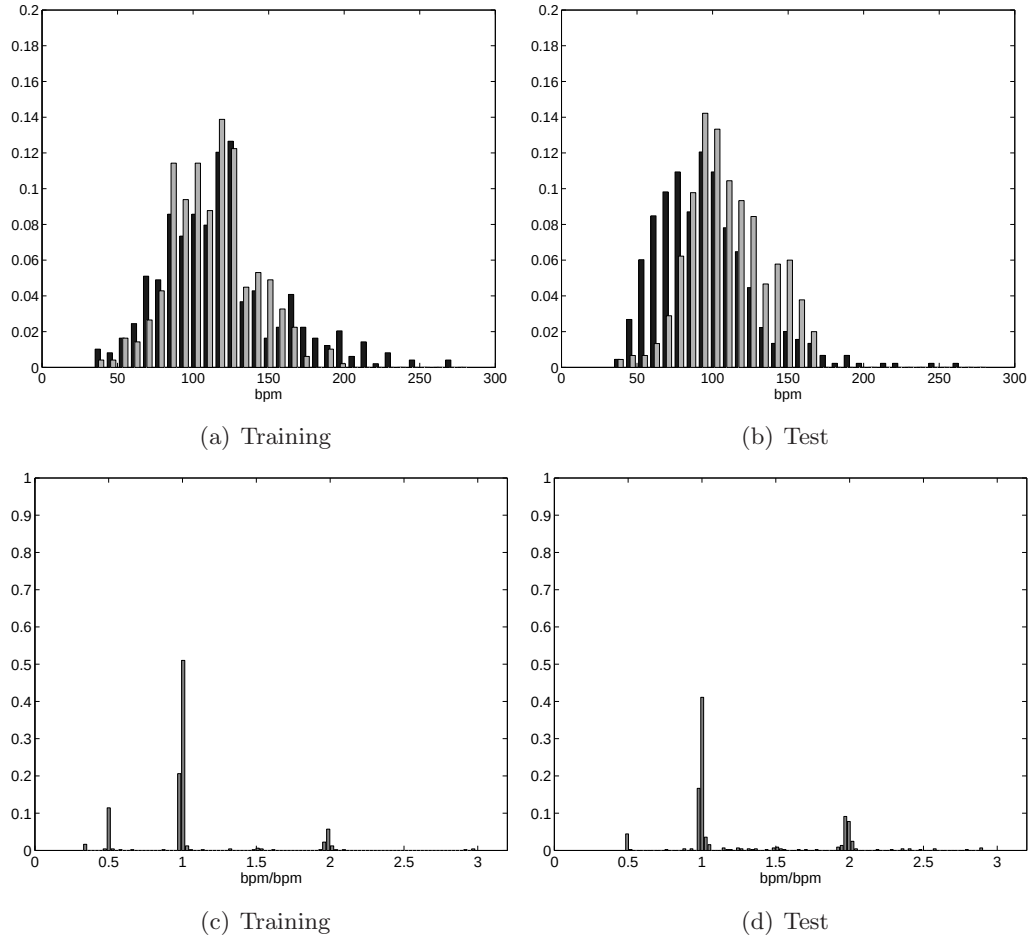
Weiterhin wird die Auswirkung der Berechnung der Periodizität der Summe der Teilbänder gegenüber der Summe der Periodizitäten der Teilbänder untersucht (siehe Seite 68).

Ergebnisse Die Evaluierung der Temposchätzung geschieht anhand des Testdatensatzes mit 450 Musikstücken mit einer mittleren Länge von 20.4 s. Die vom Autor ermittelten Referenzwerte wurden anstelle der originalen Referenzwerte verwendet, um Konsistenz mit dem Trainingsdatensatz zu bewahren, da systematische Abweichungen bei der Tempobestimmung auftraten. Bei der Bewertung wurden Abweichungen von maximal 4% toleriert.

In Tabelle 5.3 sind die Ergebnisse der Temposchätzung nach Auswahl geeigneter Werte für die im vorangegangenen Abschnitt untersuchten Parameter für die drei Einzelverfahren und das Gesamtergebnis anhand der Anzahl der Richtig- und Falschklassifikationen für Trainings- und Testdatensatz aufgelistet.

	Low-level		Beathistogramm		High-level		Voting	
	E_r	E_f	E_r	E_f	E_r	E_f	E_r	E_f
Training	70.82	6.73	64.08	13.88	50.61	25.71	72.86	5.51
Test	58.44	13.78	41.78	31.56	32.00	50.44	61.56	12.22

Tabelle 5.3: Erkennungsrate der Temposchätzung für Trainings- und Testdatensatz der einzelnen Verfahren in Prozent.



(a) Training (b) Test

(c) Training (d) Test

Abbildung 5.2: Evaluierung der Temposchätzung: a) und b) Verteilung der Referenz- (dunkel) und Schätzwerte (hell), c) und d) Histogramm der Verhältnisse aus Schätzung und Referenzwert.

Der Unterschied zwischen den Ergebnissen der Verarbeitung der Trainings- und Testdaten ist auch durch die Zusammenstellung der Datensätze bedingt. Eine genauere Betrachtung beider Datensätze zeigt, dass der Testdatensatz eine größere Herausforderung an den Analysealgorithmus stellte auf Grund der größeren Anzahl von Musikstücken mit

- expressivem Tempo
- schlechterer Audioqualität
- wenig perkussiven Elementen

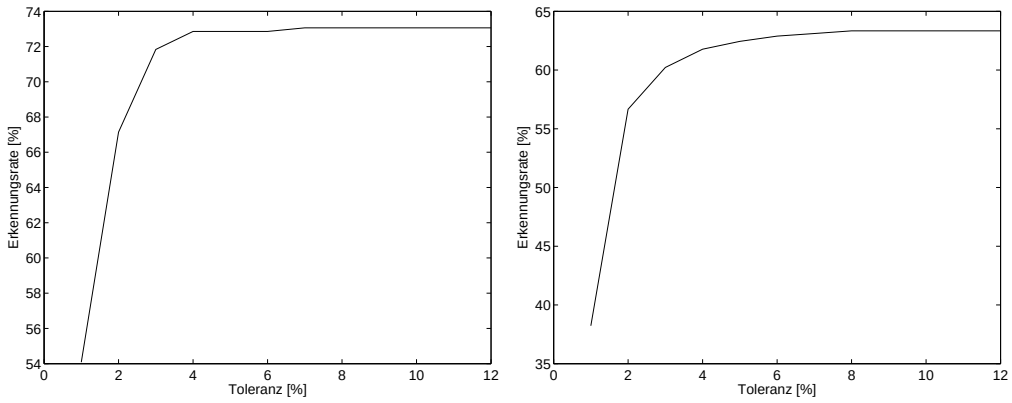
Weiterhin ist im Testdatensatz griechische und osteuropäische Folklore mit 31% deutlich überrepräsentiert [GAD+06], das Genre „Rock/Pop“ mit 15% hingegen nur wenig

vertreten.

Die Ergebnisse der Temposchätzung der Low-level-Analyse sind anhand einer Histogrammdarstellung der Werte und der Verhältnisse des Schätzwertes zum Referenzwert in Abbildung 5.2 dargestellt.

Neben Tempooktavfehlern treten Verwechselungen der metrischen Ebenen auf durch Deutung der Takt- oder Tatumebene als Zählzeit oder Verwechselung der Zählzeit mit anderen metrischen Ebenen in ternären Takten. Dazu gehören die Schätzung eines schnellen $\frac{3}{4}$-Taktes als langsamer Takt mit ternärer Mikrotime, die Schätzung eines langsamen Taktes mit ternärer Mikrotime als schnellen $\frac{3}{4}$-Takt, die Ermittlung des 1.5-fachen Tempos und binärer Mikrotime bei Musikstücken mit ternärer Mikrotime und die Ermittlung des 0.66-fachen Tempos und ternärer Mikrotime bei Musikstücken mit binärer Mikrotime.

Die Anzahl der Richtig- und Falschklassifikation in Abhängigkeit von der in der Auswertung zugelassenen Toleranz zwischen Referenzwert und Schätzwert illustriert Abbildung 5.3.



(a) Training (b) Test

Abbildung 5.3: Abhängigkeit der Erkennungsrate der Temposchätzung von der Toleranz der Bewertung.

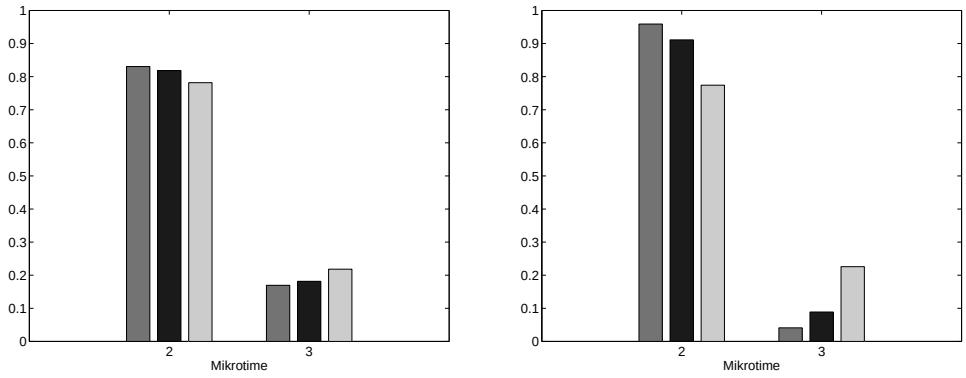
In Tabelle 5.4 ist der Einfluss der Segmentierung auf die Temposchätzung dargestellt. Dazu werden die Erkennungsraten verglichen, die aus der Analyse des repräsentativsten Segments und des gesamten Audiosignals ermittelt werden. Als repräsentativstes Segment wird das Segment ermittelt, für welches die Summe des Akzentsignals maximal ist. Die Temposchätzung der Low-level- und High-level-Analyse profitieren von der Segmentierung.

	Low-level		Beathistogramm		High-level	
	mit Seg.	ohne Seg.	mit Seg.	ohne Seg.	mit Seg.	ohne Seg.
Training	70.82	65.51	64.08	66.73	50.61	48.36
Test	58.44	54.67	41.78	49.11	32.00	27.33

Tabelle 5.4: Einfluss der Segmentierung auf die prozentuale Anzahl der Richtigklassifikationen der Temposchätzung: Das Ergebnis der Analyse *mit* Segmentierung wird durch das repräsentativste Segment bestimmt. Ohne Segmentierung wird das gesamte Audiosignal verarbeitet.

Mikrotime

Zur Evaluierung wurden die für die Temposchätzung beschriebenen Datensätze verwendet, deren Referenzwerte von einem Testhörer ermittelt wurden. Die in Abbildung 5.4 dargestellte Verteilung der Referenzwerte zeigt, dass die überwiegende Zahl an Musikstücken ein binäres Metrum auf Tatumebene aufweist.



(a) Training (b) Test

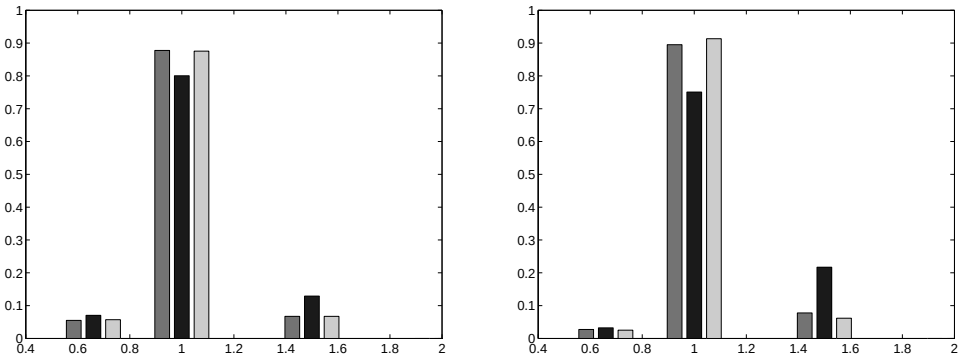
Abbildung 5.4: Verteilung der Referenzwerte (linker Balken) und der Ergebnisse der Low-level-Extraktion (mittlerer Balken) und High-level-Extraktion (rechter Balken) für Trainigsdatensatz (Abbildung a) und Testdatensatz (Abbildung b) in Prozent für die Mikrotimeschätzung.

Die Erkennungsraten der Verfahren zur Mikrotimeschätzung sind in Tabelle 5.2 dargestellt.

Die Verteilungen der Verhältnisse aus Schätzung und Referenz für Trainings- und Test-

	Low-level	High-level	Voting
Training	87.76	69.59	87.55
Test	87.11	56.89	88.89

Tabelle 5.5: Erkennungsraten der Mikrotimeschätzung für Trainings- und Testdatensatz in Prozent.



(a) Training (b) Test

Abbildung 5.5: Histogramm der Verhältnisse aus Schätzung und Referenz der Mikrotimeschätzung für Trainingsdatensatz (Abbildung a) und Testdatensatz (Abbildung b) der Low-level-Schätzung (linker Balken), High-level-Schätzung (mittlerer Balken) und des Gesamtergebnisses (rechter Balken).

datensatz sind in Abbildung 5.5 dargestellt. Eine Steigerung der Erkennungsrate durch die Auswertung zweier Verfahren kann nur unter der Voraussetzung erzielt werden, dass eine signifikante Anzahl der vom besseren Verfahren falschklassifizierten Objekte vom anderen Verfahren richtig klassifiziert werden und ein hochwertiges Konfidenzmaß verwendet wird. Durch die High-level-Analyse wurde bei der Verarbeitung der Trainingsdaten die Mikrotime von nur 3.2% der von der Low-level-Analyse falschklassifizierten Teststücke richtig geschätzt. Die Auswertung beider Verfahren ergab hier keine Verbesserung der Erkennungsleistung.

Taktart

Die Ermittlung der Taktart ist auf die Klassen $\frac{3}{4}$-, $\frac{4}{4}$-, $\frac{5}{4}$-, $\frac{7}{4}$- und $\frac{9}{4}$-Takt beschränkt. Auf Grund der Ähnlichkeit der Merkmalsausprägung wird nicht zwischen $\frac{3}{4}$- und $\frac{6}{8}$-Takt, $\frac{2}{4}$- und $\frac{4}{4}$-Takt sowie $\frac{n}{4}$- und $\frac{n}{8}$-Takt differenziert. Die Erkennungsraten der Taktartschätzung sind in Tabelle 5.6 dargestellt.

	Low-level	High-level	Voting
Training	88.16	85.92	89.80
Test	88.89	72.44	92.22

Tabelle 5.6: Erkennungsraten der Taktartschätzung für Trainings- und Testdatensatz in Prozent.

Die Ergebnisse der Schätzung der Taktart sind für den Testdatensatz anhand der Verwechslungsmatrix in Tabelle 5.7 illustriert. Es wird deutlich, dass die automatisierte Schätzung der Taktart unzuverlässiger arbeitet als die einfache Annahme, dass jedes Lied einen $\frac{4}{4}$-Takt hat.

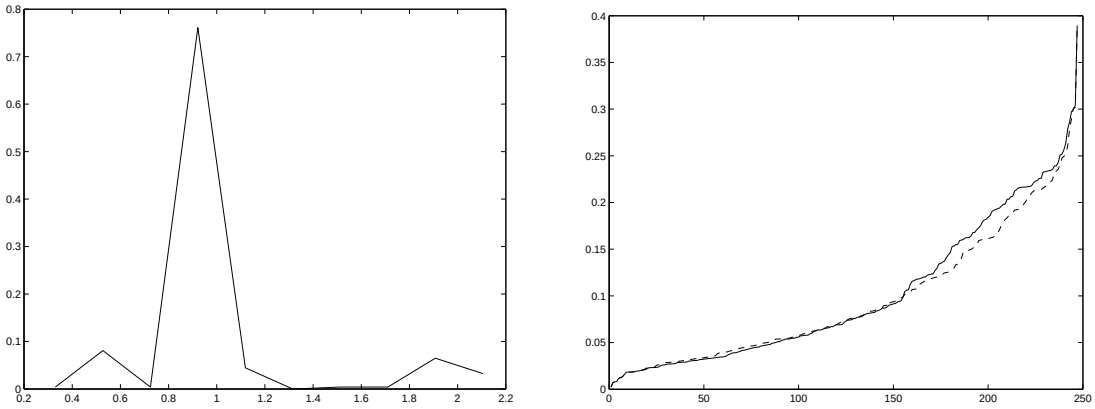
	$\frac{3}{4}$-Takt	$\frac{4}{4}$-Takt	$\frac{5}{4}$-Takt	$\frac{7}{4}$-Takt	$\frac{9}{4}$-Takt
$\frac{3}{4}$-Takt	0.44	1.33	0	0	0
$\frac{4}{4}$-Takt	5.33	91.78	0.44	0.22	0
$\frac{5}{4}$-Takt	0	0.22	0	0	0
$\frac{7}{4}$-Takt	0	0	0	0	0
$\frac{9}{4}$-Takt	0	0.22	0	0	0

Tabelle 5.7: Verwechslungsmatrix der Taktartschätzung des Testdatensatzes. Spaltenweise sind die geschätzten Klassen und zeilenweise die Referenzwerte angegeben.

Tatum- und Beattracking

Zur Evaluation wurden 251 Teststücke ausgewählt und die Referenzwerte ermittelt, indem die Ergebnisse eines Extraktionsverfahrens manuell korrigiert wurden. Die Auswahl berücksichtigt Musikstücke von unterschiedlichem Schwierigkeitsgrad, so dass starke perkussive Instrumentierung neben klassischer Musik analysiert wurde. Durch eine zu geringe Anzahl von detektierten NEZ konnte für vier Teststücke kein Tatumraster detektiert werden.

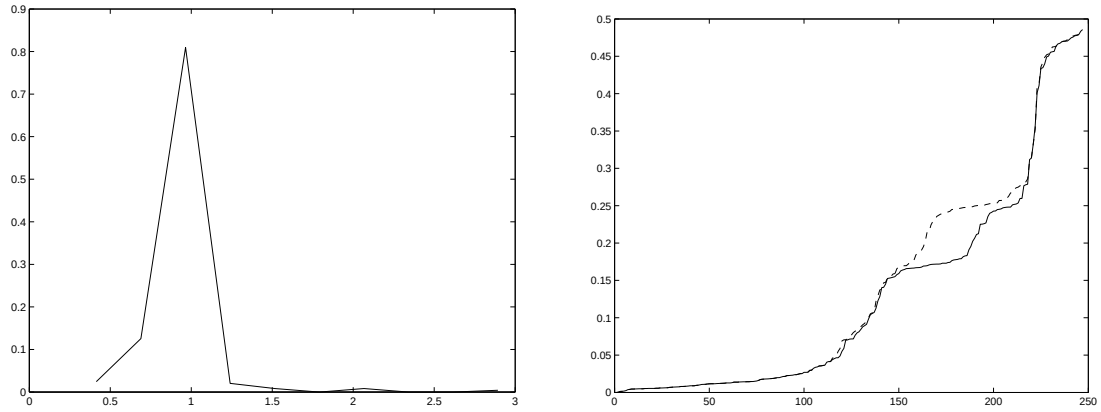
Das Histogramm der Quotienten aus Referenz und Schätzung der Tatumperiode und die mittlere relative Abweichung zwischen Referenzraster und geschätztem Raster und zwischen geschätztem und referenziertem Raster ist in Abbildung 5.6 dargestellt.



(a) (b)

Abbildung 5.6: Evaluierung des Tatumtracking: (a) Histogramm der Quotienten aus Referenz und Schätzung der Tatumperiode, (b) mittlere relative Abweichung zwischen Referenzraster und geschätztem Raster (Linie) und zwischen geschätztem und referenziertem Raster (Strich).

Abbildung 5.7 stellt ein Histogramm der Quotienten aus Referenz und Schätzung der Beatperiode und die mittlere relative Abweichung zwischen Referenzraster und geschätzem Raster und zwischen geschätztem und referenziertem Raster dar.



(a) (b)

Abbildung 5.7: Evaluierung des Beattracking: (a) Histogramm der Quotienten aus Referenz und Schätzung der Beatperiode, (b) mittlere relative Abweichung zwischen Referenzraster und geschätztem Raster (Linie) und zwischen geschätztem und referenziertem Raster (Strich).

5.3 Drumpattern

Zur Evaluierung der automatischen Extraktion von Drumpattern wurden neun Testhörern 92 Ausschnitte aus 40 Musikstücken mit einer Länge von mindestens sechs Sekunden und anschließend die aus den Teststücken automatisch extrahierten Drumpattern präsentiert. Die Teststücke entstammten verschiedenen musikalischen Genres, zum Beispiel Rock, Pop, Latin und Soul, und waren vorselektiert, so dass bei der Analyse der Teststücke keine maßgeblichen Fehler in der Instrumentenerkennung und Tatumschätzung auftraten.

Die Testhörer bewerteten die subjektiv empfundene Qualität der extrahierten Drumpattern anhand einer Skala von eins bis fünf, korrespondierend zu einer schlechten beziehungsweise perfekten Repräsentation. Zusätzlich wurden die Zuhörer darauf hingewiesen, die Bewertung unabhängig von Lautstärkeinformationen vorzunehmen. Die Bewertungen der Testhörer sind in Abbildung 5.8 dargestellt.

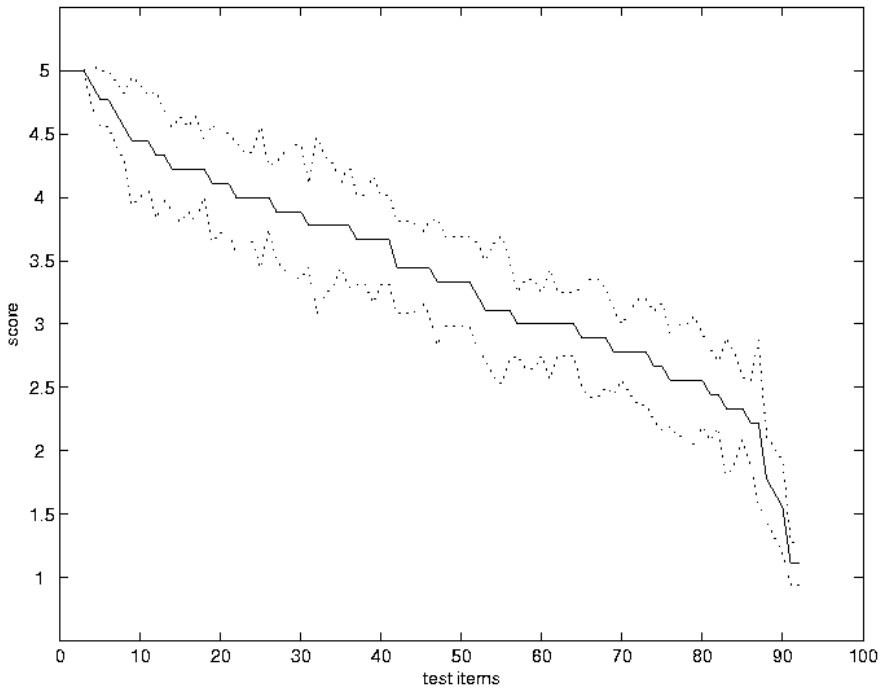


Abbildung 5.8: Evaluierung der automatischen Extraktion von Drumpattern mit neun Testhörern aus [GUDC04]: Für jedes Teststück sind Mittelwert und Standardabweichung der Bewertung der Zuhörer größensortiert dargesellt.

5.4 Eingängigkeit und Intensität

Eingängigkeit

Zur Ermittlung einer Referenz zur Evaluierung des Eingängigkeitsmaßes wurde ein Hörtest durchgeführt. Aufgrund der Unschärfe des Eingängigkeitsbegriffes wurden die VPn aufgefordert, die „rhythmische Geradheit“ von 100 Ausschnitte aus Musikstücken mit einer Dauer von 10 s zu beurteilen. Diese wird als umgekehrt proportional zur Komplexität angenommen [Tou02].

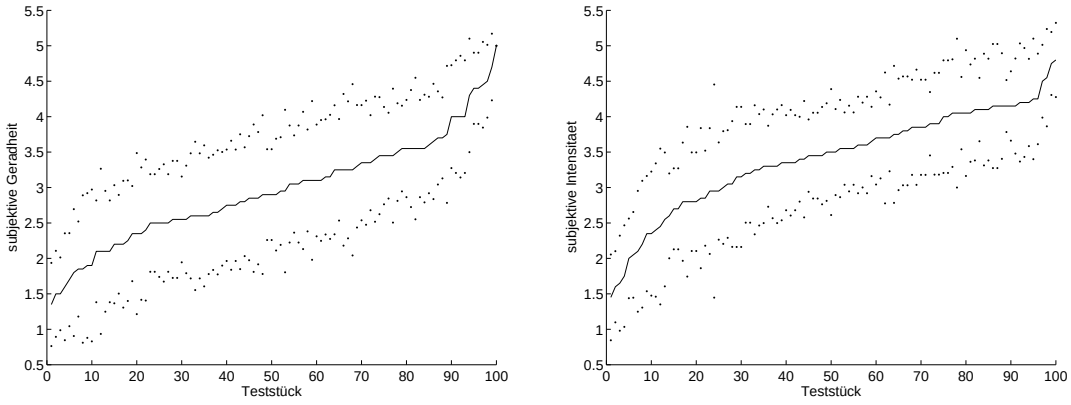
Die den zwanzig VPn gestellte Aufgabe lautete:

Der Rhythmus ist nach seiner Geradheit zu bewerten. Als gerader Rhythmus ist beispielsweise das Metronom zu bewerten, bei dem jede Zählzeit einen Klang auslöst. Ein Rhythmus wird ungerader, je synkopierter, komplizierter er wird. Die Anzahl unterschiedlicher Betonungen nimmt zu und es treten Akzente zwischen den Hauptzählzeiten auf. Zur Bewertung stehen 5 Stufen zur Wahl, 1 - ungerade bis 5 - gerade.

Eine Übersicht über die beteiligten musikalischen Genre enthält Tabelle 5.8. Die Ergebnisse der Befragung ist für die einzelnen Teststücke in Abbildung 5.9 dargestellt.

Genre	Anzahl der Teststücke	Genre	Anzahl der Teststücke
Elektronische Musik	25	Hip-Hop	7
Rock	17	Jazz & Latin	14
Klassik	9	Pop	28

Tabelle 5.8: Genreübersicht des Hörtests zur Bestimmung der rhythmischen Intensität.

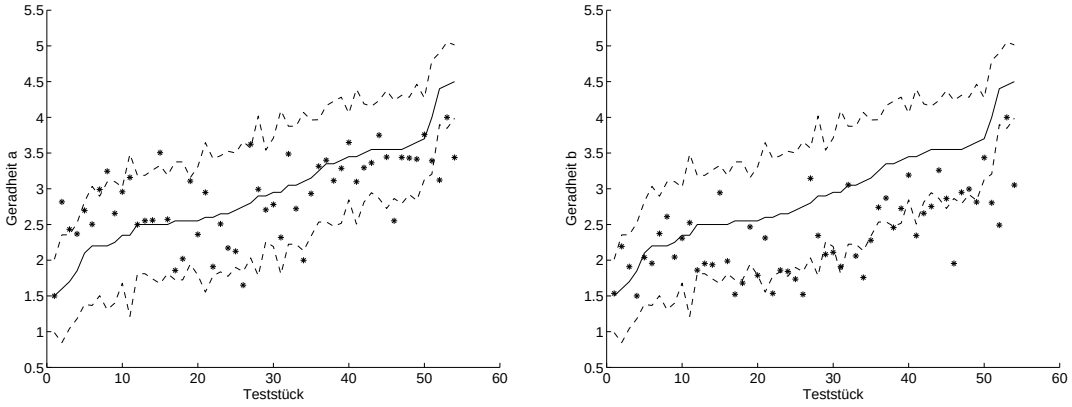


(a) (b)

Abbildung 5.9: Ergebnisse des Hörtests zur Ermittlung von Eingängigkeit und Intensität: Mittelwert und Umgebung der Standardabweichung der Testhörerurteile zur Befragung bezüglich a) Eingängigkeit und b) Intensität.

Hier, wie auch im folgenden Abschnitt beschriebenen Versuch (Bestimmung der Intensität), wird deutlich, dass die Aussagen der Testhörer eine kleinere Varianz bei extremen Bewertungen aufweist.

Zur Evaluierung wurden die Teststücke ausgewählt, die die Kriterien bezüglich der metrischen Struktur erfüllen (siehe Seite 80). Die Berechnung wird weiterhin auf die Musikstücke beschränkt in denen mindestens drei perkussive Instrumente detektiert wurden, da hier eine genügende Relevanz der perkussiven Instrumente angenommen wird.



(a) (b)

Abbildung 5.10: Ergebnisse der automatisierten Ermittlung von Eingängigkeit: additive (links) und muliplikative (rechts) Verknüpfung der Teilmaße.

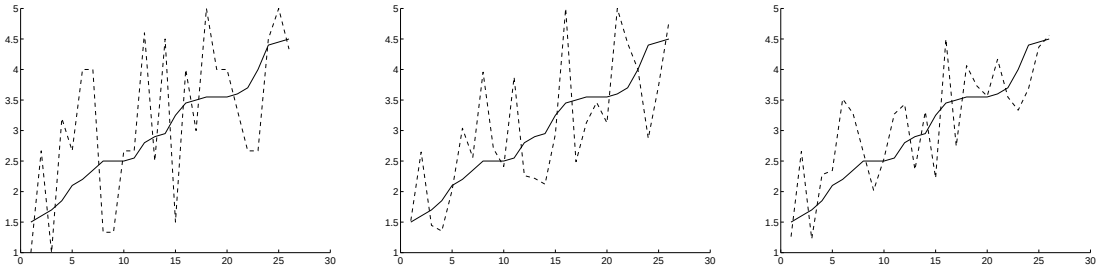
Der geringe Umfang des Datensatzes und die Unzuverlässigkeit der Ausgangsdaten durch Fehler in der Instrumentendetektion, beim Tatumtracking, der Ermittlung der Patternhistogramme und der metrischen Positionen lassen nur eingeschränkt Schlussfolgerungen über die Gültigkeit des präsentierten Ansatzes zu und sind keine Grundlage für eine Ermittlung einer optimalen Verknüpfung der Teilmaße. Eine weitere methodische Beschränkung liegt in der Tatsache begründet, dass der musikalische Rhythmus nicht allein durch die hier ausgewerteten Klänge geprägt ist.

Die Korrelationskoeffizienten zwischen den Mittelwerten der Zuhörerbewertungen und den Eingängigkeitsmaßen E_m beziehungsweise E_p betragen 0.55 beziehungsweise 0.5, die Mittelwertbildung beider Teilmaße ergab einen Korrelationskoeffizienten von 0.71. Bei ausschließlicher Berücksichtigung der Testdaten, bei denen ein gewisser Grad an Übereinstimmung zwischen den Testhörern auftrat, betrugen die Korrelationskoeffizienten 0.56, 0.66 beziehungsweise 0.76. Die Eingängigkeitsmaße sind den Zuhörerbewertungen in Abbildung 5.11 gegenübergestellt.

Intensität

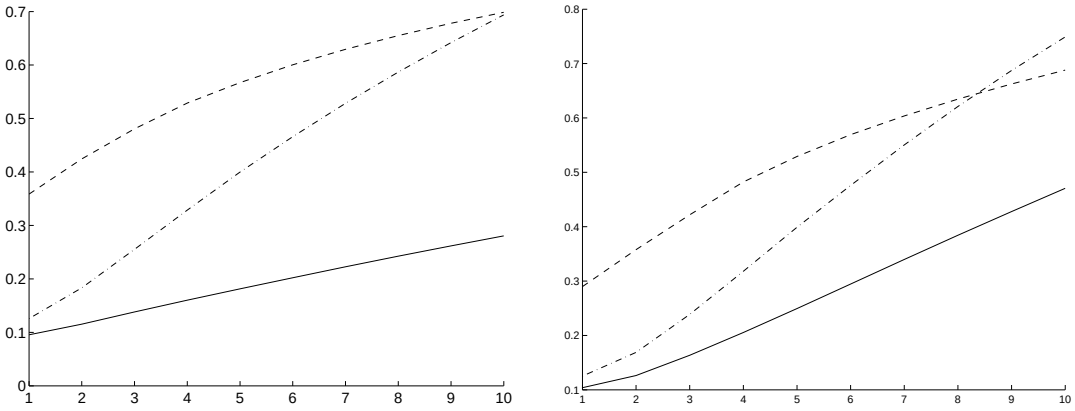
Zur Evaluierung des Intensitätsmaßes wurden zunächst synthetische Testdaten aus drei Mehrspuraufnahmen erstellt, bei denen die Schlagzeugspur mit zehn verschiedenen Verstärkungen gemischt wurde.

Die mittlere Korrelation aus Verstärkung und Mittelwert beziehungsweise Standardabweichung des Akzentsignals beträgt 0.993 beziehungsweise 0.933. Mittelwert und Standardabweichung des Akzentsignals sind in Abhängigkeit vom Verstärkungsfaktor in Abbildung 5.12 dargestellt.



(a) (b) (c)

Abbildung 5.11: Evaluierung des Eingängigkeitsmaßes: metrische Eingängigkeit E_m (links), aus der Patternähnlichkeit ermitteltes Eingängigkeitsmaß E_p (mitte) und Mittelwert beider (rechts).



(a) (b)

Abbildung 5.12: Intensitätsberechnung aus dem Akzentsignal: Mittelwert (links) und Standardabweichung (rechts) des Akzentsignals in Abhängigkeit vom Verstärkungsfaktor der Schlagzeugspuren für drei synthetische Teststücke.

Weiterhin wurden Referenzdaten in einem Hörtest mit 20 VPn durchgeführt, denen 100 Ausschnitte aus Musikstücken mit einer Dauer von 10 s mit folgender Vorgabe präsentiert wurden:

Die rhythmische Intensität soll bewertet werden. Balladen haben eine eher schwache rhythmische Intensität, während dagegen Rockstücke eine starke rhythmische Intensität haben. Musikstücke mit wenigen Perkussionsinstrumenten haben ebenso eine geringe Intensität im Vergleich zu perkussionsreichen Stücken. Zur Bewertung stehen 5 Stufen zur Wahl, 1 - schwach bis 5 - stark.

Die Korrelationskoeffizienten zwischen Hörerbewertung und Mittelwert beziehungsweise Standardabweichung des Akzentsignal betragen 0.66 beziehungsweise 0.69. Abbildung 5.13 stellt die Ergebnisse der Hörerbewertung und der Extraktion aus dem Akzentsignal gegenüber.

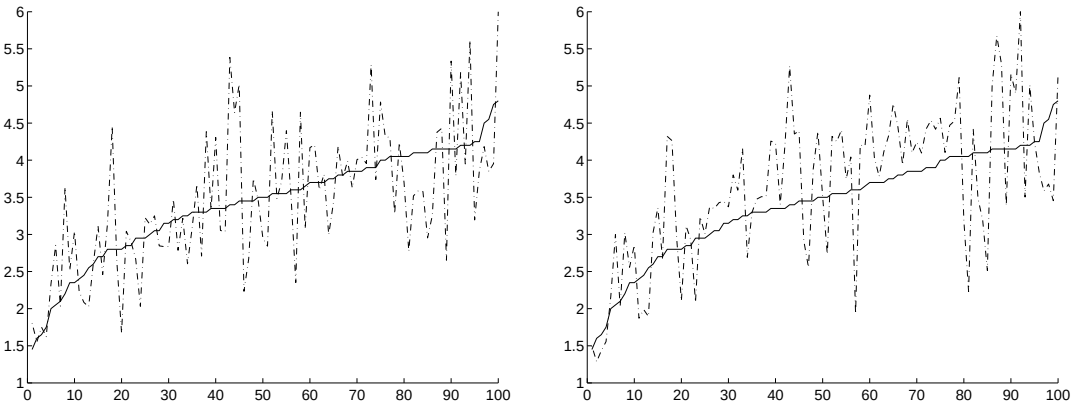


Abbildung 5.13: Intensität aus dem Akzentsignal: Mittelwert (links) und Standardabweichung (rechts) des Akzentsignals (Strich-Punkt-Linie) und Hörerbewertung der Intensität (Linie) größensortiert nach der Hörerbewertung für 100 Teststücke.

5.5 Genreerkennung

Zur Evaluation des vorgeschlagenen Ansatzes zur Genreerkennung wurde ein Testdatensatz verwendet, der für jedes der elf Genres fünfzehn Musikstücke beinhaltet. In der Testphase wurde eine Erkennungsrate von 67.5% erzielt. Detaillierte Ergebnisse sind in der Verwechslungsmatrix (englisch *confusion matrix*) 5.9 angegeben.

Die Verwechslungsmatrix verdeutlicht die Tatsache, dass verschiedene Genres ähnliche Merkmalsausprägungen besitzen, beispielsweise Walzer und Jazz/Swing oder Disco/House und Techno. Die Ursachen für Fehlklassifikationen liegen jedoch häufig in der Extraktion der Merkmale Tempo, Mikrotime, Taktart und der Instrumentenklassifikation sowie in den im Abschnitt 4.6 dargestellten Beschränkungen des Verfahrens. Eine Erweiterung der Taxonomie der Genres und das Hinzufügen einer Rückweisungsklasse sind jedoch Voraussetzungen für die Anwendung des dargestellten Ansatzes in MIR-Anwendungen.

5.6 Vergleich mit anderen Verfahren

Die Aussagekraft von Erkennungsraten für den Vergleich mit anderen Ansätzen und Verfahren ist begrenzt, wenn die verwendeten Datensätze und Bewertungsmethoden verschieden sind. Dieser Abschnitt dient dem Vergleich mit anderen Verfahren zur Temposchätzung und zum Beattracking.

Implementierungen des hier beschriebenen Ansatzes zur Temposchätzung wurden zu den von der MIR-Gemeinschaft organisierten offenen Vergleichen *ADC 2004* [GAD+06] und *MIREX 2005* [Uhl05] eingereicht.

Am *ADC 2004* nahmen sieben Forscherteams mit insgesamt zwölf Implementierungen teil. Die Teilnehmer waren aufgefordert, Implementierungen von Algorithmen einzureichen, die einen Schätzwert des Tempos ausgeben. Der Datensatz setzte sich aus 3199 Musikstücken unteschiedlicher Genres mit Längen zwischen 2 und 30 Sekunden zusammen. Die Referenzwert des Tempos lagen in einem Bereich zwischen 24 und 242 bpm. Der Beitrag des Autors beschränkte sich auf eine Implemetierung der Low-Level-Analyse zur Schätzung der metrischen Merkmale, wie sie in Abschnitt 4.2.4 beschrieben ist.

Die Ergebnisse sind in Tabelle D.1 angegeben und zeigen zwei Erkennungsraten. Die erste Erkennungsrate wurde berechnet aus der Anzahl der Stücke mit richtig geschätztem Tempo. Die zweite Erkennungsrate bewertet Schätzungen mit Tempooktavfehlern als richtige Ergebnisse. Sie zeigt, für wieviele Stücke das geschätzte Tempo dem Referenzwert, oder dem Doppelten, Halben, Dreifachen oder einem Drittel des Referenzwertes entsprechen. Abweichungen von bis zu 4 % des Referenzwertes wurden bei der Bewertung toleriert.

Der Beitrag des Autors erreichte die zweitbeste Erkennungsrate für die Temposchätzung ohne Tolerierung von Tempooktavfehlern (EOT) und die viertbeste Erkennungsrate für die Temposchätzung mit Tolerierung von Tempooktavfehlern (EMT). Der Unterschied im Vergleich mit anderen Verfahren nach EOT und EMT kann mit der im Folgenden beschriebenen Hypothese interpretiert werden.

Die in dieser Arbeit entwickelte Methode zur Temposchätzung baut auf einer Abfolge von Analyseschritten auf, wobei die Tatumschätzung und die Bestimmung des Periodizitätenprofils die Grundlagen der Temposchätzung bilden und einen wesentlichen Einfluß auf das Ergebnis haben. Ein Fehler, der in diesen Analyseschritten auftritt, planzt sich bis zum Endergebnis fort und wird häufig zu einer fehlerhaften Temposchätzung führen. Die dann auftretenden Abweichungen vom Referenztempo wirken sich in der Regel nicht als Tempooktavfehler aus, sondern führen zu Ergebnissen, die in keinem ganzzahligen Verhältnis zum Referenzwert stehen. Dieser Umstand wurde in der weiteren Entwicklung der Temposchätung berücksichtigt und führte zur zusätzlichen Auswertung der Temposchätzung auf Grundlage von Beathistogrammen, wie in den Abschnitten 4.2.5 und 4.4 beschrieben ist. Es ist weiterhin anzumerken, dass Fehler in der Tatumschätzung auch zu Ergebnissen mit richtigem Tempo und falscher Mikrotime führen können.

Am *MIREX 2005* nahmen neun Forscher beziehungsweise Teams mit insgesamt 13 Beiträgen teil. Die Teilnehmer waren aufgefordert, Implementierungen von Algorithmen einzureichen, die die zwei wahrscheinlichsten Hypothesen des Tempos und der zeitlichen Position des ersten Beats (Beatphase) ermitteln. Der Datensatz bestand aus 140 Musikstücken. Zur Bewertung der Verfahren wurde anhand der Ergebnisse der Schätzung des Tempos und der Beatphase ein Bewertungsmaß ermittelt. Die Ergebnisse sind in Tabelle D.2 dargestellt.

Zu diesem Vergleich wurden zwei Implementierungen [Uhl05] vom Autor eingereicht, die das zweit- und drittbeste Ergebnis gemessen am Bewertungsmaß erreichten. Die erste Implementierung basiert auf dem in Abschnitt 4.2.4 beschriebenen Vorgehen und entspricht damit prinzipiell dem zum *ADC 2004* eingereichten Verfahren mit der Erweiterung um die in den Abschnitten 4.2.5 und 4.4 beschriebene Auswertung des Beathistogramms.

Die zweite Implementierung ergänzt das in der ersten Implementierung angewandte Vorgehen um die High-Level-Analyse (siehe Abschnitt 4.3.5). Auf Grundlage der automatisiert extrahierten Instrumenteninformationen wurde die Periodizitätenfunktion bestimmt und zur Temposchätzung verwendet. Die ermittelten Ergebnisse flossen in Abhängigkeit des Konfidenzmasses in das Endergebnis ein.

Die zweite Implementierung führte zu keiner Verbesserung des Ergebnisses im Vergleich zur ersten. Diese Ergebnisse weichen von der in dieser Arbeit durchgeführten Evaluierung ab, die zeigt, daß die zusätzliche Auswertung der High-Level-Analyse zu Verbesserungen der Erkennungsrate führt. Die Ursachen für diesen Unterschied können durch unterschiedliche Testdatensätze bedingt sein, da die hier beschriebene High-Level-Analyse vorraussetzt, dass in der Instrumentierung der zu analysierenden Musikstücke perkussive Instrumente eine Rolle spielen. Fehler in der eingereichten Implementierung der Konfidenzmaße und deren Auswertung können jedoch ebenso dazu geführt, dass die High-Level-Analyse keinen Einfluß auf das Ergebnis hatte. Fehler in der Implementierung der High-Level-Analyse allein würden dagegen zu einer Verschlechterung der Ergebnisse der zweiten Implementierung im Vergleich zur ersten führen, so dass diese Vermutung ausgeschlossen werden kann.

	DH	HR	SF	DB	JS	RP	HM	LA	WA	PP	TE
DH	66					7	7				20
HR		66	27					7			
SF		33	53			13					
DB				87			13				
JS					66		13		20		
RP	20	13	7			60					
HM				13		13	73				
LA	20		7					73			
WA					33				66		
PP				7		7	13			73	
TE	33					7					60

Tabelle 5.9: Verwechslungsmatrix der Genreerkennung in Prozent. Spaltenweise sind die geschätzten Klassen und zeilenweise die Referenzwerte angegeben. Die Abkürzungen der Genres entsprechen denen in Tabelle 4.5.

6 Zusammenfassung

In dieser Arbeit wurden verschiedene Verfahren zur rhythmischen Analyse eines Musiksignals entwickelt und mit bekannten Verfahren zu einem Komplex zusammengefasst mit dem Ziel, einen Beitrag zur Entwicklung von leistungsfähigen MIR-Systemen zu leisten.

Zu den bekannten Verfahren gehören die Detektion von Noten, die Segmentierung des Musiksignals in charakteristische Abschnitte, die Ermittlung metrischer Merkmale anhand von signalnahen Merkmalen. Der eigene Beitrag liegt in der Entwicklung eines Verfahrens zur metrischen Analyse unter Berücksichtigung mehrerer hierarchischer Ebenen. Dieses Verfahren beinhaltet die Darstellung der im Musiksignal auftretenden Periodizitäten als ganzzahlige Vielfache der Tatumperiode, deren musiktheoretische Motivation im Kapitel 2 dargelegt wurde. Das beschriebene Verfahren wird zur Auswertung von signalnahen Merkmalen und von automatisiert extrahierten Instrumenteninformationen eingesetzt.

Die in anderen Gebieten der maschinellen Analyse verbreitete Segmentierung des Signals in charakteristische Abschnitte (*regions of interest*) wurde als Vorverarbeitungsschritt zur Musikanalyse eingeführt. Die gesuchten Merkmale werden aus den einzelnen Segmenten extrahiert, und fließen in Abhängigkeit von der Repräsentativität der Segmente in das Gesamtergebnis ein.

Weiterhin wurde die Extraktion charakteristischer Drumpattern aus automatisiert detektierten Instrumenten vorgestellt. Diese Merkmale sind für einen bedeutenden Teil populärer Musik gültig und bieten eine kompakte und repräsentative Darstellung zur Beschreibung der rhythmischen Eigenschaften. Am Beispiel der Genreklassifikation wurde schließlich die Eignung dieser Merkmale für MIR-Anwendungen gezeigt.

Ansätze zur Ermittlung der empfundenen Intensität und Eingängigkeit von Musik aus signalnahen und symbolischen Informationen wurden vorgestellt, um intuitive Suchanfragen ohne das Wissen um musikalische Begriffe zu ermöglichen. Besonders die Evaluierung dieser Methoden zeigte die Schwierigkeit, die mit der Extraktion von Merkmalen verbunden ist, für die keine exakte Definition vorliegt.

Ein wichtiger Aspekt bei der Analyse und deren Evaluierung ist die Ermittlung der Referenzwerte. Es wird deutlich, dass die Regeln zur Ermittlung der Referenzwerte für metrische Eigenschaften wie zum Beispiel Tempo und Taktart häufig verschiedene Möglichkeiten erlauben, so dass die von verschiedenen Testhörern ermittelten Referenzwerte unterschiedlich ausfallen können. Anhand eines Vergleiches von Referenzwerten wurden systematische Unterschiede für die Ermittlung von Tempowerten zwischen

verschiedenen Testhörern gezeigt. Die Problematik der Mehrdeutigkeit der Definitionen rhythmischer Merkmale ist weiterhin mit Literaturhinweisen belegt.

Die Verfahren wurden anhand einer umfangreichen Datenbasis mit realen Musikstücken verschiedenster Genres getestet. Die Auswertung der Erkennungsleistung lässt die folgenden Schlussfolgerungen zu:

- Die Ermittlung des musikalischen Tempos als primären Merkmals von Rhythmus erfordert die Analyse der gesamten metrischen Struktur des Musiksignals. Die Auswahl der zum Tempo korrespondierenden metrischen Ebene ist nicht immer eindeutig.
- Musikalisches Wissen kann verstärkt in die Entscheidungsfindung einbezogen werden, wenn Instrumenteninformationen ausgewertet werden können. Dadurch können Fehler, die bei der Auswertung signalnaher Merkmale auftreten, vermieden werden. Andererseits wirken sich die Fehler nachteilig aus, die bei der Extraktion der der High-level-Analyse zugrundeliegenden Informationen auftreten. Die Auswertung der Konfidenzmaße der Ergebnisse ermöglicht eine Steigerung der Erkennungsleistung.
- Die Auswertung der Analyse mehrerer Schätzverfahren kann die Erkennungsleistung der automatisierten Verfahren verbessern. Wie die Auswertung der Tatumschätzung zeigt, gilt diese Aussage auch, wenn die Verfahren ähnliche Informationen auswerten.
- Die Segmentierung des zu analysierenden Signals in charakteristische Abschnitte auf Grundlage von signalnahen Merkmalen als Vorverarbeitungsschritt verbessert die Erkennungsleistung.
- Die Erkennungsleistung eines geübten Zuhörers wird mit maschinellen Verfahren noch nicht erreicht.

Die erfolgreiche Teilnahme der entwickelten Verfahren am *ADC 2004* und *MIREX 2005* in der Disziplin *Audio Tempo Extraction* zeigen Leistungsfähigkeit der entwickelten Verfahren zur Temposchätzung im internationalen Vergleich. Möglichkeiten zur Verbesserung der Verfahren werden im folgenden Kapitel aufgezeigt.

7 Ausblick

Die in der vorliegenden Arbeit präsentierten Ergebnisse zeigen, dass die Analyse von Musik bezüglich der rhythmischen Eigenschaften kein gelöstes Problem darstellt und die Verarbeitung verschiedener Musikstücke Verbesserungen und Erweiterungen der Verfahren erfordern. Neben der Arbeit an den Extraktionsalgorithmen stellen auch die Verbesserung der Beschreibung der Problemstellung und der Definitionen der zu extrahierenden Merkmale eine weitere Voraussetzung für den Erfolg der Arbeit an MIR-Systemen dar. Dieses Kapitel zeigt aus Sicht des Autors mögliche Ansätze zur Steigerung der Erkennungsleistung auf.

Extraktion der Akzentsignale Die primäre Voraussetzung für eine erfolgreiche Erkennung der rhythmischen Merkmale ist die robuste Extraktion von Akzenten aus dem Audiosignal. Die in dieser Arbeit verwendeten Akzentsignale spiegeln die Intensität eines empfunden Akzentes nur bedingt wieder und bieten Raum für Verbesserungen. Bedingt durch die Tatsache, dass der beschriebene Ansatz auf der Berechnung von dynamischen Akzenten in Teilbändern basiert und nur bedingt melodische und harmonische Akzente berücksichtigt, werden bedeutende Informationen, die für die Rhythmuswahrnehmung eine große Rolle spielen, nicht ausgewertet. Besonders für die Ausprägung der metrischen Struktur auf höheren Ebenen scheinen die tonalen Eigenschaften von großer Bedeutung zu sein. Dieses Vorgehen führt in der Konsequenz zu der Frage, welcher Akzent beim Auftreten konkurrierender Akzente die Wahrnehmung von Rhythmus leitet.

Ermittlung der metrischen Struktur Die hier vorgeschlagene Ermittlung der metrischen Struktur basiert auf der Ermittlung des Tatums und einer Repräsentation der im Musiksignal auftretenden Periodizitäten als ganzzahlige Vielfache der Tatumperiode. In Musikstücken mit wechselnder Mikrotime und einem wenig ausgeprägten Tatumraster versagt dieser Ansatz häufig und wird durch die Auswertung des Beathistogramms ersetzt. Da die hier verwendete Auswertung des Beathistogramms keine Informationen über andere metrische Ebenen liefert, ist eine diesbezügliche Erweiterung anzustreben.

Die Berechnung von Periodizitäten im Akzentsignal ist ein musiktheoretisch begründetes Konzept. Bei Rhythmen mit expressivem Tempo kann ein Zuhörer dem Tempoverlauf mit kleiner Verzögerung folgen. Die hier verwendete AKF wie alle anderen merkmalsbasierten Ansätze auch verarbeitet das Signal dagegen blockweise. Verfahren mit einer *multiple agent architecture* sind für die Verarbeitung von Rhyth-

men mit expressivem Tempo prädestiniert, haben sich jedoch bisher nicht durchgesetzt. Zwei Gründe dafür sind möglicherweise einerseits der Umstand, dass expressives Tempo selten auftritt und andererseits der ereignisbasierte Ansatz dieser Verfahren. Die Entwicklung eines merkmalsbasierten Verfahrens zur Periodizitätenberechnung unter Berücksichtigung von Tempovariationen stellt einen wichtigen Schritt zur Verbesserung automatisierter Verfahren dar.

Auswertung von Instrumenteninformationen Die High-level-Analyse umfasst eine Reihe von Verarbeitungsschritten und damit verschiedene Fehlerursachen. Eine Verbesserung der einzelnen Komponenten der Verarbeitung wird zweifelsfrei positive Auswirkungen auf das Endergebnis haben. Dazu gehören das Tatumtracking und die Quantisierung der NEZ, die Instrumentenerkennung, die Ermittlung der Patternlänge und des Patterns sowie dessen Auswertung.

Die Auswertung mehrkanaliger Audiosignale (z.B. Stereoaufnahmen) kann insbesondere im Hinblick auf die Instrumentenerkennung zu Verbesserungen führen, auch wenn für einen Zuhörer einkanalige Signale oft nicht weniger schwer zu separieren sind. Da die Verfahren der Instrumentenerkennung robuster arbeiten, je größer der Signal-Rausch-Abstand, ist die Auswertung der einzelnen Kanäle einer echten Mehrkanalaufnahme eine einfache Möglichkeit zur Verbesserung der Erkennungsleistung.

Weitere Ansätze zur Verbesserung der Instrumentenerkennung stellen die redundante Analyse und die Entwicklung einer vereinfachten Taxonomie der Instrumentenklassen dar. Eine vereinfachte Taxonomie der Instrumente auf der Basis der Instrumentenklänge im Gegensatz zur Instrumentenkonstruktion ist für verschiedene MIR-Anwendungen vorteilhaft, beispielsweise für *recommandation engines* oder die Ermittlung des musikalischen Genres. Diese Überlegung ist motiviert durch die Tatsachen, dass unterschiedliche Instrumente ähnliche Klänge hervorrufen können, durch die Bearbeitung des Audiosignals die Klangcharakteristik manipuliert werden kann, und durch unterschiedliche Spielweisen mit einem Instrument verschiedene Klänge erzeugt werden können. Für das Hören eines Musikstückes ist der Klang und nicht das verwendete Werkzeug entscheidend.

MIR mit symbolischen Repräsentationen Die in dieser Arbeit beschriebenen Verfahren werten ausschließlich Audiosignale aus. Die Analyse von symbolischen Repräsentationen von Musiksignalen, wie zum Beispiel MIDI-Dateien oder Darstellungen im GUIDO Notation Format [HHRK98], wird in dieser Arbeit nicht betrachtet, da diese Darstellungen nur für relativ wenige Musikstücke, gemessen an der Gesamtheit, zur Verfügung stehen, und so eine geeignete Beschränkung der Aufgabenstellung erreicht wurde. Die Auswertung von symbolischen Repräsentationen ist jedoch vorstellbar für verschiedene Anwendungen und führt zu zusätzlichen Informationen zur Beschreibung des Musiksignals. Da für eine Vielzahl von Audiodateien die Metadaten über den Au-

tor, den Titel und die Veröffentlichung durch *Audio-Fingerprint-Technologien* [AHH+01] oder Datenbanken wie *Gracenote Media Database*[1] ermittelt werden können, ist die Suche nach symbolischen Darstellungen des Musikstückes möglich, aus denen ein Vielzahl von musikalischen Merkmalen extrahiert werden können.

Definition der Begriffe und Aufgabenstellungen Bei der Ermittlung der metrischen Eigenschaften treten Fehler auf, die häufig auf eine falsche Zuordnung des Merkmals Tatum, Beat und Takt zu den ermittelten metrischen Ebenen zurückzuführen ist. Dabei spielen Tempooktavfehler und die Verwechselung der Taktart eine große Rolle. Die mehrdeutigen Definitionen für die rhythmischen Merkmale (siehe Abschnitt 2.3) und die verschiedenen Möglichkeiten der Interpretation der metrischen Struktur und Auswahl des musikalischen Tempos (siehe die Beschreibung der Ergebnisse des Hörversuchs in Abschnitt 5.2) führen zu Fehlklassifikationen und erschweren die Ermittlung einer Ähnlichkeit von Musikstücken auf Grundlage der nur unscharf definierten Begriffe. Die Repräsentation der metrischen Struktur von Musikstücken ohne die Interpretation der metrischen Ebenen (z.B. die Bestimmung, welche metrische Ebene dem Tempo entspricht) ist für Anwendungen wie beispielsweise die Ermittlung der Ähnlichkeit von Musikstücken ausreichend und robuster. Eine solche Darstellung beinhaltet die Tempi und empfundene Intensität der Pulse auf allen metrischen Ebenen.

Abschließend ist festzustellen, dass die Analyse von Musik mittels digitaler Rechentechnik und dem derzeitigen Entwicklungsstand der Verfahren sich nicht in einer Tiefe und mit einer Zuverlässigkeit betreiben lässt, die ein Musikhörer erreichen kann. Wie auch bei Problemstellungen aus der Verarbeitung von Bild- und Sprachsignalen deutlich wird, ist die natürliche der künstlichen Intelligenz in vielerlei Hinsicht überlegen. Jedoch gibt es verschiedenen Aufgaben, für deren Bewältigung automatisierte Verfahren notwendig sind, beispielsweise die Analyse umfangreicher Datenbanken.

Die Globalisierungstendenzen und die Zunahme der Informations- und Kommunikationstechnik führen neben einem Anstieg des Umfangs an Informationen auch zu einer Expansion der angebotenen Musikvielfalt. Die hier dargestellten Methoden sind hauptsächlich für die Analyse von populärer Musik entwickelt, sie sind jedoch in ihrer Anwendung nicht beschränkt. Es ist aber zu berücksichtigen, dass inmitten der gewaltigen musikalischen Vielfalt Musikstücke auftreten, deren Analyse kritisch zu betrachten ist. Gerade die Schaffung illusionärer Eindrücke, die Auslotung perzeptueller Grenzen und das Übertreten von bestehenden Konventionen gehören oft zur Motivation eines Musikschaffenden. Da sich die Entwicklung der Methoden der Musikanalyse an der Art und Weise, wie ein Zuhörer Musik wahrnimmt orientiert, sind natürliche Grenzen durch das Interpretationsvermögen eines erfahrenen Zuhörers gesetzt.

[1] Die bis 2004 als *CDDB* bekannte Datenbank erlaubt die Identifikation eines digitalen Tonträgers anhand der Anzahl und Länge der enthaltenen Musikstücke.

Literaturverzeichnis

[ABDR03] Alonso, Miguel ; Badeau, Roland ; David, Bertrand ; Richard, Gaël: Musical tempo estimation using noise subspace projections. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2003

[AD90] Allen, Paul E. ; Dannenberg, Roger B.: Tracking musical beats in real-time. In: *Proceedings of the International Computer Music Conference, ICMC*, 1990

[ADR03] Alonso, Miguel ; David, Bertrand ; Richard, Gaël: A study of tempo tracking algorithms from polyphonic music signals. In: *Proceedings of the 4th COST 276 Workshop*, 2003

[AHH+01] Allamanche, Eric ; Herre, Jürgen ; Hellmuth, Oliver ; Fröba, Bernhard ; Cremer, Markus: AudioID: Towards Content-Based Identification of Audio Material. In: *Proceedings of the 110th AES Convention*, 2001

[AP03] Abdallah, Samer ; Plumbley, Mark: Unsupervised onset detection: A probabilistic approach using ICA and a hidden Markov classifier. In: *Proceedings of the Music Processing Colloquium*, 2003

[Ben84] Benjamin, W.: A theory of musical meter. In: *Music Perception* 1 (1984), S. 355–413

[BG98] Behne, Klaus-Ernst ; Gembris, Heiner: *Die Generative Theorie der Tonalen Musik*. LIT Verlag, 1998

[Bil93] Bilmes, Jeffrey A.: *Timing is of essence: perceptual and computational techniques for representing, learning and reproducing expressive timing in percussive rhythms*, Massachusetts Institute of Technology, Diplomarbeit, 1993

[BL03] Burred, Juan J. ; Lerch, Alexander: A hierarchical approach to automatic musical genre classification. In: *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx*, 2003

[Bra98] Brandenburg, Karlheinz: Perceptual coding of high quality digital audio. In: *Applications of digital signal processing to audio and acoustics*. Kluwer Academic Publishers, 1998

[Bre90] Bregman, Albert S.: *Auditory scene analysis.* MIT Press, 1990

[Bre98] Bregman, Albert S.: Psychological data and computational ASA. In: *Computational Auditory Scene Analysis.* Lawrence Erlbaum Associates, 1998

[Bro98] Brown, Judith C.: Determination of the meter of musical scores by autocorrelation. In: *Journal of the Acoustical Society of America* 94 (1998), Nr. 4, S. 1953–1957

[BS03] Bello, Juan P. ; Sandler, Mark: Phase-based note onset detection for music signals. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-03*, 2003

[CG91] Cassell, Annette ; Green, Patrick: Wahrnehmung. In: *Einführung in die Kognitionspsychologie.* Reinhardt Verlag, 1991

[CGG+06] Cano, Pedro ; Gomez, Emilia ; Gouyon, Fabien ; Koppenberger, M. ; Ong, B. ; Streich, Sebastian ; Wack, N.: ISMIR 2004 audio description contest. In: *IEEE Transactions on Speech and Audio Processing* (2006)

[Che94] Chernoff, John M.: *Rhythmen der Gemeinschaft.* Trickster Verlag, 1994

[CK99] Carterette, Edward C. ; Kendall, Roger A.: Comparative music perception and cognition. In: *The Psychology of Music.* Academic Press, 1999

[CK02] Cemgil, Ali T. ; Kappen, Bert: Integrating tempo tracking and quantization using particle filtering. In: *Proceedings of the International Computer Music Conference, ICMC*, 2002

[CK03] de Cheveigné, Alain ; Kawahara, Hideki: YIN, a fundamental frequency estimator for speech and music. In: *Journal of the Acoustical Society of America* 111 (2003), Nr. 4, S. 1917–1930

[CKDH01] Cemgil, Ali T. ; Kappen, Bert ; Desain, Peter ; Honing, Henk: On tempo tracking: Tempogram representation and Kalman filtering. In: *Journal of New Music Research* 28 (2001), Nr. 4, S. 259–273

[CM63] Cooper, Grosvenor ; Meyer, Leonard B.: *The rhythmic structure of music.* The University of Chicago Press, 1963

[Com94] Comon, Pierre: Independent Component Analysis, a new concept? In: *Signal Processing, Elsevier* 36 (1994), Nr. 3, S. 287–314

[CW00] Casey, Michael A. ; Westner, Alex: Separation of mixed audio sources by Independent Subspace Analysis. In: *Proceedings of the International Computer Music Conference, ICMC*, 2000

[DBDS03a] Duxbury, Chris ; Bello, Juan P. ; Davis, Mike ; Sandler, Mark: A combined phase and amplitude based approach to onset detection for audio segmentation. In: *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS-03*, 2003

[DBDS03b] Duxbury, Chris ; Bello, Juan P. ; Davis, Mike ; Sandler, Mark: Complex domain onset detection for musical signals. In: *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx*, 2003

[DDS01] Duxbury, Chris ; Davies, Mike ; Sandler, Mark: Separation of transient information in musical audio using multiresolution analysis techniques. In: *Proceedings of the COST G-6 Conference on Digital Audio Effects, DAFx*, 2001

[Deu99] Deutsch, Diana: Grouping mechanism in music. In: *The Psychology of Music*. Academic Press, 1999

[DG02] Dixon, Simon ; Goebl, Werner: Pinpointing the beat: Tapping to expressive performances. In: *Proceedings of the 7th International Conference on Music Perception and Cognition, ICMPC*, 2002

[DGW02] Dixon, Simon ; Goebl, Werner ; Widmer, Gerhard: Real time tracking and visualization of musical expression. In: *Proceedings of the International Conference on Music and Artificial Intelligence, ICMAI*, 2002

[DH89] Desain, Peter ; Honing, Henk: The quantization of musical time: A connectionist approach. In: *Computer Music Journal* 13 (1989), Nr. 3, S. 56–66

[DH94] Desain, Peter ; Honing, Henk: Foot-tapping: A brief introduction. In: *Proceedings of the International Computer Music Conference, ICMC*, 1994

[DH00] Desain, Peter (Hrsg.) ; Honing, Henk (Hrsg.): *Rhythm perception and production*. Swets & Zeitlinger, 2000

[Dix97] Dixon, Simon: Beat induction and rhythm recognition. In: *Proceedings of the Australian Joint Conference on Artificial Intelligence*, 1997

[Dix99] Dixon, Simon: A beat tracking system for audio signals. In: *Proceedings of the Conference on Mathematical and Computational Methods in Music*, 1999

[Dix00] Dixon, Simon: A lightweight multi-agent musical beat tracking system. In: *Proceedings of the Pacific Rim International Conference on Artificial Intelligence, PRICAI*, 2000

[Dix01a] Dixon, Simon: Automatic extraction of tempo and beat from expressive performances. In: *Journal of New Music Research* 30 (2001), Nr. 1

[Dix01b] Dixon, Simon: An empirical comparision of tempo trackers. In: *Proceedings of the 8th Brazilian Symposium on Computer Music*, 2001

[Dix01c] Dixon, Simon: Learning to detect onsets of acoustic piano tones. In: *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, 2001

[DMT01] Daudet, Laurent ; Molla, Stéphane ; Torrésani, Bruno: Transient detection and encoding using wavelet coefficient trees. In: *Proceedings of the GRETSI Symposium on Signal and Image Processing*, 2001

[Dow03a] Downie, J. Stephen: Music information retrieval. In: *Annual Review of Information Science and Technology*. Information Today, 2003

[Dow03b] Downie, J. Stephen: Towards a scientific evaluation of music information retrieval systems. In: *Proceedings of the 4th International Conference on Music Information Retrieval, ISMIR*, 2003

[Dow05] Downie, J. Stephen: The 2005 music information retrieval evaluation exchange (MIREX2005): Preliminary overvie. In: *Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR*, 2005

[DP04] Davies, Matthew E. P. ; Plumbley, Mark D.: Causal tempo tracking of audio. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[DPB00] Drake, Carolyn ; Penel, Amandine ; Bigand, Emmanuel: Why musicians tap slower than nonmusicians. In: *Rhythm perception and production*. Swets and Zeitlinger, 2000

[DPG03] Dixon, Simon ; Pampalk, Elias ; Goebl, Werner: Classification of dance music by periodicity patterns. In: *Proceedings of the 4th International Conference on Music Information Retrieval, ISMIR*, 2003, S. 159–165

[DSD02] Duxbury, Chris ; Sandler, Mark ; Davies, Mike: A hybrid approach to musical note onset detection. In: *Proceedings of the 5th International Conference on Digital Audio Effects, DAFx*, 2002

[DU04] Dittmar, Christian ; Uhle, Christian: Further steps towards drum transcription of polyphonic music. In: *Proceedings of the AES 116th Convention*, 2004

[DW03] Dalinghaus, Klaus ; Weyde, Tillman: Structure recognition on sequences with a neuro-fuzzy system. In: *Proceedings of the International Conference in Fuzzy Logic and Technology*, 2003

[Edl95] Edler, Bernd: *Äquivalenz von Transformation und Teilbandzerlegung in der Quellencodierung*, Universität Hannover, Diss., 1995

[EG95] Eddins, David A. ; Green, David M.: Temporal integration and temporal resolution. In: *Hearing*. Academic Press, 1995

[FC03] Foote, Jonathan T. ; Cooper, M. L.: Media segmentation using self-similar decomposition. In: *Proceedings of SPIE Storage and Retrieval for Multimedia Databases* Bd. 5021, 2003, S. 167–175

[Fit04] FitzGerald, Derry: *Automatic drum transcription and source separation*, Dublin Institute of Technology, Diss., 2004

[FLC03a] FitzGerald, Derry ; Lawlor, Bob ; Coyle, Eugene: Drum transcription using automatic grouping of events and Prior Subspace Analysis. In: *Proceedings of the 4th European Workshop on Image Analysis or Multimedia Interactive Services*, 2003

[FLC03b] FitzGerald, Derry ; Lawlor, Bob ; Coyle, Eugene: Prior Subspace Analysis for drum transcription. In: *Proceedings of the AES 114th Convention*, 2003

[Fle01] Fleissner, G.: Rhythmizität, zirkadiane Rhythmik und Schlaf. In: *Neurowissenschaft*. Springer-Verlag, 2001

[Fle03] Fleischmann, Kristoffer: *Untersuchung zur Detektion von Noteneinsätzen in monophoner und polyphoner Musik*, Technische Universität Ilmenau, Diplomarbeit, 2003

[Foo00] Foote, Jonathan T.: Automatic audio segmentation using a measure of audio novelty. In: *Proceedings of the Internaional Conference on Multimedia and Expo, ICME*, 2000

[Fra82] Fraisse, Paul: Rhythm and tempo. In: *The Psychology of Music*. Academic Press, 1982

[Fri04] Frieler, Klaus: Beat and meter extraction using gaussified onsets. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[FU01] Foote, Jonathan T. ; Uchihashi, S.: The beat spectrum: A new approach to rhythm analysis. In: *Proceedings of the Internaional Conference on Multimedia and Expo, ICME*, 2001

[GAD+06] Gouyon, Fabien ; Alonso, Miguel ; Dixon, Simon ; Klapuri, Anssi ; Tzanetakis, George ; Uhle, Christian: An experimental comparision of audio tempo induction algorithms. In: *IEEE Transactions on Speech and Audio Processing* (2006)

[GD04] Gouyon, Fabien ; Dixon, Simon: Dance music classification: a tempo-based approach. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[GD05] Gouyon, Fabien ; Dixon, Simon: A review of automatic rhythm desciption systems. In: *Computer Music Journal* 29 (2005), Nr. 1

[GDPW04] Gouyon, Fabien ; Dixon, Simon ; Pampalk, Elias ; Widmer, Gerhard: Evaluating rhythmic desriptors for musical genre classification. In: *Proceedings of the AES 25th International Conference*, 2004

[GH01] Gouyon, Fabien ; Herrera, Perfecto: Exploration of techniques for automatic labeling of audio drum tracks' instruments. In: *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, 2001

[GH03a] Gouyon, Fabien ; Herrera, Perfecto: A beat induction method for musical audio signals. In: *Proceedings of the 4th WIAMIS Special Session on Audio Segmentation and Digital Music*, 2003

[GH03b] Gouyon, Fabien ; Herrera, Perfecto: Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors. In: *Proceedings of the AES 114th Convention*, 2003

[GHC02] Gouyon, Fabien ; Herrera, Perfecto ; Cano, Pedro: Pulse-dependent analysis of percussive music. In: *Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002

[GK05] GK, Gesellschaft für Kognitionswissenschaft e.V. *Allgemeine Information.* http://www.gk-ev.de/. 2005

[GM94] Goto, Masataka ; Muraoka, Yoichi: A beat tracking system for acoustic signals of music. In: *ACM Proceedings*, 1994

[GM95] Goto, Masataka ; Muraoka, Yoichi: Music understanding at the beat level - Real-time beat tracking for audio signals. In: *Proceedings of the IJCAI-95 Workshop on Computational Auditory Scene Analysis*, 1995

[GM96] Goto, Masataka ; Muraoka, Yoichi: Beat tracking based on multiple-agent architecture - a real-time beat tracking syste for audio signals. In: *Proceedings of the 2nd International Conference on Multiagent Systems*, 1996

[GM97a] Goto, Masataka ; Muraoka, Yoichi: Issues in evaluating beat tracking systems. In: *Proceedings of the IJCAI-97 Workshop on Issues in AI and Music*, 1997

[GM97b] Goto, Masataka ; Muraoka, Yoichi: Real-time rhythm tracking for drumless audio signals - chord change detection for musical decisions. In: *Proceedings of the IJCAI-97 Workshop on Computational Auditory Scene Analysis*, 1997

[GM02] Gouyona, Fabien ; Meudic, Benoit: Towards rhythmic content processing of musical signals: fostering complementary approaches. In: *Journal of New Music Research* (2002)

[Gol02] Goldstein, E. B.: *Wahrnehmungspsychologie.* 2. dt. Auflage. Spectrum Verlag, 2002

[Gor87] Gordon, J. W.: The perceptual attack time of musical tones. In: *Journal of the Acoustical Society of America* 82 (1987), S. 88–105

[Got98] Goto, Masataka: Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. In: *Speech Communication* 27 (1998), S. 311–335

[Got01] Goto, Masataka: An audio-based real-time beat-tracking system for music with and without drums. In: *Journal of New Music Research* 30 (2001), Nr. 2, S. 159–171

[GPD00] Gouyon, Fabien ; Pachet, François ; Delerue, Olivier: On the use of zero crossing rate for an application of classification of percussive sounds. In: *Proceedings of COST G-6 Conference on Digital Audio Effects, DAFx*, 2000

[Gra59] Grabner, H.: *Allgemeine Musiklehre.* Bärenreiter-Verlag, 1959

[GUDC04] Gruhne, Matthias ; Uhle, Christian ; Dittmar, Christian ; Cremer, Markus: Extraction of drum patterns and their description within the MPEG-7 High-Level-Framework. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[Har75] Hartigan, John A.: *Clustering Algorithms.* Wiley and Sons, 1975

[HDG03] Herrera, Perfecto ; Dehamel, Amaury ; Gouyon, Fabien: Automatic labeling of unpitched percussion sounds. In: *Proceedings of the AES 114th Convention*, 2003

[Hei05] Heinz, Thorsten: *Ein physiologisch gehörrechtes Verfahren zur automatisierten Melodietranskription*, Technische Universität Ilmenau, Diss., 2005

[Hem97] Hempel, C.: *Neue Allgemeine Musiklehre*. Schott Musik International, 1997

[HH99] Hyvärinen, A. ; Hoyer, Patrick: Independent Subspace Analysis shows emergence of phase and shift invariant features from natural images. In: *Proceedings of the International Joint Conference on Neural Networks, IJCNN*, 1999

[HHRK98] Hoos, Holger H. ; Hamel, Keith A. ; Renz, Kai ; Kilian, Jürgen: The GUIDO notation format -a novel approach for adequatley representing score level music. In: *Proceedings of the 1998 International Computer Music Conference, ICMC*, 1998, S. 451–454

[Hin90] Hinich, Melvin J.: Detecting a transient signal by bispectral analysis. In: *IEEE Transactions on Acoustis, Speech and Signal Processing* 38 (1990), Nr. 7, S. 127–1283

[Hir59] Hirsh, I. J.: Auditory perception of temporal order. In: *Journal of the Acoustical Society of America* 31 (1959), S. 759–767

[HM03] Hainsworth, S. ; Macleod, M.: Onset detection in musical audio signals. In: *Proceedings of the International Computer Music Conference*, 2003

[HS61] Hirsh, I. J. ; Sherrick, C. E.: Perceived order of different sense modalities. In: *Journal of Experimental Psychology* 62 (1961), S. 423–432

[HSG04] Herrera, Perfecto ; Sandvold, Vegard ; Gouyon, Fabien: Percussion-related semantic descriptors of music audio files. In: *Proceedings of the AES 25th International Conference on Metadata*, 2004

[HT05] Hannon, Erin E. ; Trehub, Sandra E.: Metrical Categories in Infancy and Adulthood. In: *Psychological Science* 16 (2005), Nr. 1, S. 48–55

[HYG02] Herrera, Perfecto ; Yeterian, Alexandre ; Gouyon, Fabien: Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In: *Proceedings of the 2nd International Conference on Music and Artificial Intelligence, ICMAI*, 2002

[IPO04] Iversen, John R. ; Patel, Aniruddh D. ; Ohgushi, Kengo: Perception of nonlinguistic rhythmic stimuli by American and Japanese listeners. In: *Proceedings of the 18th International Congress on Acoustics*, 2004

[Iye98] Iyer, Vijay S.: *Microstructures of feel, macrostructures of sound: embodied cognition in West-African and African-American musics*, University of California, Diss., 1998

[JA03] Jensen, Kristoffer ; Anderson, Tue H.: Real-time beat estimation using feature extraction. In: *Proceedings of the Computer Music Modelling and Retrieval Symposium*, 2003

[Jou01] Jourdain, Robert: *Das wohltemperierte Gehirn*. Spektrum Akademischer Verlag, 2001

[KBT04] Kapur, Ajay ; Benning, Manjinder ; Tzanetakis, George: Query by beatboxing - Music Information Retrieval for the DJ. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[KCZS03] Keiler, Florian ; CanKaradogan ; Zölzer, Udo ; Schneider, Albrecht: Analysis of transient musical sounds by auto-regressive modelling. In: *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx*, 2003

[KEA04] Klapuri, Anssi P. ; Eronen, Antti J. ; Astola, Jaakko T.: Analysis of the meter of acoustic musical signals. (2004). – to appear

[KEA05] Klapuri, Anssi ; Eronen, Antti ; Astola, Jaakko: Analysis of the meter of acoustic musical signals. In: *IEEE Transcations on Speech and Audio Processing* (2005)

[Kla99] Klapuri, Anssi: Sound onset detection by applying psychoacoustic knowledge. In: *IEEE Interntional Conference on Acoustic, Speech and Signal Processing, ICASSP* (1999)

[Kla03] Klapuri, Anssi. *Musical meter estimation and music transcription*. presented at the Cambridge Music Colloquium, Cambridge University, UK. 2003

[Kla04] Klapuri, Anssi: *Signal processing methods for the automatic transcription of music*, Tampere University of Technology, Diss., 2004

[KP04] Kapanci, Emir ; Pfeffer, Avi: A hierarchical approach to onset detection. In: *Proceedings of the International Computer Music Conference, ICMC*, 2004

[Kru00] Krumhansl, Carol L.: Rhythm and pitch in music perception. In: *Psychological Bulletin* 126 (2000), Nr. 1, S. 159–179

[Lar01] Laroche, Jean: Estimating tempo, swing and beat locations in audio recordings. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2001

[Lar03] Laroche, Jean: Efficient tempo and beat tracking in audio recordings. In: *Journal of the Audio Engineering Society* 51 (2003), Nr. 4, S. 226–233

[LDR04] Leveau, Pierre ; Daudet, Laurent ; Richard, Gael: Methology and tools for the evaluation of automatic onset detection algorithms in music. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[Lem95] Leman, Marc: *Music and schema theory.* Springer Verlag, 1995

[LiL82] Longuet-iggins, H. C. ; Lee, C. S.: The perception of musical rhythms. In: *Perception* 11 (1982), S. 115–128

[LJ83] Lerdahl, Fred ; Jackendoff, Ray: *A generative theory of tonal music.* MIT Press, 1983

[LK94] Large, Edward W. ; Kolen, John F.: Resonance and the perception of musical meter. In: *Connection Science* 6 (1994), Nr. 1, S. 177–208

[Log00] Logan, Beth: Mel frequency cepstral coefficients for music modelling. In: *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, 2000

[LZ76] Lempel, Abraham ; Ziv, Jacob: On the complexity of finite sequences. In: *IEEE Transactions on Information Theory* 22 (1976), Nr. 1, S. 75–81

[LZ77] Lempel, Abraham ; Ziv, Jacob: A univeral algorithm for sequential data compression. In: *IEEE Transactions on Information Theory* 23 (1977), Nr. 3, S. 337–343

[Mar91] Marron, Eddy: *Rhythmiklehre.* AMA-Verlag, 1991

[MB94] Maher, Robert C. ; Beauchamp, James W.: Fundamental frequency estimation of musical signals using a two-way mismatch error procedure. In: *Journal of the Acoustical Society of America* 95 (1994), Nr. 4, S. 2254–2263

[MB96] Masri, Paul ; Bateman, Andrew: Improved modelling of attack transients in music analysis-resynthesis. In: *Proceedings of the International Computer Music Conference, ICMC*, 1996

[McD98a] McDonald, SKot: Biologicalesque transcription of percussion. In: *Proceedings of the Australian Computer Music Conference*, 1998

[McD98b] McDonald, SKot. *Telling guitars to fuzz off: Adaptive Filtering to improve event detektion in the presence of irregula broad band AM.* Departmental Conference. 1998

[Meu02a] Meudic, Benoit: Automatic meter extraction from MIDI files, 2002

[Meu02b] Meudic, Benoit: A causal algorithm for beat-tracking, 2002

[Meu03] Meudic, Benoit: Musical pattern extraction: From repetition to musical structure. In: *Proceedings of the International Workshop Computer Music Modeling and Retrieval*, 2003

[MIR05] MIREX. *MIREX Audio Onset Detection Results.* http://www.music-ir.org/evaluation/mirex-results/audio-onset/index.html. 2005

[MKP02] Marolt, Matija ; Kavcic, Alenka ; Privosnik, Marko: Neural networks for note onset detection in piano music. In: *Proceedings of the International Computer Music Conference, ICMC*, 2002

[Moe02] Moelants, Dirk: Preferred tempo reconsidered. In: *Proceedings of the 7th International Conference on Music Perception and Cognition, ICMPC*, 2002

[Moo82] Moore, Brian C. J.: *An introduction to the psychology of hearing.* 2. Ausgabe. Academic Press, 1982

[Moo95] Moore, Brian C. J.: Frequency analysis and masking. In: *Hearing.* Academic Press, 1995

[MPE01] MPEG, ISO/IEC JTC1/SC29/WG11: *Multimedia Content Description Interface - part 4: Audio, Final Committee Draft 15938-4.* 2001

[Mus05] Music Genome Project. *Pandora.* http://www.pandora.com/. 2005

[MW05] McLeod, Philip ; Wyvill, Geoff: A smarter way to find pitch. In: *Proceedings of International Computer Music Conference, ICMC*, 2005

[NM99] van Noorden, Leon ; Moelants, Dirk: Resonance and the perception of musical pulse. In: *Journal of New Music Research* 28 (1999), Nr. 1, S. 43–66

[Ori01] Orife, Iroro Fred O.: *RIDDIM: A rhythm analysis and decomposition tool based on Independent Subspace Analysis*, Dartmouth College, Hanover, New Hampshire, Diplomarbeit, 2001

[Par94] Parncutt, Richard: A perceptual model of pulse salience and metrical accent in musical rhythms. In: *Music Perception* 11 (1994), Nr. 4, S. 409–464

[Par04] Pardo, Bryan: Tempo tracking with a single oscillator. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[PAT04] Pikrakis, Aggelos ; Antonopoulos, Iasonas ; Theodoridis, Sergios: Music meter and tempo tracking from raw polyphonic audio. In: *Proceedings of International Conference on Music Information Retrieval, ISMIR*, 2004

[PC95] Plack, Christopher C. ; Carlyon, Robert P.: Loudness perception and intensity coding. In: *Hearing*. Academic Press, 1995

[PD03] Patel, Aniruddh D. ; Daniele, Joseph R.: An empirical comparision of rhythm in language and music. In: *Cognition* 87 (2003), S. B35–B45

[Pie99] Pierce, John R.: The Nature of Musical Sound. In: *The psychology of music*. Academic Press, 1999

[Pin04] van Pinxteren, Markus: *Implementierung eines Verfahrens zur Segmentierung von polyphonen Audiosignalen*, Technische Universität Ilmenau, Diplomarbeit, 2004

[PK90] Palmer, C. ; Krumhansl, C.L.: Mental representation for musical meter. In: *Journal of Experimental Psychology: Human Perception and Performance* 15 (1990), Nr. ?, S. 728–741

[PK02] Paulus, Jouni ; Klapuri, Anssi: Measuring the similarity of rhythmic patterns. In: *Proceedings of the 3rd International Conference on Music Information Retrieval, ISMIR*, 2002

[PK03a] Paulus, Jouni ; Klapuri, Anssi: Conventional and periodic N-grams in the transcription of drum sequences. In: *Proceedings of the International Conference on Multimedia and Expo, ICME*, 2003

[PK03b] Paulus, Jouni ; Klapuri, Anssi: Model-based event labeling in the transkription of perkussive audio signals. In: *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx*, 2003, S. 73–77

[Plu94] Plumbley, M.: Algorithms for Non-Negative Independent Component Analysis. In: *IEEE Transactions on Neural Networks* 14 (1994), Nr. 3

[Pos04] Pospeschill, Markus: *Konnektionismus und Kognition*. Kohlhammer Verlag, 2004

[Pov94] Povel, Dirk J.: A theoretical framework for rhythm perception. In: *Psychological Research* 45 (1994), S. 315–337

[Pre04] Pressing, Jeffrey: Cognitive complexity and the structure of musical patterns. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[Püs88] Püschel, Dirk: *Prinzipien der zeitlichen Auflösung beim Hören*, Georg-Augst-Universität Göttingen, Diss., 1988

[Pus05] Puschnerus, Arne. *Tempo.* Persönliche Kommunikation. 2005

[Rap01] Raphael, Christopher: Automated rhythm transcription. In: *Proceedings of the 2nd International Conference on Music Information Retrieval, ISMIR*, 2001

[Ras79] Rasch, R. A.: Synchronization in performed ensemble music. In: *Acustica* 43 (1979), S. 121–131

[RDF04] RDF, Core Working Group / W3C Semantic Web Activity. *Resource description framewok - primer.* http://www.w3.org/TR/rdf-primer/. 2004

[RGM94] Rosenthal, David ; Goto, Masataka ; Muraoka, Yoichi: Rhythm tracking using multiple hypothesis. In: *Proceedings of the International Computer Music Conference, ICMC*, 1994

[RJ01] Rodet, Xavier ; Jaillet, Florent: Detection and modeling of fast attack transients. In: *Proceedings of the International Computer Music Conference, ICMC*, 2001

[Roe00] Roederer, Juan G.: *Physikalische und psychoakustische Grundlagen der Musik.* 3. Ausgabe. Springer Verlag, 2000

[Ros92] Rosenthal, D. A.: Emulation of human rhythm perception. In: *Computer Music Journal* 16 (1992), Nr. 1, S. 64–76

[Ros00] Rossing, Thomas D.: *The science of percussion instruments.* World Scientific, 2000

[RWD02] Repp, Bruno H. ; Windsor, W. L. ; Desain, Peter: Effects of tempo on the timing of simple musical rhythms. In: *Music Perception* 19 (2002), Nr. 4, S. 565–594

[Sac53] Sachs, Curt: *Rhythm and tempo: A study in music history.* Columbia University Press, 1953

[SBR03] Supper, Ben ; Brookes, Tim ; Rumsey, Francis: A new approach to detecting auditory onsets within a binaural stream. In: *Proceedings of the AES 114th Convention*, 2003

[Sch85] Schloss, Walter A.: *On the automatic transcription of percussive music - from acoustic signal to high-level analysis*, Stanford University, Diss., 1985

[Sch98] Scheirer, Eric D.: Tempo and beat analysis of acoustic musical signals. In: *Journal of the Acoustical Society of America* (1998)

[Sch00] Scheirer, Eric D.: *Music-listening systems*, Massachusetts Institute of Technology, Diss., 2000

[Sep01] Seppänen, Jarno: Tatum grid analysis of musical signals. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2001

[Ser97] Serra, Xavier: Musical Sound Modelling with Sinusoids plus Noise. In: *Musical Signal Processing*. Swets and Zeitlinger, 1997

[Set93] Sethares, William A.: Local consonance and the relationship between timbre and scale. In: *Journal of the Acoustical Society of America* 94 (1993), Nr. 3

[Sil00] Sillanpää, Jukka: Drum stroke recognition / Tampere University of Technology. 2000. – Forschungsbericht

[SKSV00] Sillanpää, Jukka ; Klapuri, Anssi ; Seppänen, Jarno ; Virtanen, Tuomas: Recognition of acoustic noise mixtures by combined bottom-up and top-down processing. In: *Proceedings of the European Signal Processing Conference, EUSIPCO*, 2000

[SMS05] Sethares, William A. ; Morris, Robin D. ; Sethares, James C.: Beat tracking of musical performances using low-level audio features. In: *IEEE Transactions on Speech and Audio Processing* 13 (2005), Nr. 2

[SP98a] Shmulevich, Ilya ; Povel, Dirk-Jan: Complexity measures of musical rhythms. In: *Proceedings of the Rhythm Perception and Production Workshop*, 1998

[SP98b] Shmulevich, Ilya ; Povel, Dirk-Jan: Rhythm complexity measures for music pattern recognition. In: *Proceedings of IEEE Workshop on Multimedia Signal Processing*, 1998

[SP03] Shannon, Ben J. ; Paliwal, Kuldip K.: A comparative study of filter bank spacing for speech recognition. In: *Proceedings of the Microelectronic Engineering Research Conference*, 2003

[Spi03] Spitzer, Manfred: *Musik im Kopf.* Schattauer, 2003

[Spo98] Sporer, Thomas: *Qualitätsbeurteilung von Audiosignalen mittels gehörangepasster Messverfahren*, Universität Erlangen-Nürnberg, Diss., 1998

[SS99] Sethares, William A. ; Staley, Malcolm W.: Periodicity transforms. In: *IEEE Transactions on Signal Processing* 47 (1999), Nr. 11

[SS01] Sethares, William A. ; Staley, Malcolm W.: Meter and periodicity in musical performance. In: *Journal of New Music Research* (2001)

[Tan93] Tanguiane, Andranick S.: *Artificial perception and music recognition.* Springer Verlag, 1993

[TEC01a] Tzanetakis, George ; Essl, Georg ; Cook, Perry: Audio Analysis using the Discrete Wavelet Transform. In: *Proceedings of the WSES International Conference on Acoustics and Music: Theory and Applications*, 2001

[TEC01b] Tzanetakis, George ; Essl, Georg ; Cook, Perry: Automatic musical genre classification of audio signals. In: *Proceedings of the 2nd International Conference on Music Information Retrieval, ISMIR*, 2001

[TEC02] Tzanetakis, George ; Essl, Georg ; Cook, Perry: Human Perception and Computer Extraction of Musical Beat Strength. In: *Proceedings of the 5th International Conference on Digital Audio Effects, DAFx*, 2002

[Tem01] Temperley, David: *The cognition of basic musical structures.* MIT Press, 2001

[Tem04] Temperley, David: An evaluation system for metrical models. In: *Computer Music Journal* 28 (2004), Nr. 3, S. 28–44

[Ter98] Terhardt, Ernst: *Akustische Kommunikation.* Springer Verlag, 1998

[The96] The MIDI Manufacturers Association: *The Complete MIDI 1.0 Detailed Specification.* Zweite Edition. 1996

[TNS03] Takeda, Haruto ; Nishimoto, Takuya ; Sagayama, Shigeki: Automatic rhythm transcription from multiphonic MIDI signals. In: *Proceedings of the 4th International Conference on Music Information Retrieval, ISMIR*, 2003

[Tou02] Toussaint, Godfried: A mathematical comparision of African, Brazilian and Cuban clave rhythms. In: *Proceedings of BRIDGES: Mathematical Connections in Art, Music and Science*, 2002

[UD04a] Uhle, Christian ; Dittmar, Christian: Drum pattern based genre classification of popular music. In: *Proceedings of the AES 25th International Conference on Metadata*, 2004

[UD04b] Uhle, Christian ; Dittmar, Christian: Generation of musical scores of percussive un-pitched instruments from automatically detected events. In: *Proceedings of the AES 116th Convention*, 2004

[UDS03] Uhle, Christian ; Dittmar, Christian ; Sporer, Thomas: Extraction of drum tracks from polyphonic music using independent subspace analysis. In: *Proceedings of 4th International Symposium on Independent Component Analysis, ICA-03*, 2003

[UH03] Uhle, Christian ; Herre, Juergen: Estimation of tempo, micro time and time signature from percussive music. In: *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx*, 2003

[Uhl05] Uhle, Christian. *Tempo induction by investigating the metrical structure of music using a periodicity signal that relates to the tatum period.* http://www.music-ir.org/evaluation/mirex-results/articles/tempo/uhle.pdf. 2005

[UZ99] Uitdenbogerd, Alexandra ; Zobel, Justin: Matching techniques for large music databases. In: *Proceedings of the 7th ACM International Multimedia Conference*, 1999, S. 57–66

[Vei03] Veigel, Alexis: Implementierung eines Verfahrens zur Segmentierung von Musiksignalen / Fraunhofer Institut für Digitale Medientechnologie. 2003. – Praktikumsbericht

[VIS04] Vacher, Michel ; Istrate, Dan ; Serignat, Jean-Francois: Sound detection and classification through transient models using wavelet coefficient trees. In: *Proceedings of the 12th European Signal Processing Conference, EUSIPCO*, 2004

[Wey01] Weyde, Tillman: Grouping, similarity and the recognition of rhythmic structure. In: *Proceedings of the International Computer Music Conference, ICMC*, 2001

[Yat95] Yates, Graeme K.: Cochlear structure and function. In: *Hearing.* Academic Press, 1995

[Yes76] Yeston, M.: *The stratification of musical rhythm*. Yale University Press, 1976

[YGO04] Yoshii, Kazuyoshi ; Goto, Masataka ; Okuno, Hiroshi G.: Automatic drum sound description for real-world music using template adaption and matching methods. In: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, 2004

[ZBH05] van Zaanen, Menno ; Bod, Rens ; Honing, Henkjan: A memory-based approach to meter induction. In: *Proceedings of the 5th Triennial ESCOM Conference*, 2005

[Zen94] Zenner, Hans-Peter: *Hören: Physiologie, Biochemie, Zell- und Neurobiologie*. Georg Thieme Verlag, 1994

[ZF99] Zwicker, Eberhard ; Fastl, Hugo: *Psychoacoustics*. 2. dt. Auflage. Springer Verlag, 1999

[ZP03] Zils, Aymeric ; Pachet, François: Extracting automatically the perceived intensity of music titles. In: *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx*, 2003

[ZPDG02] Zils, Aymeric ; Pachet, François ; Delerue, Olivier ; Gouyon, Fabien: Automatic extraction of drum tracks from polyphonic music signals. In: *Proceedings of the 2nd International Conference on Web Delivering of Music, WEDELMUSIC*, 2002

Verzeichnis der Abkürzungen

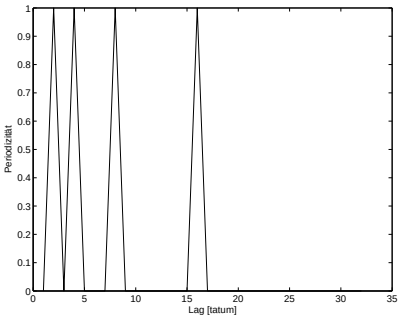
ABF	Amplitudenbasisfunktionen
ADC	Audio Description Contest
AKF	Autokorrelationsfunktion
AMDF	Average Magnitude Difference Function
ASA	Auditory Scene Analysis
ASE	Audio Spectrum Envelope
BD	Bassdrum (Große Trommel)
BM	Basilarmembran
BS	Beatstärke
BTS	Beattracking-System
CASA	Computational Auditory Scene Analysis
CD	Compact Disc
CMNDF	Cumulative Mean Normalized Difference Function
bpm	Beats per minute
BTS	Beat Tracking System
DCT	Diskrete Kosinus Transformation
DFT	Diskrete Fourier Transformation
EDS	Extractor Discovery System
ERB	Equivalent Rectangular Bandwidth
GGT	Größter Gemeinsamer Teiler
GZV	Ganzzahiges Vielfaches
HMM	Hidden Markov Models
HWR	Half-wave rectification (Einweg- bzw. Halbwellen-Gleichrichtung)
HZ	Haarzellen
ICA	Independent Component Analysis
IDMT	Institut für Digitale Medientechnik
IIS	Institut für Integrierte Schaltungen
IOI	Inter Onset Interval
JND	Just Noticeable Difference

LDA	Lineare Diskriminanzanalyse
LDS	Leistungsdichtespektrum
MIDI	Musical Instruments Digital Interface
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation Exchange
NNICA	Non-Negative Independent Component Analysis
QBB	Query-by-beatboxing
SACF	Summary Autocorrelation Function (Summierte Autokorrelationsfunktion)
SFM	Spectral Flatness Measure (Spektrales Flachheitsmaß)
SD	Snaredrum (Kleine Trommel)
SDF	Squared Difference Function
SPL	Sound Pressure Level (Schalldruckpegel)
STFT	Short Time Fourier Transform (Kurzzeit-Fouriertransformation)
TPRW	Three-Point Running Window
TWME	Two-Way Mismatch Error
VP	Versuchsperson
WQ	Weber'scher Quotient

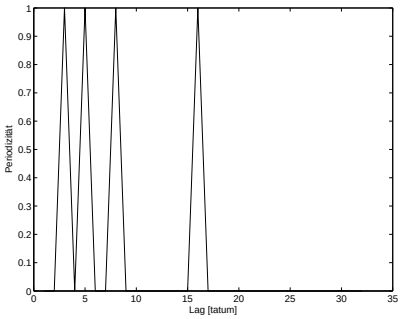
Symbolverzeichnis

A	Akzentsignal
D	Differenziertes und halbweggleichgerichtes Hüllkurvensignal
E	Hüllkurvensignal
H_P	Patternhistogramm
k	Konfidenzmaß
m	Mikrotime
M	Drumpattern
p_B	Beatperiodenlänge
p_T	Tatumperiodenlänge
Q	zeitlich quantisierte Instrumenteninformation
r_B	Beatraster
r_T	Tatumraster
R	Metrisches Template
S	Ähnlichkeitsmaß
V	Periodizitätenprofil

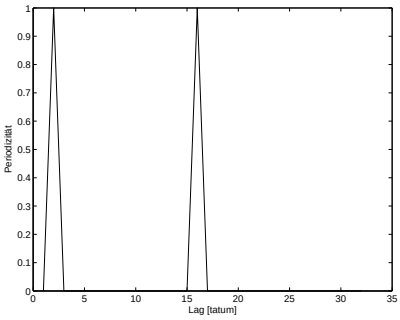
A Metrische Templates



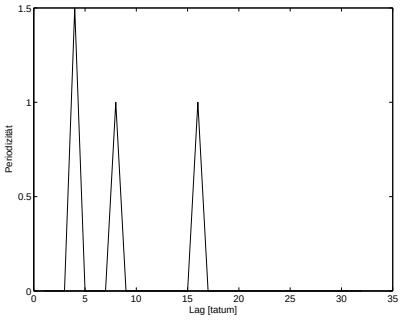
(a)



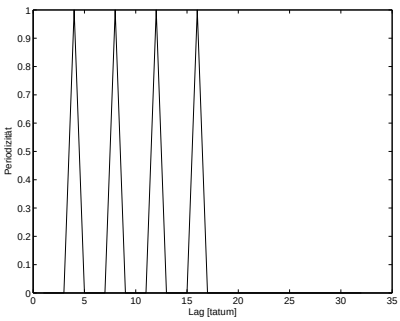
(b)



(c)

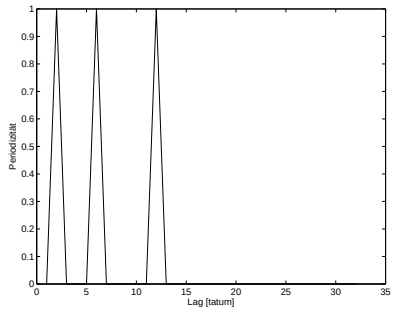


(d)

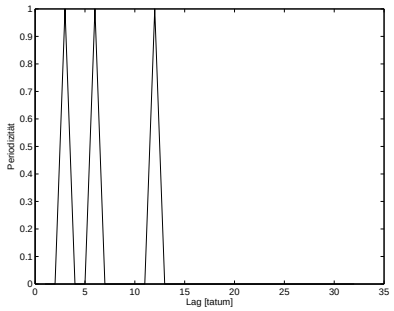


(e)

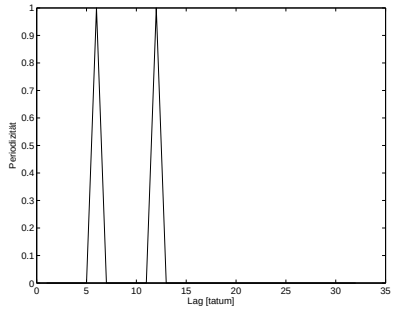
Abbildung A.1: Metrische Templates der Klasse 1: 4/4-Takt mit binärer Mikrotime.



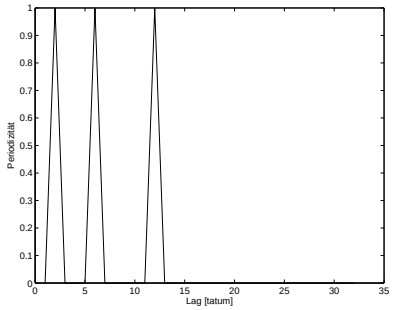
(a)



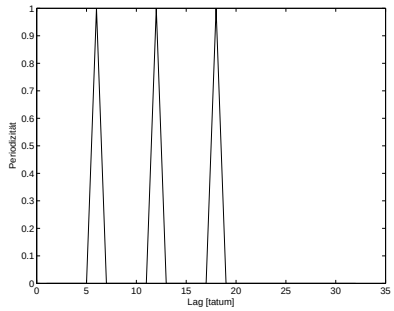
(b)



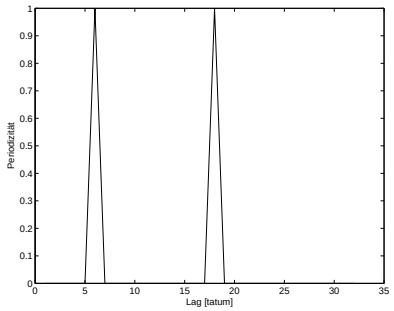
(c)



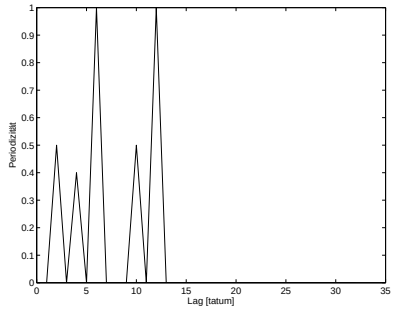
(d)



(e)



(f)



(g)

Abbildung A.2: Metrische Templates der Klasse 2: 4/4-Takt mit ternärer Mikrotime, 3/4- und 6/8-Takt mit binärer Mikrotime.

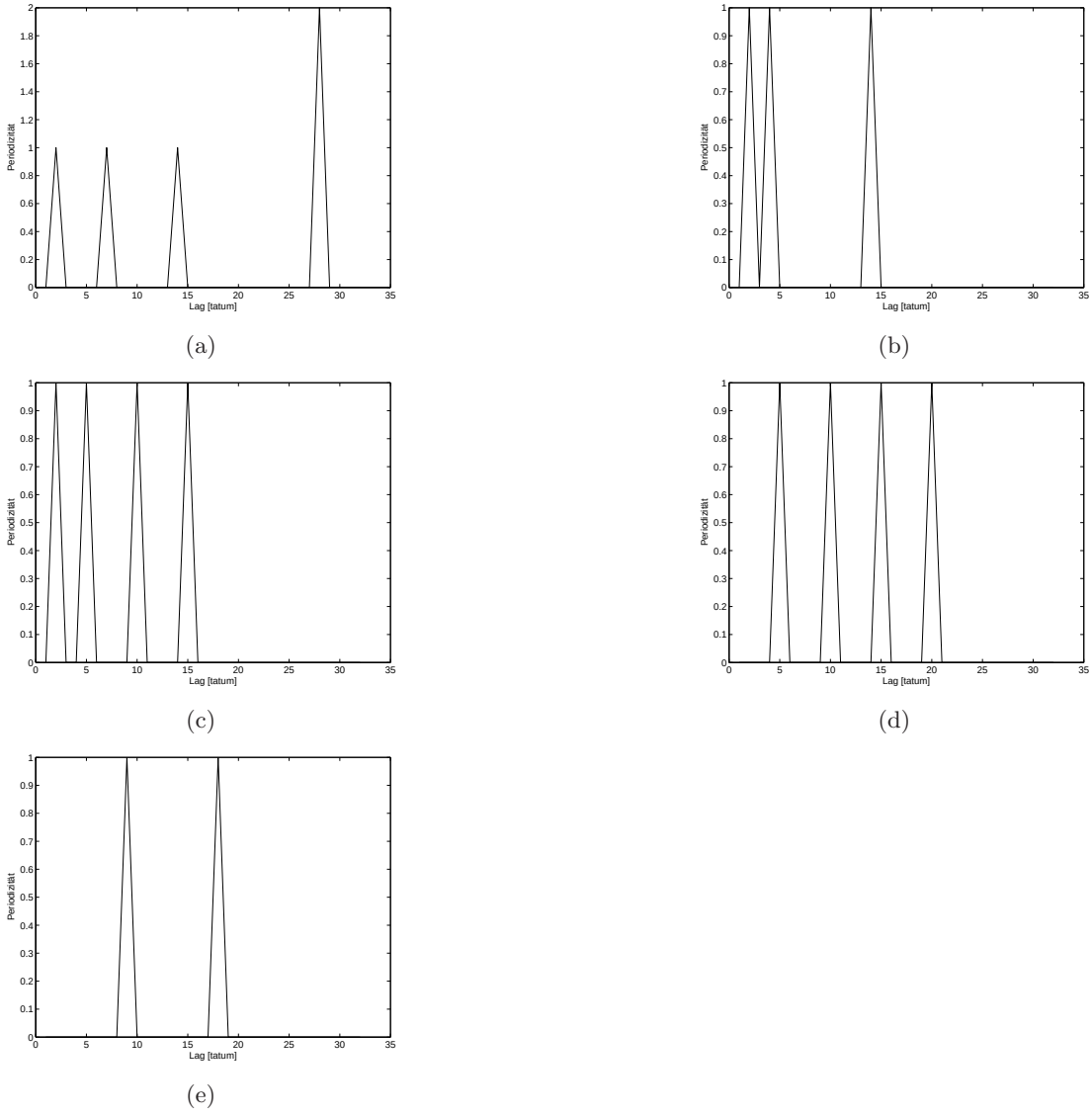


Abbildung A.3: Metrische Templates der Klasse 3: 3/4-Takt mit ternärer Mikrotime und zusammengesetzte Taktarten.

B Tempo Test

	Kategorie	A	B	C	D	E	F	G	H
1	schnell	213	213	107	105	106	106	212	211
2		198	196	98	98	196	196	196	196
3		242	232	117	114	234	112	235	234
4	moderat	150	143	144	72	143	144	146	146
5		115	115	116	57	115	115	115	115
6		120	120	120	59	120	120	121	120
7		74	74	74	74	73	75	148	74
8	Übereinstimmung	86	85	86	86	86	86	86	85
9	ternär, 2 Ebenen	44	133	45	44	134	45	134	45
10		52	158	48	52	157	52	158	53
11		43	42	126	42	127	42	133	42
12		69	210	70	70	70	70	209	210
13		42	40	122	40	116	40	116	122
14		45	136	137	45	136	136	136	45
15		0	120	120	40	119	120	120	40
16	ternär, 3 Ebenen	82	82	82	82	82	82	247	123
17		62	93	93	93	186	93	189	186
18		57	170	170	56	57	57	115	56
19		77	76	78	76	76	114	230	77
20		47	93	144	47	140	47	141	140
21	additiver Takt	129	129	130	65	129	194	130	131

Tabelle B.1: Ergebnisse des Hörtests zur Tempobestimmung mit 8 Testhörern: Die Teststücke sind mit Ausnahme von (8) nach Eigenschaften der metrischen Struktur kategorisiert.

C Parametertest

Die folgenden Tabellen geben eine Übersicht über die Entwicklung der Erkennungsraten der Temposchätzung in Abhängigkeit einer Auswahl von Parametern. Zur Beurteilung der Erkennungsleistung sind die Anzahl der Richtig- und Falschklassifikationen E_r beziehungsweise E_f sowie das Verhältnis der Anzahlen der Ermittlung des doppelten und halben Tempos E_d beziehungsweise E_h angegeben.

Distanzmaß	Erkennungsrate der High-level-Analyse [%]		
	E_r	E_f	E_d/E_h
Hamming	50.41	26.33	0.28
L1-Norm	43.67	30.20	0.10
L2-Norm	43.27	30.41	0.12
AKF	49.98	28.78	0.27
SDF	43.27	30.41	0.12

Tabelle C.1: Einfluss des Ähnlichkeitsmaßes zur Berechung der Periodizität in symbolischen Darstellungen auf die Erkennungsrate.

c	Low-level			Beathistogramm			High-level		
	E_r	E_f	E_d/E_h	E_r	E_f	E_d/E_h	E_r	E_f	E_d/E_h
0	67.96	8.16	1.02	61.02	19.18	1.02	50.00	26.33	0.29
1	68.37	6.94	0.98	62.65	17.35	1.04	49.59	26.94	0.26
2	69.59	6.33	1.03	63.06	16.94	0.96	51.22	25.92	0.30
3	71.02	6.73	1.02	64.08	15.92	0.96	50.41	26.33	0.28
4	70.00	6.94	1.05	64.08	15.71	1.02	51.22	25.71	0.31

Tabelle C.2: Einfluss der Wichtung der Teilbänder zur Bildung des Akzentsignals auf die prozentuale Erkennungsrate.

d	Low-level			High-level		
	E_r	E_f	E_d/E_h	E_r	E_f	E_d/E_h
0	68.36	7.55	1.02	50.41	25.92	0.30
0.5	70.82	6.94	0.98	50.61	25.71	0.30
1.0	68.78	7.14	1.00	50.20	25.52	0.31
1.5	68.16	7.75	0.98	50.20	25.52	0.31

Tabelle C.3: Einfluss der Konstante d bei der Wichtung der Tatumschätzungen auf die prozentuale Erkennungsrate.

σ	Low-level			Beathistogramm			High-level		
	E_r	E_f	E_d/E_h	E_r	E_f	E_d/E_h	E_r	E_f	E_d/E_h
0.0	61.22	7.14	1.25	54.08	11.02	1.34	44.29	33.47	0.18
0.5	69.80	6.73	1.13	64.08	13.88	1.08	49.39	27.14	0.29
1.0	71.02	6.73	1.02	64.08	15.92	0.96	50.41	26.33	0.28
1.5	71.02	6.94	0.96	62.86	17.96	1.00	50.82	26.12	0.27
2.0	71.02	7.35	0.93	61.84	19.18	0.98	50.82	26.53	0.26

Tabelle C.4: Einfluss der A-prior-Wahrscheinlichkeit für das Auftreten von Beatperioden auf die prozentuale Erkennungsrate.

Erkennungsrate	einkanaliges Akzentsignal	mehrkanaliges Akzentsignal
Temposchätzung	70.82	70.82
Taktart	88.16	82.45

Tabelle C.5: Erkennungsraten für Periodizitätenberechnung im ein- und mehrkanaligen Akzentsignal für Tempo- und Taktschätzung mittels Low-level-Analyse des Trainingsdatensatzes.

D Ergebnisse des Vergleichs mit anderen Verfahren

Teilnehmer	RT1 [%]	RT2 [%]	Zeit
Klapuri	67.29	85.01	0.5
Uhle	51.61	76.11	0.1
Anonymous	45.26	81.21	15
Dixon (auco)	38.82	82.3	1
Scheirer	37.85	68.08	0.4
Alonso (sppr)	36.29	69.77	0.1
Dixon (indu)	31.76	73.6	0.02
Tzan (medsumbands)	31.22	55.51	2
Tzan (medmultibands)	30.76	50.73	2
Alonso (auco)	27.78	57.89	0.1
Dixon (trac)	26.56	74.3	0.1
Tzan (histsumbands)	25.22	54.67	2

Tabelle D.1: Ergebnisse des ADC 2004 Audio Description Contest. Die verglichenen Implementierungen sind nach der Prozentzahl richtig geschätzter Tempi (Spalte RT1) sortiert. Spalte RT2 enthält die Prozentzahl der Schätzungen, dem Doppelten, Halben, Dreifachen oder einem Drittel des richtigen Tempos oder dem richtigen Tempo entsprechen. Weiterhin ist die durchschnittlich benötigte Rechenzeit bezogen auf die Länge des Audiosignals angegeben.

Rg.	Teilnehmer	Bewertung (Std.-abweichung)	RT1 [%]	RT2 [%]	RP1 [%]	RP2 [%]	Zeit [s]
1	Alonso, David, Richard	0.689 (0.231)	95.00	55.71	25.00	5.00	2875
2	Uhle 1	0.675 (0.273)	90.71	59.29	32.14	7.14	1160
3	Uhle 2	0.675 (0.272)	90.71	59.29	32.86	6.43	2621
4	Gouyon, Dixon 1	0.670 (0.252)	92.14	56.43	40.71	7.86	3303
5	Peeters	0.656 (0.223)	95.71	47.86	27.86	4.29	2159
6	Gouyon, Dixon 2	0.649 (0.253)	92.14	51.43	37.14	5.71	2050
7	Gouyon, Dixon 4	0.645 (0.294)	87.14	55.71	48.57	10.71	1357
8	Eck	0.644 (0.300)	86.43	53.57	37.14	5.71	1665
9	Davies, Brossier	0.628 (0.284)	86.43	48.57	26.43	4.29	1005
10	Gouyon, Dixon 3	0.607 (0.287)	87.14	47.14	36.43	6.43	1388
11	Sethares	0.597 (0.252)	90.71	37.86	30.71	0.71	70975
12	Brossier	0.583 (0.333)	80.71	51.43	28.57	2.14	180
13	Tzanetakis	0.538 (0.359)	71.43	50.71	28.57	3.57	7173

Tabelle D.2: Ergebnisse des MIREX 2005 Audio Tempo Extraction Contest. Die verglichen Implementierungen sind nach dem Rang (Rg.) sortiert, der aus dem Bewertungsmaß (Spalte 3) folgt. Angegeben sind die Prozentzahlen für „Mindestens ein Tempo korrekt“ (RT1), „Beide Tempi korrekt“ (RT2), „Mindestens eine Phase korrekt“ (RP1), „Beide Phasen korrekt“ (RP2) und die Rechenzeit, die für alle Teststücke benötigt wurde.

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit am Fraunhofer Institut für Digitale Medientechnologie in Ilmenau. Ich möchte an dieser Stelle dem Leiter des Instituts, Prof. Dr.-Ing. Karlheinz Brandenburg, für die Möglichkeit der Bearbeitung dieses Themas und die Erstellung des Erstgutachtens meinen Dank aussprechen.

Für die Erstellung der weiteren Gutachten danke ich den Herren Prof. Dr.-Ing. Jürgen Wernstedt und Dr.-Ing. Jürgen Herre.

Bei meiner Arbeit unterstützten mich verschiedene Personen. Wertvolle Hinweise bei Fragen verschiedenster Art erhielt ich von Dr.-Ing. Thomas Sporer. Für die kritische Durchsicht der Arbeit bedanke ich mich bei Dr. Werner Uhle. Meinen Kollegen am Institut danke ich für die kreative Zusammenarbeit. Für die Implementierung des Verfahrens zur Klassifikation der perkussiven Instrumente gebührt Christian Dittmar Anerkennung. Zur Implementierung der entwickelten Algorithmen wurde das von Wolfgang Hirsch entwickelte Signalverarbeitungssystem verwendet.

Allen Studenten, deren Medienprojekte, Studien- und Diplomarbeiten einen Beitrag zur vorliegenden Arbeit leisteten, möchte ich an dieser Stelle ebenso danken.