Automatische Erkennung von Onsets in Musiksignalen zur Steuerung von Beattracking-Systemen

Magisterarbeit
eingereicht am 18.8.2004

[redacted]

Betreuer:
Alexander Lerch, zplane.development
Tim Flohrer, zplane.development
Prof. Dr. Stefan Weinzierl

Fachgebiet Kommunikationswissenschaft
Institut für Sprache und Kommunikation
Fakultät I der Technischen Universität Berlin

Die selbständige und eigenhändige Anfertigung versichere ich an Eides statt.

Berlin, den

_____

Ingmar-L. Klich

Kurzfassung

Die vorliegende Arbeit beschreibt die Untersuchung und Optimierung von Verfahren, um automatisch in einem unbekannten Musiksignal so genannte *Onsets* zu erkennen. Ein Onset bezeichnet den Eintritt eines musikalischen Ereignisses, wie beispielsweise den Beginn eines von einem Musikinstrument gespielten Tones oder Schlages. Die korrekte Erkennung von Onsets bildet die Grundlage für das *Beattracking*, einer Disziplin des *Music Information Retrieval (MIR)*. Hierbei geht es darum, automatisch aus einem Musiksignal Tempo und Lage (Phase) der Schlagzeiten (engl. Beat) zu bestimmen, eventuell auch die Taktart zu berechnen. Eine automatische Onset-Erkennung wird neben dem Beattracking auch in verschiedenen anderen Systemen aus dem Forschungsbereich des MIR benötigt; Beispiele hierfür sind Harmonieerkennung und automatische Transkription von Notenwerten aus dem Audiosignal. So können Ergebnisse dieser Arbeit durchaus auch zur Entwicklung dieser Systeme beitragen.

Mehrere aus der Literatur bekannte Verfahren zur automatischen Onset-Erkennung werden beschrieben und allgemeine Stärken und Schwächen diskutiert. Basierend darauf wird ein neuer Ansatz zur Optimierung dieser Ansätze untersucht. Die Ergebnisse dienen zur Auswertung der Algorithmen, um im Anschluss ein System zu entwickeln, dass möglichst robust gegenüber Genre und Komplexität der Musiksignale am Eingang reagiert und damit die Voraussetzungen erfüllt, in Beattracking-Systemen eingesetzt zu werden. Hierzu werden die bekannten Verfahren durch zusätzliche Funktionen erweitert, die dem Algorithmus ermöglichen sollen, sich optimal an das Eingangssignal anzupassen. Zusätzlich zu der Erkennung und Extraktion der Onset-Zeiten, werden diese hinsichtlich ihrer Intensität und ihrer Relevanz im rhythmischen Gefüge eingestuft. Dies geschieht im Hinblick darauf, dem Beattracking-System weitere Informationen zu liefern und es bei der Differenzierung wichtiger und weniger wichtiger Onsets sowie einer eventuellen Takterkennung zu unterstützen.

Abschließend erfolgt Evaluierung und Test des Gesamtsystems. Hierbei wird zunächst die Methodik diskutiert und eine Methode zur anschließenden Evaluierung des Systems entwickelt. In der Evaluierung wird das Gesamtsystem sowohl mit den bereits vorhandenen automatischen Systemen verglichen, als auch gegen ein MIDI-Signal und die Fähigkeiten einer menschlichen Wahrnehmung getestet. Als Unterscheidungsmerkmal der automatischen Systeme steht dabei die Robust-

heit gegenüber Genre, Instrumentation und Stimmenanzahl des zu verarbeitenden Musiksignals im Vordergrund. Zu dieser Evaluierung wird selbst erstelltes Testmaterial verwendet, das unter dem Aspekt zusammengestellt wurde, die Fähigkeiten möglichst objektiv aufzuzeigen. Den Abschluss der Evaluierung bildet einen kurzes Fazit über die angewandte Evaluierungsmethode.

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Motivation . . . 1
1.2 Zielsetzung und Überblick . . . 3

2 Grundlagen **5**
2.1 Der Onset-Begriff . . . 5
2.1.1 Onset im Kontext des Beattracking . . . 6
2.1.2 Zeitliche Auflösung der Onset-Wahrnehmung . . . 8
2.2 Signaltheoretische Grundlagen . . . 9
2.2.1 Das digitale Audiosignal . . . 9
2.2.2 Filter . . . 10
2.2.3 Fouriertransformation . . . 11
2.2.4 Stochastische Größen . . . 15
2.2.5 Auto- und Kreuzkorrelation . . . 16
2.2.6 Lineare Prädiktion . . . 17

3 Verfahren der Onset-Erkennung **19**
3.1 Allgemeiner Ablauf . . . 19
3.2 Literaturübersicht . . . 20
3.3 Algorithmen im Zeitbereich . . . 21
3.3.1 Signalenergie und Hüllkurve . . . 21
3.3.2 Stärken der Verfahren . . . 25
3.3.3 Bekannte Schwächen . . . 25
3.3.4 Optimierung . . . 27
3.4 Algorithmen im Frequenzbereich . . . 30
3.4.1 Distanzen im Betragsspektrum . . . 31
3.4.2 Distanz in der Komplexen Ebene . . . 32
3.4.3 Merkmale frequenzbasierter Verfahren . . . 34
3.5 Ansatz mit Linearer Prädiktion . . . 36
3.5.1 Idee . . . 36
3.5.2 Algorithmus . . . 36
3.5.3 Anwendbarkeit . . . 38
3.6 Peakpicking . . . 41

4 Entwicklung des optimierten Algorithmus **43**
4.1 Auswahl des Basisalgorithmus 43
4.1.1 Zeit- versus Frequenzbereich 43
4.1.2 Vergleich der Verfahren im Frequenzbereich 44
4.1.3 Struktur des Basisalgorithmus 50
4.2 Onset-Erkennung in Frequenzbändern 51
4.2.1 Das Problem der Bandaufteilung 51
4.2.2 Bandaufteilung auf Basis von Feature-Extraktion 54
4.2.3 Optimiertes Peakpicking 56
4.2.4 Intensität des Onsets 57
4.2.5 Zusammenfassung der Frequenzbänder 58
4.2.6 Erweiterte Struktur des Systems 60
4.3 Onset-Klassifizierung zur Takterkennung 61
4.3.1 Schwere Zählzeiten 61
4.3.2 Korrelationsmethode 61
4.4 Gesamtsystem 64
4.4.1 Datenformat 65

5 Evaluierung **67**
5.1 Methodik zur Evaluierung 67
5.2 Entwicklung der Evaluierungsmethode 68
5.2.1 Ziel dieser Evaluierung 68
5.2.2 Allgemeine Probleme zur Evaluierung von Systemen zur automatischen Onset-Erkennung 69
5.2.3 Evaluierungskriterien 69
5.2.4 Auswahl und Erstellung des Testmaterials 72
5.2.5 Durchführung 73
5.3 Ergebnisse der Evaluierung 74
5.3.1 Statistik erkannter Onsets 74
5.3.2 Ergebnisse der Erkennungsmaße 75
5.3.3 Ergebnisse der statistischen Abweichungsmaße 76
5.4 Fazit der Evaluierungsmethode 78

6 Resümee **79**
6.1 Zusammenfassung 79
6.2 Ausblick 80

Kapitel 1

Einleitung

1.1 Motivation

Rhythmus ist eine wichtige Information in Musik. Er beschreibt die Ordnung und Gliederung des zeitlichen Verlaufes der Klangereignisse durch unterschiedliche Dauern, Akzente sowie wechselnde Klänge und Klangfarben. Der Hörer empfindet den Rhythmus als regelmäßig wiederkehrende Muster, gruppiert hierzu einzelne Ereignisse und bestimmt daraus subjektiv Takt und Tempo. Wichtiger Bestandteil des Rhythmus sind die *Schlagzeiten* - allgemein als *Beat* (engl.) oder früher als *Tactus* bezeichnet. Der Beat ist der Puls des Musikstückes, markiert durch äquidistante Schläge in einem gewissen Tempo. Unter *Beattracking* versteht man den Vorgang, in dem aus einem laufenden Musikstück die Schlagzeiten ermittelt werden. Ein Mensch ist selbst bei geringen musikalischen Vorkenntnissen schnell in der Lage, den Beat beispielsweise durch Fußschläge, Kopfwippen oder Tanzen zu verfolgen und sich bei Temposchwankungen entsprechend anzupassen. Er verarbeitet dabei kleine kurzzeitige Abweichungen oder synkopische Betonungen und ist im rhythmischen Empfinden meist unabhängig von Genre und Qualität der Musikdarbietung.

Diese Fähigkeiten sollen nun so gut wie möglich auf die maschinelle Informationsverarbeitung übertragen werden, denn Erzeugung und Verarbeitung von Musik geschieht mehr und mehr in digitaler Form mit Rechnern. Durch den rasanten Fortschritt der Computertechnik und die Verbreitung des Internets existiert eine vielseitige und leistungsfähige technische Basis, Informationen aus Musiksignalen maschinell und automatisiert zu extrahieren. Die Fähigkeiten der Algorithmen zur computergestützten Verarbeitung von Audioinformation sind jedoch zum gegenwärtigen Zeitpunkt auf relativ klar strukturierte Problembereiche begrenzt und kommen an die menschliche Leistungsfähigkeit und Robustheit nicht heran. Die Forschung ist daher bestrebt, zur inhaltlichen Analyse komplexer Musikdaten die Art und Weise der menschlichen Verarbeitung durch Modellbildung zu approximieren.

Das Modell, um aus den rhythmischen Mustern den Beat zu extrahieren, basiert auf der Erkennung so genannter *Onsets (engl.)*. Das sind die Einsatzzeiten der in dem Stück vorkommenden Klangereignisse, wie zum Beispiel gespielte Töne oder Schläge. Für die Einordnung ins zeitliche Gefüge ist der Beginn eines musikalischen Ereignisses die wichtigste Information. Onsets sind daher ausschlaggebend für die rhythmische Empfindung und bilden die Grundlage für das Beattracking. Diese Arbeit thematisiert die automatische Erkennung von Onsets in Musiksignalen, eine Problematik, die nicht nur für automatische Beattracking-Systeme von Interesse, sondern beispielsweise auch für die automatische Transkription oder die Harmonieerkennung von Bedeutung ist. Derartige Systeme sind ein Teilbereich des *Music Information Retrieval (MIR)*. MIR ist ein Forschungsgebiet, das bereits seit Ende 1960 existiert und mit der zunehmenden Verbreitung des Internet und entsprechender digitaler Formate (zum Beispiel MP3) an Bedeutung gewonnen hat. Die zu lösenden Kernprobleme umfassen Technologien, die dem Menschen einen effizienten Zugriff auf umfangreiche Musikkollektionen ermöglichen. Hierzu sollen die Informationen direkt aus den Audiodaten extrahiert werden, um diese auf einer höheren Ebene zu interpretieren und Daten über charakteristische und klassifizierende Merkmale zu erhalten.

Die Algorithmen werden aber nicht nur zu Kategorisierungszwecken benötigt. Es gibt eine Vielzahl von Anwendungen im Bereich der computergestützten Musikproduktion oder künstlerischen Performance, in denen derartige Algorithmen verwendet werden können oder bereits erfolgreich eingesetzt werden. Als Beispiel seien hier DJ-Programme, Programme zur Umwandlung von Audio nach MIDI oder zum Segmentieren und Quantisieren von Musiksignalen genannt. Algorithmen zum automatischen Beattracking können beispielsweise zur Playlist-Generierung, zum so genannten Beatmixing oder zur automatischen Erzeugung von Metadaten (zum Beispiel ID-Tags bei MP3) eingesetzt werden. Weitere Anwendungen sind Tempokorrektur sowie die Synchronisierung von Audiospuren. Die Programme sind meist in größere Frameworks zur Audioverarbeitung eingebunden und integrieren ihrerseits wieder einen Algorithmus zur Onset-Erkennung.

Die Weiterentwicklung von informationsverarbeitenden Algorithmen wie der automatischen Onset-Erkennung basiert darauf, diese „intelligenter“ zu machen. Die grundlegende Struktur eines intelligenten Algorithmus beruht auf der starken Interaktion zwischen den elementaren Funktionseinheiten oder Modulen. So strebt man an, dass sich Module entweder gegenseitig oder sich selbst dynamisch anpassen, indem die Parametrisierungen steuerbar realisiert oder den Algorithmen gewisse Lernfähigkeiten verliehen werden. Weiterhin bedient man sich der Modellierung menschlicher kognitiver Fähigkeiten oder physiologischer Gegebenheiten, um dem Menschen nähere Vorgehensweisen der Informationsaufnahme und -verarbeitung zu entwerfen und daraus Methoden zur qualitativen Verbesserung und Anreicherung herkömmlicher Algorithmen abzuleiten. Beispiele sind Modelle der Psychoakustik.

Ein wichtiges Hauptziel der automatischen Musiksignalverarbeitung ist, dass die Algorithmen möglichst robust gegenüber Genre und Qualität der Musiksignale am Eingang reagieren. Meist ist es nicht schwer, ein System auf bestimmte Eingangssignale hin zu optimieren. Ebenso leicht findet man jedoch Gegenbeispiele, bei denen es versagt. Wie so oft ist ein ideales System nicht realisierbar, so dass man einen Kompromiss eingehen muss, dessen Güte die Robustheit der Systeme ausmacht. Andere Ziele der Optimierungen automatischer informationsverarbeitender Systeme sind Rechenperformanz und Echtzeit-Fähigkeit.

1.2 Zielsetzung und Überblick

Das Ziel dieser Arbeit ist die Entwicklung eines optimierten Verfahrens zur automatischen Onset-Erkennung in Musiksignalen auf der Basis aus der Literatur bekannter Ansätze. Für die Optimierung sind folgende Kriterien maßgeblich: Die Robustheit gegenüber Genre und Qualität der Musiksignale am Eingang sowie die Einsatzfähigkeit zur Steuerung automatischer Beattracking-Systeme. Weiterhin sind Recheneffizienz und sparsamer Ressourcenverbrauch zu beachten, damit die Möglichkeit besteht, das System später als Software zu implementieren. Die Echtzeitfähigkeit ist für diese Arbeit zwar wünschenswert, stellt aber kein dringend notwendiges Optimierungskriterium dar. Das System erhält als Eingabe ein digitales Musiksignal im PCM Format - wobei zur Vereinfachung von Signalen mit einer Abtastrate von 44.1 kHz ausgegangen wird - und soll am Ausgang zusätzlich zu dem Zeitpunkt auch Informationen zur Klassifizierung der in dem Musikstück vorkommenden Onsets liefern. Der Algorithmus soll neben dem Musiksignal keine weiteren Eingaben wie beispielsweise Genre oder Anzahl der Stimmen fordern, sondern sich an das Eingangssignal individuell anpassen und die hierzu nötigen Informationen aus dem Musiksignal extrahieren. Für diese Zielsetzung stellen sich zwei grundlegende Teilaufgaben:

- **Evaluierung bekannter Verfahren**. Zum Thema automatische Onset-Erkennung ist bereits eine Menge an Forschung betrieben worden. In der Literatur werden viele verschiedene Ansätze beschrieben und deren Leistungsfähigkeit durch vielversprechende Ergebnisse aufgezeigt. In dieser Arbeit sollen die vielversprechendsten und naheliegensten Ansätze gesammelt und im Hinblick auf die Erfüllung der obigen Maßgaben verglichen werden. Das Ziel dieser Teilaufgabe ist es, die Ausgangslage für eine nachfolgende Optimierung zu schaffen, das heißt, das oder die Verfahren auszuwählen, die in das endgültige System als Basisalgorithmus integriert und dessen oder deren Parametrisierungen optimiert werden.

- **Entwicklung eines optimierten Verfahrens**. Für die Optimierung stellt sich zunächst die Frage, an welchen Stellen des Basisalgorithmus eine Optimierung in Hinblick auf die Maßgaben möglich und sinnvoll ist. Ergänzend

dazu wird nach Verfahren gesucht, die in anderen Gebieten erfolgreich eingesetzt werden und deren Idee auf das Problem in dieser Arbeit übertragen werden kann. Die Optimierung gliedert sich in zwei Hauptprobleme:

Zum einen muss die Architektur des optimierten Systems erarbeitet werden. Die gewählten Verfahren werden zu einem Basissystem integriert, das durch zusätzliche Funktionen erweitert wird. Diese Funktionen sollen die dynamische Anpassung des Systems realisieren, um es optimal auf das Eingangssignal einzustellen. Am Ende werden von dem Gesamtsystem Ausgaben erwartet, die für das automatische Beattracking notwendige Informationen liefern. So sollen dem Beattracker neben der Zeitinformation der Onsets auch Werte über Intensität und Relevanz im Hinblick einer Takterkennung mitgeliefert werden.

Das zweite Problem ist die sinnvolle, effektive und effiziente Parametrisierung der integrierten Verfahren. Die Optimierung eines Parameters hinsichtlich eines bestimmten Kriteriums wirkt sich häufig nachteilig auf andere Eigenschaften des System aus. Es gilt also, den für die Anwendung günstigsten Kompromiss zu finden.

Zur Bearbeitung der Aufgaben wurden sämtliche Algorithmen in MATLAB implementiert, um sich einerseits ein von den Beschreibungen der Autoren unabhängiges Bild über die Fähigkeiten der Verfahren zu verschaffen und andererseits eine eventuelle Implementierung des Gesamtsystems als Software vorzubereiten. Die Arbeit ist wie folgt gegliedert:

In Kapitel 2 werden die für die Arbeit wichtigen Grundlagen besprochen. Der erste Abschnitt behandelt hierbei die Erörterung des Onset-Begriffs im Kontext des Beattracking. Der zweite Teil fasst die relevanten signaltheoretischen Grundlagen zusammen und führt in die verwendete Notation ein.

Kapitel 3 beschreibt die untersuchten Verfahren und bewertet ihre allgemeinen Fähigkeiten. Weiterhin wird ein neuer Ansatz zur Optimierung der Verfahren vorgestellt und dessen Nutzen diskutiert. Auf den Ergebnissen aus diesem Kapitel basiert die in Kapitel 4 behandelte Entwicklung eines robusten Algorithmus zur automatischen Onset-Erkennung. Hier werden zunächst die Fähigkeiten der vorgestellten Verfahren verglichen und der vielversprechendste Ansatz als Ausgangspunkt der nachfolgenden Optimierung gewählt. Das Kapitel stellt nicht nur den optimierten Algorithmus als Ergebnis dar, sondern gibt Aufschluss über die einzelnen Optimierungsschritte. Es wird weiterhin das Ausgabeformat des entwickelten Systems definiert.

Kapitel 5 ist das Evaluierungskapitel. Der erste Teil erörtert die Evaluierungsmethodik mit der Entwicklung von Evaluierungskriterien und der Auswahl des Testmaterials. Im zweiten Teil folgt die Evaluierung selbst mit Durchführung und Auswertung der Tests. Im Anschluss daran wird ein kurzes Fazit zur angewandten Evaluierungsmethode gezogen. Den Abschluss der Arbeit bildet Kapitel 6 mit Zusammenfassung und Ausblick.

Kapitel 2

Grundlagen

2.1 Der Onset-Begriff

Vor der Untersuchung und Entwicklung von Verfahren zur Erkennung von Onsets in Musiksignalen, soll zunächst der Begriff des Onsets im Kontext dieser Arbeit erläutert werden. Ein musikalischer Onset bezeichnet im Allgemeinen den Beginn eines musikalischen Klangereignisses. Welche Zeit damit genau gemeint ist, kann man sich anhand eines so genannten *ADSR-Modells* klarmachen. Hier wird ein Klang durch einen Hüllkurvenverlauf dargestellt und in die vier aufeinander folgenden Phasen *Einschwingen* oder *Anschlag (Attack)*, *Dämpfung (Decay)*, *Haltephase (Sustain)* und *Ausschwingen (Release)* eingeteilt (vergleiche Abbildung 2.1).

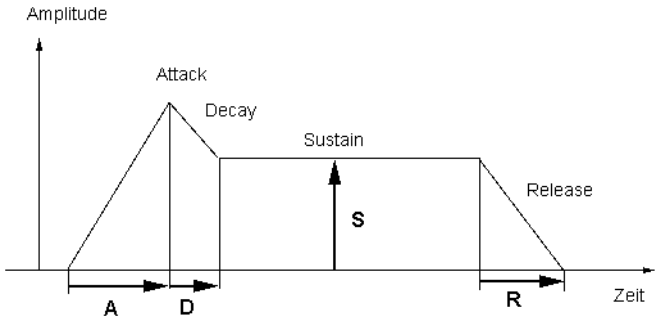

Abbildung 2.1: ADSR-Modell

In Abhängigkeit von der Erzeugung haben unterschiedliche Klänge verschieden starke Ausprägungen der einzelnen Phasen. So findet man beim Klavier beispielsweise einen recht markanten Anschlag, während es praktisch keine Haltephase gibt, da der Ton nach dem Anschlagen sofort beginnt zu verklingen, so dass Dämpfung und Ausschwingen quasi ineinander übergehen. Im Gegensatz dazu enthält der Klang einer Orgel nahezu keine Dämpfungs- und Ausschwingphase.

Der Orgelton schwingt ein, bleibt relativ konstant auf dem Haltepegel und würde bei Loslassen der Taste auch sofort verstummen, wenn der umgebene Raum frei von Reflexionen wäre. Ein Onset bezeichnet die Einschwingphase, das ist die Zeitspanne vom absoluten physikalischen Beginn des Klangereignisses bis zum Erreichen des Maximums in der Hüllkurve. Diese kann impulsartig und kaum wahrnehmbar ausfallen, wie beim Schlag einer Trommel, oder einen langen gradienten Anstieg aufweisen, wie beispielsweise beim Streichen einer Geigensaite. Die Dauer des Einschwingvorgangs variiert stark in Abhängigkeit des Klangerzeugers. Bei natürlichen Musikinstrumenten liegt diese im Bereich von 5 bis 800 ms.

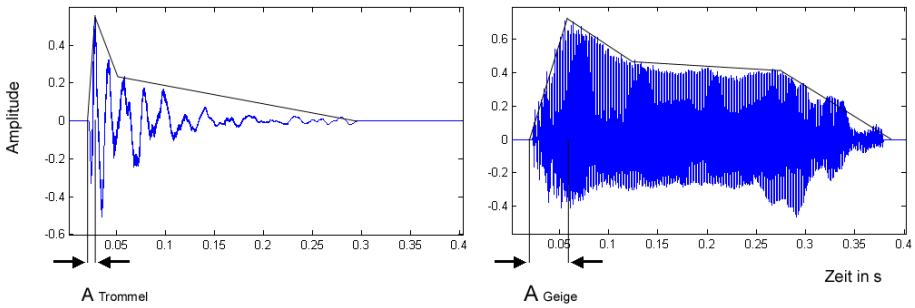

Abbildung 2.2: Vergleich der Einschwingvorgänge eines Trommelschlags $A_{Trommel}$ (links) und Geigentons A_{Geige} (rechts)

2.1.1 Onset im Kontext des Beattracking

Wie bereits erwähnt, wird eine automatische Onset-Erkennung für verschiedene Anwendungen zur Informationsverarbeitung in Musiksignalen benötigt. Veröffentlichungen der jeweiligen Themengebiete verwenden gleichermaßen den Begriff Onset als Eintritt eines musikalischen Ereignisses, verarbeiten jedoch einen speziellen diskreten Zeitpunkt. Welcher Zeitpunkt der oben genannten Zeitspanne des Einschwingvorgangs einem Onset entspricht, wird meist nicht weiter behandelt. Für die Entwicklung von Algorithmen fehlt somit eine wichtige Information, die mitunter zu falschen Ansätzen und dadurch zu unbefriedigenden Ergebnissen führen kann. Bei genauerer Betrachtung fällt auf, dass verschiedene Problemstellungen unter Onset im Detail ganz unterschiedliche Information bzw. Zeitpunkte verstehen und der jeweilige Ansatz von genau diesem Detail abhängt. So benötigt beispielsweise eine Onset-Erkennung für einen automatischen Transkriptionsalgorithmus nach MIDI genau den physikalischen Beginn einer Note, da Steuersignale generiert werden sollen, die neben der Tonhöhe exakt den zeitlichen Verlauf, das heißt Anfang, Ende und damit die Dauer einer Note rekonstruieren müssen.

Im Kontext von Beattracking-Systemen ist der Zeitpunkt nicht so exakt definierbar. Ein automatischer Beattracker errechnet den Beat aus der diskreten

Folge der Onset-Zeitpunkte des Musiksignals. Wichtig ist, dass diese diskreten Zeiten möglichst gut die in dem Stück enthaltenen rhythmischen Informationen repräsentieren. Orientiert man sich hier an der menschlichen Wahrnehmung, ist für die rhythmische Empfindung meist nicht der physikalische Beginn eines Klangereignisses ausschlaggebend. Psychoakustische Untersuchungen zur Rhythmuswahrnehmung in [ZF99] zeigen beispielsweise, dass sich die wahrgenommene Einsatzzeit eines Klanges innerhalb eines rhythmischen Gefüges je nach Dauer und Gestalt der Einschwingphase und des weiteren Verlaufs gegenüber dem physikalischen Beginn nach hinten verschieben kann. Weiterhin wird ein Modell beschrieben, das die rhythmische Wahrnehmung auf den Verlauf der empfundenen Lautheit des Signals zurückführt. Hiernach ist ein rhythmisches Ereignis für Sprache und Musik durch die Maxima des Verlaufs der wahrgenommenen Lautheit charakterisiert. In [Sch78] wird festgestellt, dass die Verschiebung zwischen wahrgenommenen und physikalischen Onset in Abhängigkeit der Form des Einschwingvorgangs der Klangereignisse stark variiert und unter bestimmten Bedingungen bis zu 80 ms betragen kann. Untersuchungen in [JV82] begründen die unterschiedlichen Verschiebungszeiten dadurch, dass ein Klang vor seiner Wahrnehmung eine bestimmte Amplitudenschwelle überschritten haben muss. Diese liegt im Bereich von 6-15 dB unter der maximalen Amplitude des Klanges.

Es existieren eine Reihe von Untersuchungen, die sich mit der exakten Lokalisation des wahrgenommenen Onset-Zeitpunktes, bzw. mit seiner Verschiebung vom physikalischen Beginn eines Klangereignisses beschäftigen. Hierbei werden jeweils die hinsichtlich eines Rhythmus wahrgenommenen Onset-Zeiten in Sprach- oder Musiksignalen bei einer Gruppe von Testpersonen gemessen und die Mittelung der resultierenden Zeiten als Referenz-Onsets festgelegt. Hierauf folgt die Bildung eines Modells der rhythmischen Wahrnehmung mit dem Ziel, eine konkrete Berechnungsvorschrift für die Verschiebungszeit zu finden. Viele Untersuchungen stammen aus dem Bereich der Sprachsignalverarbeitung und berechnen die Verschiebung anhand der Längen unterschiedlicher Segmente rhythmisch gesprochener Silben. Zu nennen sind die so genannten *P-Center-Modelle* in [Mar81] und [BPM87]. Bei letzterem wird zusätzlich zu den Längen noch der Einfluss der spezifischen Lautheit (siehe [ZF99]) einbezogen. Für Details der einzelnen Verfahren siehe obige Quellen.

Die Ergebnisse dieser Modellierungen weisen zum Teil hohe Übereinstimmungen mit den Referenz-Zeiten auf. Da diese Verfahren jedoch auf der Basis von Sprachsegmenten wie beispielsweise Konsonanten arbeiten, ist die direkte Übertragung auf Musiksignale nicht möglich. In [Jan95] wird weiterhin hinzugefügt, dass Verfahren, bei denen Sprachsilbensegmente betrachtet werden, ein Problem für die maschinelle Sprachverarbeitung darstellen, da diese auf den menschlichen Fähigkeiten zur Segmentierung von Sprachelementen beruhen. Es wird hier deshalb nach Möglichkeiten gesucht, die nötigen Informationen aus den Rohdaten des Sprachsignals zu extrahieren. Dazu werden unter anderem der Hüllkurvenverlauf des Signals sowie dessen erste und zweite Ableitung berechnet und die

Referenz-Onsets mit zeitlich korrespondierenden Maxima der jeweiligen Verläufe verglichen. Eine erstaunlich gute Übereinstimmung ergibt sich für die Maxima der zweiten Ableitung. Dieser Ansatz wird zwar auf Sprachsignale angewendet, ist jedoch sehr einfach und für diese Arbeit aufgrund der leichten Integrierbarkeit nahe liegend.

Andere Modelle sind allgemeiner und beziehen sich auf Sprache und Musik. Das Modell des *Einsatzzeitpunktes (EZP)* nach Schütte [Sch78] zum Beispiel bildet die Verschiebung des wahrgenommenen Onsets basierend auf psychoakustischen Untersuchungen durch eine gewisse Trägheit des Gehörs bei der Verarbeitung von Schallereignissen nach. Diese Trägheit wird hierbei mit Hilfe eines bestimmten Tiefpassfilters realisiert und der Onset als derjenige Zeitpunkt festgelegt, nachdem der modellierte träge Anstieg am Beginn eines Klangereignisses einen gewissen Schwellenwert erreicht hat. Die Schwelle errechnet sich dabei aus dem Maximum des Hüllkurvenverlaufes des Ereignisses und einem heuristisch optimierten konstanten Faktor. Die Ergebnisse sind für Musiksignale sehr vielversprechend. Ein Versuch in [Jan95], dieses Verfahren an Sprachsignalen anzuwenden schlug jedoch fehl.

2.1.2 Zeitliche Auflösung der Onset-Wahrnehmung

Psychoakustische Untersuchungen ergeben, dass das Gehör bei der Wahrnehmung von Klangereignissen eine gewisse Trägheit aufweist, so dass aufeinander folgende Klangeinsätze nur dann unterschieden werden können, wenn sie einen bestimmten zeitlichen Abstand einhalten. Die Ergebnisse in der Literatur schwanken je nach Ziel und Durchführung der Untersuchungen im Bereich von 8-50 ms, manchmal auch mehr. Der Abstand wird in vielen Zusammenhängen als *gerade noch wahrnehmbarer Unterschied (JND)*[1] bezeichnet. Nach [ME49] wird der versetzte Beginn zweier identischer Klänge als gleichzeitiger Einsatz wahrgenommen, wenn der Abstand weniger als 8 ms beträgt. Wird der Abstand der Einsatzzeiten in einer Folge diskreter Klangereignisse immer weiter verkürzt, so beträgt die Grenze bis der Hörer die Folge als kontinuierliches Klangsignal auffasst 50 ms[Han89]. In [AF93] wird an einer rhythmisch angeordneten Klangfolge psychoakustisch untersucht, wie weit ein einzelnes Klangereignis verschoben werden kann, bis der Hörer eine Störung des Rhythmus empfindet. Der durchschnittliche JND liegt hier bei 10 ms für Klänge bis zu 240 ms Länge und bei ungefähr 5 Prozent der Klangdauer für längere Klangereignisse und ist weiterhin unabhängig von der musikalischen Vorbildung der Versuchspersonen.

[1] engl. Just Notable Difference

2.2 Signaltheoretische Grundlagen

Dieser Abschnitt beschäftigt sich mit den signaltheoretischen Grundlagen, die für das Verständnis der Signalverarbeitung in dieser Arbeit hilfreich sind. Es werden hierbei sowohl relevante Definitionen geliefert als auch nachfolgende Notationen eingeführt.

2.2.1 Das digitale Audiosignal

Die Musiksignale, die in dieser Arbeit betrachtet werden, befinden sich in digitaler Form. Je nach Genre und Produktionstechnik entstanden diese entweder direkt durch synthetische Klangerzeugung, zum Beispiel durch digitale Synthesizer oder mittels Digitalisierung aufgenommener Analogsignale. In den folgenden Betrachtungen geht es daher fast ausschließlich um digitale Signale.

2.2.1.1 Digitalisierung

Analoge Signale weisen einen kontinuierlichen Zeit- und Wertebereich auf. Bei der Digitalisierung wird das kontinuierliche analoge Signal $x(t)$ mit einer bestimmten Frequenz, der Abtast-Frequenz f_s, abgetastet und dadurch in ein zeitdiskretes Signal $x[n]$ umgewandelt. Der kontinuierliche Wertebereich des analogen Signals wird in eine feste Anzahl diskreter Amplitudenstufen unterteilt, jedem Abtastwert des zeitdiskreten Signals ein diskreter Amplitudenwert zugeordnet und somit das zeit- und wertdiskrete Digitalsignal gebildet (vergleiche Abbildung 2.3).

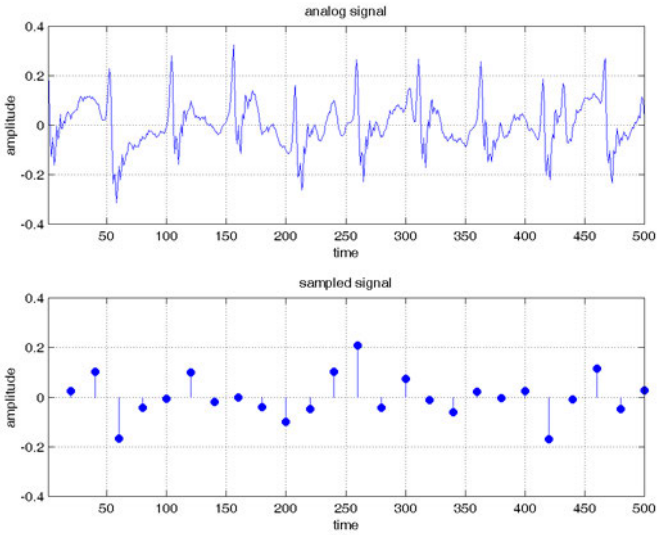

Abbildung 2.3: Vergleichende Darstellung eines kontinuierlichen (oben) und eines abgetasteten Signalverlaufs (unten)

2.2.1.2 Signalenergie

Für das diskrete Signal $x[n]$ ist die Signalenergie innerhalb eines Intervalls $[0, N]$ definiert durch:

$$E = \sum_{n=0}^{N} |x[n]|^2 \,. \tag{2.1}$$

2.2.2 Filter

Neben der Analyse, Interpretation und Darstellung von Signalen, besteht die Aufgabe der digitalen Signalverarbeitung vor allem in deren Manipulation, d.h. den Verlauf der Signalfunktion gezielt zu verändern. Zu den wichtigsten Möglichkeiten zählen die Abschwächung oder Verstärkung bestimmter Frequenzen oder Frequenzbereiche. Dies wird im allgemeinen als Filterung bezeichnet und die Strukturen als digitale Filter. Man unterscheidet prinzipiell zwei Arten von Filterstrukturen:

- Transversalfilter oder nichtrekursive Filter (FIR-Filter)[2],
- Rekursive Filter (IIR-Filter)[3].

Beide unterscheiden sich grundsätzlich in ihrer Struktur und darin, dass FIR-Filter im Gegensatz zu den IIR-Filtern eine endlich lange Impulsantwort aufweisen. Für diese Arbeit sind lediglich FIR-Filter von Interesse.

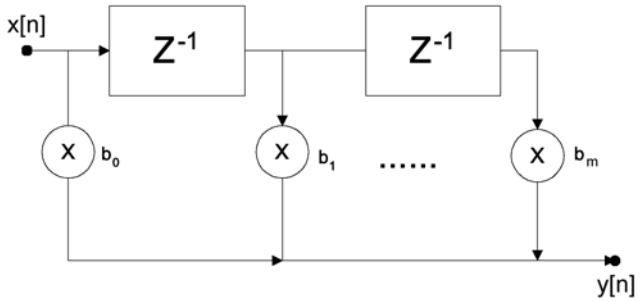

Abbildung 2.4: Struktur eines FIR Filters

FIR-Filter zeichnen sich durch ihre iterative Struktur aus. Abbildung 2.4 zeigt die typische Struktur eines Transversalfilters. z^{-1} repräsentiert eine Verzögerung um einen Abtastwert. Vorangegangene Abtastwerte werden gewichtet zu dem

[2]engl. Finite Impulse Response
[3]engl. Infinite Impulse Response

aktuellen hinzuaddiert. Mathematisch beschreibt man den Vorgang folgendermaßen:

$$y[n] = \sum_{m=0}^{M} b_m x[n-m] = x[n] * b[n]. \tag{2.2}$$

Der Operator $*$ wird als Faltung bezeichnet. Nichtrekursive Filter haben eine endliche Impulsantwort, die exakt den Filterkoeffizienten entspricht.

2.2.3 Fouriertransformation

Die einfachste Darstellung eines Signals ist der Verlauf der Amplitude über der Zeit, das *Zeitsignal*. Dieser steht eine weitere, in der Signalverarbeitung sehr wichtige Repräsentation gegenüber, in der das Signal als Linearkombination von Sinus-Funktionen (oder allgemeiner komplexen Exponentialfunktionen) beschrieben wird. Letzteres wird Darstellung im *Frequenzbereich* oder auch *Spektrum* genannt. Das mathematische Verfahren zur Berechnung des Spektrums eines Signals ist die *Fouriertransformation*.

2.2.3.1 Diskrete Fouriertransformation

Für ein diskretes Signal $x[n]$ ist die Fouriertranformation folgendermaßen definiert:

$$X[e^{j\omega}] = \sum_{n=-\infty}^{\infty} x[n] e^{-jn\omega}. \tag{2.3}$$

f: Frequenz
$\omega = 2\pi \frac{f}{f_s}$: Normierte Kreisfrequenz

Die Fouriertransformierte eines diskreten Signals ist immer periodisch mit der Abtastfrequenz und, wie der Gleichung zu entnehmen, kontinuierlich. In der rechnerbasierten Signalverarbeitung sind diese beiden Fakten nicht erwünscht. Die Signale müssen hier sowohl diskreter und begrenzter Natur sein.

Diese Probleme werden durch die *Diskrete Fouriertransformation (DFT)* gelöst. Für ein Signal $x[n]$ mit einer Länge von N Abtastwerten ist die N-Punkt-DFT definiert als:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}. \tag{2.4}$$

Die DFT liefert ein komplexes Ergebnis, das sich auch in Betrag und Phase darstellen lässt. Die Umrechnung von X[k] in Betrag $A[k]$ und Phase $\Phi[k]$ erfolgt nach:

$$A[k] = \sqrt{\Re(X[k])^2 + \Im(X[k])^2}, \quad \Phi[k] = \arctan\left(\frac{\Im(X[k])}{\Re(X[k])}\right). \tag{2.5}$$

Das diskrete Spektrum X[k] kann als die an N äquidistanten Stützstellen abgetastete Fouriertransformation des Signals interpretiert werden. Diese Stützstellen werden oft auch als Frequenz-Bins bezeichnet. Für diskrete reellwertige Signale, wie zum Beispiel Audiosignale, ist das diskrete Spektrum symmetrisch um den mittleren Frequenz-Bin $k = N/2$. Dieser korrespondiert mit der höchsten Frequenz, die von der DFT dargestellt werden kann, der so genannten *Nyquist-Frequenz*[4]. Allgemein ist die von Bin k repräsentierte Frequenz gegeben durch:

$$f[k] = \frac{kf_s}{N}, \qquad 0 \leq k \leq N/2. \tag{2.6}$$

Aus dieser Gleichung kann gesehen werden, dass die Nyquist-Frequenz gleich der halben Abtast-Frequenz ist.

Die DFT weist weiterhin folgende Eigenschaften auf:

- Linearität: $DFT(ax[n] + by[n]) = aX[k] + bY[k]$,
- Einer Multiplikation im Zeitbereich entspricht einer Faltung im Frequenzbereich und umgekehrt.

2.2.3.2 Kurzzeit Fouriertransformation

Zur Berechnung der DFT existiert ein sehr effizienter Algorithmus falls die Länge N eine Potenz von zwei ist, die *Schnelle Fouriertransformation (FFT)*[5]. In dieser Arbeit werden ausschließlich digitale Audiosignale betrachtet. Auch wenn diese eine begrenzte Länge aufweisen, ist eine direkte Anwendung der DFT nicht sinnvoll, da der Berechnungsaufwand mit steigendem N exponentiell anwächst. Für große N ist dies selbst mit Verwendung der FFT völlig unpraktikabel.

Der größte Nachteil der DFT ist jedoch der Verlust sämtlicher Zeitinformationen. Die DFT liefert zusammengefasst die spektralen Eigenschaften des gesamten Signals. Diese verändern sich im Verlauf der Zeit in der Regel jedoch ständig, was bei der DFT nicht berücksichtigt wird.

Zur Lösung der genannten Probleme wird das Signal nicht als Ganzes, sondern blockweise transformiert. Man wählt hierbei eine Blockgöße mit einem akzeptablen Berechnungsaufwand, für den man weiterhin annimmt, dass die spektrale Charakteristik in diesem einigermaßen stationär bleibt. Dieser Ausschnitt aus dem Signal wird auch Fenster genannt und bedeutet formaler betrachtet eine Multiplikation des Signalausschnitts mit einer Fensterfunktion $w[n]$. Das Fenster der Länge L wird nun entlang des Signals mit einer bestimmten Schrittweite H verschoben und in jedem Schritt die FFT berechnet. H muss dabei nicht zwangsweise die gleiche Länge aufweisen wie die Fenstergöße. Verwendet man kleinere Schrittweiten, überlappen sich aufeinander folgende Blöcke, was oft in Form eines

[4] nach Nyquist, Entdecker des Abtasttheorems
[5] engl. Fast Fourier Transform

Überlappungsfaktors *(engl. overlap)* angegeben wird. Das gefensterte Signal kann nun notiert werden als:

$$x_i[n] = x[iH + n]\ w[n], \qquad 0 \leq n \leq L - 1. \tag{2.7}$$

i: Laufindex des jeweiligen Fensters (Bereich hängt von der Signallänge ab)

Diese blockweise Transformation wird *Kurzzeit-Fouriertransformation (STFT)*[6] genannt (in diesem Zusammenhang vielleicht genauer Kurzzeit-DFT) und stellt die zeitabhängige Fouriertransformation durch Berechnung der DFT jedes Fensters dar. Unter der Verwendung der obigen Notation ist die STFT gegeben durch:

$$X[i,k] = \sum_{n=0}^{L-1} x[iH + n]w[n]e^{-j\frac{2\pi}{N}kn}, \qquad 0 \leq k \leq N - 1. \tag{2.8}$$

N: Länge der DFT
L: Länge des Fensters
H: Schrittweite

Offensichtlich bestimmen die Parameter Fenstergröße und Schrittweite sowohl Zeit- als auch Frequenzauflösung. Hierbei gilt es einen Kompromiss zu finden, da sich eine höhere Zeitauflösung immer zu Ungunsten der Frequenzauflösung und umgekehrt auswirkt. Eine anschauliche und oft angewandte Visualisierung der STFT ist das Spektrogramm. Wie in 2.5 zu sehen, wird hier die Frequenz über der Zeit aufgetragen und die jeweilige Energie des Betrags logarithmisch durch eine bestimmte Farbgebung angezeigt.

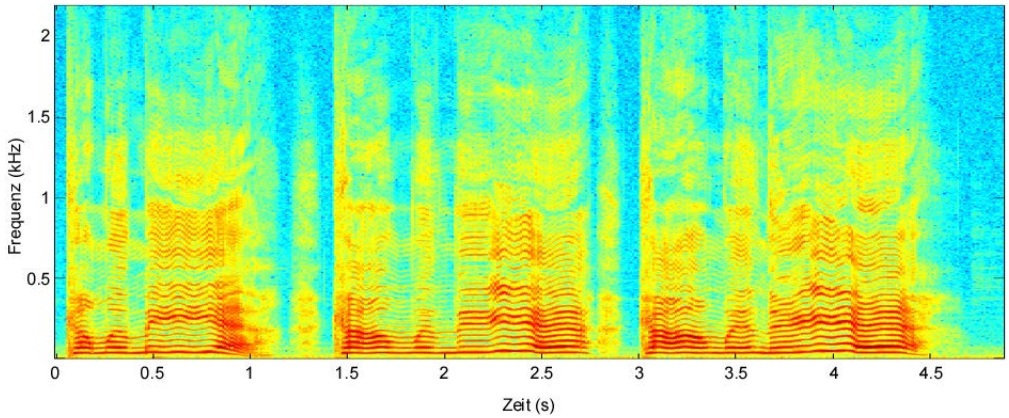

Abbildung 2.5: Spektrogrammdarstellung eines Audiosignals; die roten Bereiche besitzen die höchste Amplitude

[6] engl. Short Time Fourier Transform

2.2.3.3 Fensterung

Wie bereits erwähnt, steckt hinter der Fensterung eine Multiplikation mit einer Fensterfunktion. Im einfachsten Fall ist dies ein *Rechteckfenster*, bei dem jeder Wert des Signals innerhalb des Fensters mit eins gewichtet wird. Da eine Multiplikation im Zeitbereich einer Faltung im Frequenzbereich entspricht, wird das Spektrum des Signalausschnitts mit dem Spektrum der Fensterfunktion gefaltet.

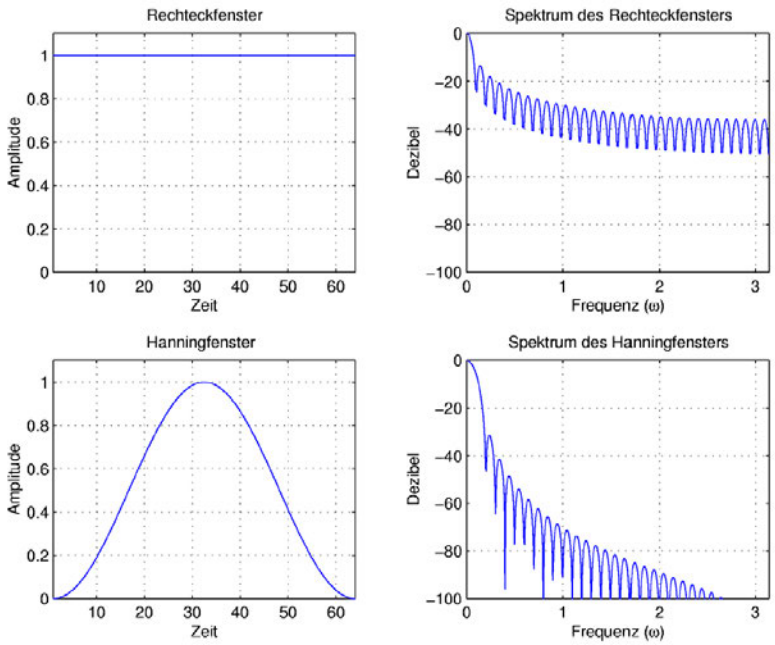

Abbildung 2.6: Zeitverlauf und Spektrum von Rechteckfenster(oben) und Hanningfenster(unten)

Das Spektrum des Rechteckfensters ist eine si-Funktion (vergleiche Abbildung 2.6) und verursacht gefaltet mit dem Signalspektrum starke Verfälschungen der hohen Frequenzen. Dies ist im Prinzip bei allen Fensterfunktionen der Fall, beim Rechteckfenster jedoch besonders ausgeprägt. Um eine geeignete Fensterfunktion auszuwählen, werden drei Parameter des Spektrums einer Fensterfunktion betrachtet:

- Breite der Hauptkeule,
- Pegel der Nebenkeulen,
- Pegelgenauigkeit.

Grundsätzlich ist man bestrebt, die Breite der Hauptkeule sowie die Höhe der Nebenkeule zu minimieren und die Pegelgenauigkeit zu maximieren. Dieses

Ziel lässt sich nicht erreichen, da die Optimierung eines Parameters eine Verschlechterung der anderen zur Folge hat. Einen guten Kompromiss stellt das in der Audiosignalverarbeitung häufig verwendete *Hanningfenster* (siehe Abbidung 2.6) dar. Für die Länge N lässt es sich wie folgt berechnen:

$$Hann_N[n] = \frac{1}{2}\left[1 - \cos\left(2\pi\frac{n}{N-1}\right)\right], \qquad n = 0, 1, 2, ...N-1. \tag{2.9}$$

2.2.4 Stochastische Größen

Natürliche Signale, wie zum Beispiel Audiosignale, lassen sich sowohl analytisch wie auch statistisch beschreiben. Da Audiosignale im Allgemeinen aber für eine analytische Beschreibung zu komplex sind und zudem oft über einen Rauschanteil verfügen, hat sich die statistische Beschreibung als sehr nützlich erwiesen. Das Signal wird hierbei als stochastischer Prozess $\mathcal{X}$ betrachtet. $P(\mathcal{X} = x)$ gibt die Wahrscheinlichkeit an, dass die Realisierung $x[n]$ des Prozesses $\mathcal{X}$ zum Zeitpunkt n den Wert x annimmt. Im Folgenden wird eine Auswahl wichtiger Größen erläutert.

2.2.4.1 Verteilungs- und Verteilungsdichtefunktion

Die Verteilungs- und Verteilungsdichtefunktion beschreiben jeweils über einer Menge von Realisierungen die Verteilung der Auftretenswahrscheinlichkeiten der einzelnen vorkommenden Werte. Die Verteilungsfunktion bei diskreten Prozessen ist wie folgt definiert:

$$F[x] = \sum_{x' \leq x} P[\mathcal{X} = x']. \tag{2.10}$$

Die Verteilungsdichtefunktion ergibt sich dann folgendermaßen:

$$p_{\mathcal{X}}[x] = P[\mathcal{X} = x]. \tag{2.11}$$

Die Verteilungsdichtefunktion besitzt weiterhin folgende wichtige Eigenschaft:

$$\sum_x p_x[x] = 1 \qquad \forall x \in \Omega. \tag{2.12}$$

Ω: Menge aller möglichen Realisierungen

2.2.4.2 Erwartungswert

Sei $\mathcal{X}$ eine diskrete Zufallsgröße mit den Realisierungen x und der Wahrscheinlichkeitsdichtefunktion $p_x[x]$. Dann ist der Erwartungswert einer Funktion $g[\mathcal{X}]$

wie folgt definiert:

$$E[g[\mathcal{X}]] = \sum_x g[x]p_x[x] \qquad \forall x \in \Omega. \tag{2.13}$$

Aus dieser allgemeinen Definition leiten sich weitere feste Größen der Wahrscheinlichkeitsrechnung ab. So ist der Mittelwert einer Zufallsgröße als

$$\mu = E[\mathcal{X}] \tag{2.14}$$

definiert.

2.2.4.3 Varianz

Die Varianz wird in der Signalverarbeitung oft als Fehlerkriterium (siehe Abschnitt 2.2.6) verwendet

$$\sigma^2 = E[(\mathcal{X} - \mu)^2]. \tag{2.15}$$

Die Varianz lässt sich auch der Gruppe der Momente zuordnen. Dabei handelt es sich um eine Klasse von Erwartungswerten, die als Funktion $g[x]$ eine Potenzfunktion $g_n[x] = x^n$ verwenden. Das Moment n-ter Ordnung lautet dann:

$$E[\mathcal{X}^n] = \sum_x x^n p_x(x) \qquad \forall x \in \Omega, \tag{2.16}$$

das zentrale (mittelwertbefreite) Moment entsprechend:

$$E[(\mathcal{X} - \mu)^n] = \sum_x (x - \mu)^n p_x(x) \qquad \forall x \in \Omega. \tag{2.17}$$

Bei der Varianz handelt sich also um ein zentrales Moment zweiter Ordnung. Das ist auch der Grund, dass statistische Verfahren, die auf der Varianz beruhen, als Verfahren der Statistik zweiter Ordnung bezeichnet werden.

2.2.5 Auto- und Kreuzkorrelation

Die Auto- und Kreuzkorrelation wird in der Signalverarbeitung sehr häufig angewendet und zeigen statistische Abhängigkeiten zwischen Signalen oder Signalabschnitten auf. Da im Weiteren nur diskrete Korrelationsfunktionen von Bedeutung sind, soll auf die kontinuierlichen Pendants nicht eingegangen werden.

Die *Autokorrelationsfunktion (AKF)* eines Signals gibt die statistische Abhängigkeit zweiter Ordnung der Abtastwerte einer Folge voneinander an. Sie wird verwendet, um Ähnlichkeiten des Signals mit sich selbst, wie zum Beispiel Periodizität, zu entdecken und ist definiert durch

$$r_{xx}[m] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=-N/2}^{N/2} x[n]x[n+m]. \tag{2.18}$$

Eine Schätzung der AKF ist

$$r_{xx}[m] = \frac{1}{N} \sum_{n=-N/2}^{N/2} x[n]x[n+m]. \tag{2.19}$$

Die AKF wird oft normiert angegeben:

$$r_{0,xx}[m] = \frac{r_{xx}[m]}{r_{xx}[0]}. \tag{2.20}$$

Somit hat die normierte AKF an dieser Stelle den Wert 1. $r_{xx}[0]$ ist gleich der Energie des Signals. Wichtige Eigenschaften der AKF sind:

- Die AKF ist symmetrisch: $r_{xx}[-m] = r_{xx}[m]$.
- Es gilt $|r_{xx}[m]| \leq r_{xx}[0]$. Somit kann der Betrag der normierten Autokorrelationsfunktion keinen Wert größer 1 annehmen.

Die *Kreuzkorrelationsfunktion (KKF)* betrachtet nun die statistischen Abhängigkeiten zweier verschiedener Signale x[n] und y[n]. Ihre Schätzung ist gegeben durch

$$r_{xy}[m] = \frac{1}{N} \sum_{n=-N/2}^{N/2} x[n]y[n+m]. \tag{2.21}$$

2.2.6 Lineare Prädiktion

Lineare Prädiktion findet besonders in der Nachrichtentechnik bei der Kodierung von Sprach- und Bildsignalen Anwendung. Ziel der Linearen Prädiktion ist es, ein Signal $x[n]$ aus seinen vorangegangenen Werten $x[n-1], x[n-2], \ldots, x[n-M]$ möglichst gut vorherzusagen. Voraussetzung dafür ist, dass dem Signal x statistische Abhängigkeiten inne wohnen, sonst ist eine Vorhersage gar nicht möglich. Abbildung 2.7 zeigt das Blockschaltbild eines linearen Prädiktors zweiter Ordnung.

Der Prädiktor besitzt im Wesentlichen die Struktur eines nichtrekursiven Filters. Der Schätzwert, auch Prädiktionssignal genannt, eines Prädiktors der Ordnung M kann für das Eingangssignal $x[n]$ wie folgt formuliert werden:

$$\hat{x}[n] = \sum_{m=1}^{M} b_m x[n-m]. \tag{2.22}$$

Der Schätzfehler, oder auch Prädiktionsfehler, der offensichtlich von den Koeffizienten b_m abhängt, ist dann:

$$e[n] = x[n] - \hat{x}[n]. \tag{2.23}$$

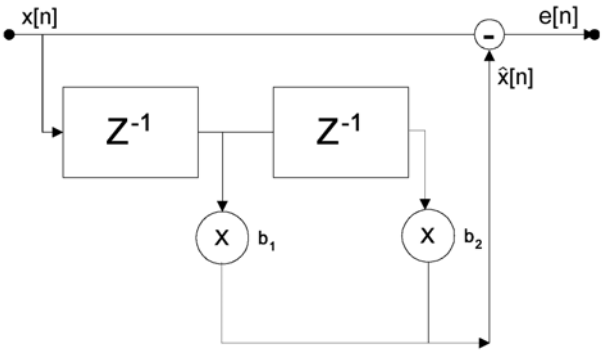

Abbildung 2.7: Struktur eines linearen Prädiktors 2.Ord.

2.2.6.1 Minimierung des Prädiktionsfehlers

Aufgabe des Prädiktors ist es, den Prädiktionsfehler zu minimieren. Dazu müssen die Prädiktorkoeffizienten optimiert werden. Bei der Linearen Prädiktion wird die Varianz des Prädiktionsfehlers als Fehlermaß genommen. Das Optimierungkriterium stellt sich nun so dar:

$$\sigma_e^2 = E[\{\mathcal{X} - \hat{\mathcal{X}}\}^2] = \frac{1}{N} \sum_{\forall n} \left[x[n] - \sum_{m=1}^{K} b_m x[n-m] \right]^2 = min. \tag{2.24}$$

Dazu setzt man die Ableitung nach den Prädiktorkoeffizienten b_l zu Null

$$\frac{\partial}{\partial b_l} \sigma_d^2 = \frac{2}{N} \sum_{\forall n} \left[x[n] - \sum_{m=1}^{M} b_m x[n-m] \right] x[n-l] = 0. \tag{2.25}$$

Nach Umstellen und vertauschen der Summation erhält man

$$\frac{1}{N} \sum_{\forall n} x[n]x[n-l] = \frac{1}{N} \sum_{m=1}^{M} b_m \sum_{\forall n} x[n-m]x[n-l]. \tag{2.26}$$

Der Term $\frac{1}{N} \sum_{\forall n} x[n]x[n-l]$ entspricht der Autokorrelationsfunktion $r_{xx}[l]$ des Signals x. Dementsprechend kann man das Ergebnis zusammenfassen:

$$r_{xx}[l] = \sum_{m=1}^{M} b_m r_{xx}[l-m]. \tag{2.27}$$

Die optimalen Filterkoeffizienten berechnen sich also aus einem Gleichungssystem der Autokorrelationsfunktion des Signals x.

Kapitel 3

Verfahren der Onset-Erkennung

Es existieren verschiedene Verfahren zur Extraktion der Onset-Zeiten aus einem unbekannten Musiksignal. Viele Autoren beschäftigten sich mit der Entwicklung effizienter Algorithmen zur Onset-Erkennung unter anderem während der Implementierung von Beattracking-Systemen. Nach Darstellung des allgemeinen Ablaufschemas und einer kurzen Literaturübersicht wird eine Auswahl dieser vorgestellt und der jeweilige Ansatz erläutert. Unterteilt wird hierbei in Algorithmen im Zeit- und Frequenzbereich. Zur grafischen Veranschaulichung von Signalen wird zum Teil von der Abtastrate abstrahiert und der Amplitudenverlauf über der Zeit in Sekunden aufgetragen. Es handelt sich aber ausschließlich um diskrete Signale.

Im Zuge der Erläuterungen wird auf Stärken und allgemeine Probleme der Algorithmen eingegangen, die in Kapitel 4 zur Entwicklung eines robusten Algorithmus als Diskussionsgrundlage dienen sollen. Der letzte Abschnitt dieses Kapitels untersucht einen neuen Ansatz zur Optimierung bestehender Verfahren. Hierbei soll das Prinzip der linearen Prädiktion auf die Problematik der Onset-Erkennung übertragen werden.

3.1 Allgemeiner Ablauf

Die Verfahren basieren auf einem gemeinsamen Grundschema wie in Abbildung 3.1 illustriert. Das digitale Musiksignal wird hierbei entsprechend eines Algorithmus in eine Zwischenfunktion, im Weiteren als *Onsetkurve* bezeichnet, transformiert, deren lokale Maxima potentielle Kandidaten für Onsets darstellen. Anschließend folgt die Auswahl der herausstechendsten Maxima, das so genannte Peakpicking. Wie später noch erwähnt wird, hat Letzteres großen Einfluss auf die Qualität der Ergebnisse. Ausgangsseitig erhält man einen Vektor mit festgelegten Onset-Zeiten.

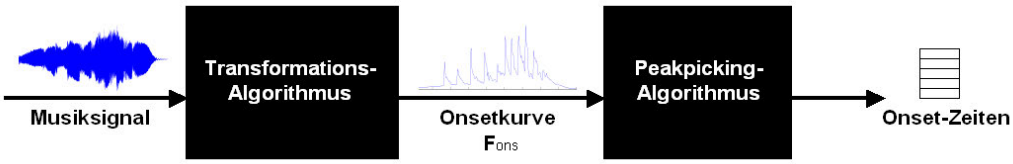

Abbildung 3.1: Grundschema der Onset-Erkennung

Diese Darstellung des Schemas ist sehr vereinfacht und es wäre möglich Pre- oder Postprocessing-Zwischenschritte hinzuzufügen. Da diese Processing-Schritte jedoch spezifisch an die jeweiligen Algorithmen angepasst sind, werden sie den Algorithmen zugeordnet, befinden sich demnach innerhalb der Blackboxes.

3.2 Literaturübersicht

Ein sehr frühes und interessantes Beispiel für Onset-Erkennung stammt aus den 30er Jahren [Sea38]. Hier wird mit Hilfe einer in einem Piano installierten Kamera die Bewegungen der Hämmer gefilmt und somit optisch Start- aber auch Endpunkt sowie die daraus resultierende Dauer einer gespielten Note festgehalten. Es handelt sich hierbei also um Onset-Erkennung im Kontext einer automatischen Transkription.

Die maschinelle Informationsverarbeitung setzte zunehmend mit dem Fortschritt in Computertechnik und Signalverarbeitung ein. Besonders in letzter Zeit wurde viel geforscht, um Informationen direkt aus dem Audiosignal per digitaler Signalverarbeitung zu extrahieren. Die einfachsten Verfahren zur Onset-Erkennung untersuchen Amplituden- oder Energieschwankungen des Signals im Zeitbereich [AT99]. Zur Optimierung werden die Signale in verschiedene Frequenzbänder aufgeteilt und dortige Energieschwankungen ausgewertet [Dix99] [Dix03] [PK02]. Die meisten Systeme verwenden FFT-basierte Differenzenmodelle [GM95], teilweise auch aufgeteilt in Bändern [Kla03], andere werten Phaseninformationen aus [BS03]. Es gibt Verfahren, die andere Transformationen verwenden, wie zum Beispiel die Wavelettransformation [Smi96a] oder die Diskrete Kosinustransformation [VM97]. In [MM02] wird ein neuronales Netz zur Onset-Erkennung in Piano-Musik eingesetzt, eine Weiterentwicklung von [Smi96b], der diesen Ansatz zu Sparachsegmentierung verwendet. [AP03] stellt ein zweistufiges System vor, bei dem die erste Stufe mit Hilfe der Independent Component Analysis eine Art Onsetkurve erstellt und über „Onset oder Nicht-Onset“ mit einem Hidden Markov Modell entschieden wird.

Andere Ansätze versuchen die FFT-basierte Analyse zu optimieren, einerseits durch verschiedene Distanzmessungen [HM03], andererseits durch harmonische

Analysen mit großen FFT-Blockgrößen, welche die Wahrscheinlichkeit von „Akkordwechseln“ feststellen sollen [GM99] [Got01]. Nicht nur auf Energiedifferenzen, sondern auf der Kombination der Betrags- und Phaseninformation basiert [DBDS03], in [DDS03] realisiert durch Betrachtung der komplexen Ebene.

Diese Arbeit betrachtet Verfahren im Zeitbereich und FFT-basierte Ansätze im Frequenzbereich, da sie im Vergleich zu den übrigen ähnliche Ergebnisse liefern aber weniger komplex und meist auch Echtzeit-fähig sind. Sie werden am häufigsten verwendet und sind gut erweiterbar. Für Zeit- und Frequenzbereich werden im Folgenden repräsentative Beispiele untersucht, die im Rahmen dieser Arbeit in MATLAB implementiert wurden, um sich ein genaues Bild über die Leistungsfähigkeit zu verschaffen.

3.3 Algorithmen im Zeitbereich

3.3.1 Signalenergie und Hüllkurve

Es werden zwei Verfahren beschrieben, bei denen die Onsets - genauer die Onsetkurven - aus dem Zeitsignal $x[n]$ berechnet werden. Die Vorgehensweise folgt hierbei jeweils dem Gedanken, die Onsets im starken Anstieg der Signalintensität zu suchen. Für die Berechnung sind die meisten Arbeitsschritte beider Verfahren gleich und werden deshalb gemeinsam in Abbildung 3.2 zusammengefasst dargestellt. Der Unterschied besteht in der ersten Stufe von Ebene 2, in der es darum geht, aus dem Zeitsignal den Verlauf der Intensität zu extrahieren. Das geschieht entweder durch Integrieren der Signalenergie oder durch Bilden einer Hüllkurve mittels Tiefpassfilterung. Betrachtet man das Blockbild als Ganzes, beschreibt es den Inhalt der ersten Blackbox aus dem Schema in Abbildung 3.1.

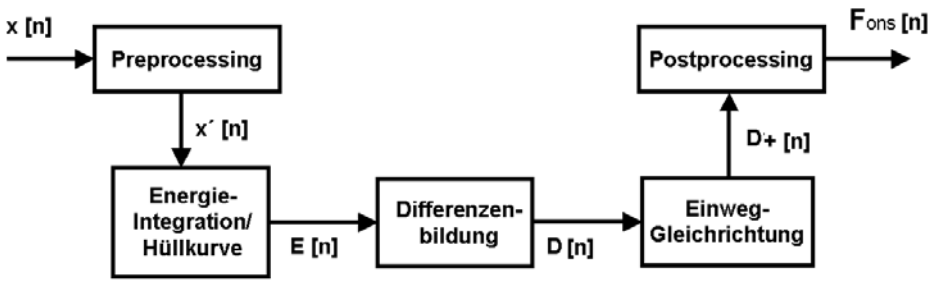

Abbildung 3.2: Blockbild der Arbeitsschritte zur Transformation in eine Onsetkurve im Zeitbereich

3.3.1.1 Preprocessing

Da das Musiksignal als Zeitsignal am Eingang anliegt, kann es im Grunde direkt verwendet werden. Es erfolgt jedoch zumeist ein Preprocessing-Schritt. Um Berechnungsaufwand zu sparen, werden hierbei Stereosignale nach mono konvertiert und häufig auch ein Downsampling vorgenommen. Oft führt man auch eine Pegelnormierung durch, um die Energieverhältnisse anzupassen.

3.3.1.2 Integration der Signalenergie

Ein simpler Algorithmus [DDS03] setzt den Verlauf der Signalintensität mit der Signalenergie in Beziehung. Der Beginn eines musikalischen Ereignisses geht somit mit einem Anstieg der Signalenergie einher. Das ist bei transienten Perkussivlauten sehr ausgeprägt der Fall. Die Energie wird entlang des Signals jeweils für ein kurzes Intervall berechnet nach

$$E_{Energie}[n] = \sum_{n'=(n-1)L}^{nL} |x[n']|^2 . \tag{3.1}$$

L: Länge des Intervalls
n: Index des Intervalls

Abbildung 3.4 zeigt das Ergebnis der Energie-Integration. Der Verlauf hat in etwa die Form einer Hüllkurve des Zeitsignals, verzerrt nach der Energie. Wie in Abschnitt 2.1 beschrieben, können Onset-Zeiten in charakteristischen Einschwingvorgängen des Hüllkurvenverlaufes gefunden werden. Anstelle der Energie-Integrierung gemäß Gleichung 3.1, kann somit auch die Berechnung einer Hüllkurve in die Kette der Arbeitsschritte treten. Tatsächlich findet dieser Ansatz auch häufiger Anwendung [Sch97] [Kla99]. Um aus dem Zeitsignal eine Hüllkurve zu berechnen und dadurch den Verlauf der Signalintensität abzubilden, ist es wichtig eine geeignete Strategie zu finden, das Zeitsignal zu integrieren.

3.3.1.3 Hüllkurve

In [Sch97] ist zur Bildung einer Hüllkurve $E_{Huell}[n]$ folgendes Verfahren beschrieben: Das Zeitsignal wird durch Betragsbildung und anschließender Tiefpassfilterung in die Hüllkurve transformiert. Für die Qualität der Hüllkurve ist die Impulsantwort H_{tp} des Filters entscheidend. Diese hat hier die Form eines halben Hanningfensters mit einer Länge von 200 ms (vergleiche Abbildung 3.3). Die Wahl basiert auf psychoakustischen Untersuchungen, die ergaben, dass jenes Filter eine dem menschlichen Gehör ähnliche Energie-Integrierung durchführt.

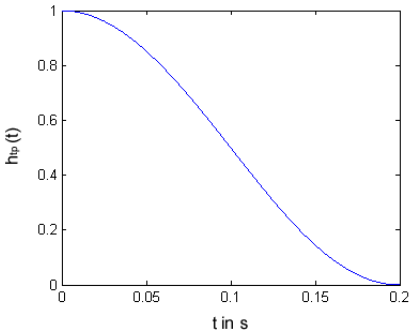

Abbildung 3.3: Impulsantwort des Tiefpassfilters

3.3.1.4 Differenzenbildung

Um einen Anstieg der Signalintensität auszumachen, muss man den Signalverlauf auf markante Änderungen hin untersuchen. Dieses wird anhand der ersten Ableitung von $E[n]$ durchgeführt, wofür im diskreten Fall jeweils die Differenz zwischen aktuellem Stützwert und seinem Vorgänger zu berechnen ist. Als Formel:

$$D[n] = E[n] - E[n-1]. \tag{3.2}$$

3.3.1.5 Einweg-Gleichrichtung

Die diskrete Ableitung zeigt die Bereiche von $E[n]$ mit den größten Änderungen auf, wodurch Onsets und Offsets charakterisiert sind. Stationäre Bereiche bleiben unbetont. Auf der Suche nach den Onset-Zeiten sind lediglich steigende Flanken von Interesse, weshalb im Anschluss an die Differenzenbildung die negativen Flanken abgeschnitten werden. Dieses wird häufig auch als Einweg-Gleichrichtung *(engl. half-wave rectification)* bezeichnet.

3.3.1.6 Postprocessing

Das Signal weist nach der Einweg-Gleichrichtung noch einen recht hohen Rauschanteil auf. Deshalb wird es in einem Postprocessing-Schritt mittels Tiefpassfilterung geglättet, um eine sinnvolle Ausgangslage für das Peakpicking zu schaffen. Als Filter eignet sich hierbei gut ein Tiefpass mit der Impulsantwort eines Hanningfensters. Es ist dabei zu beachten, dass diese Filterung einen zeitlichen Versatz zu der tatsächlichen Onset-Zeit erzeugt, der im nachhinein kompensiert werden muss. Als Ergebnis erhält man die Onsetkurve F_{ons} (Abildung 3.4), die zur weiteren Verarbeitung dem Peakpicking-Algorithmus übergeben wird.

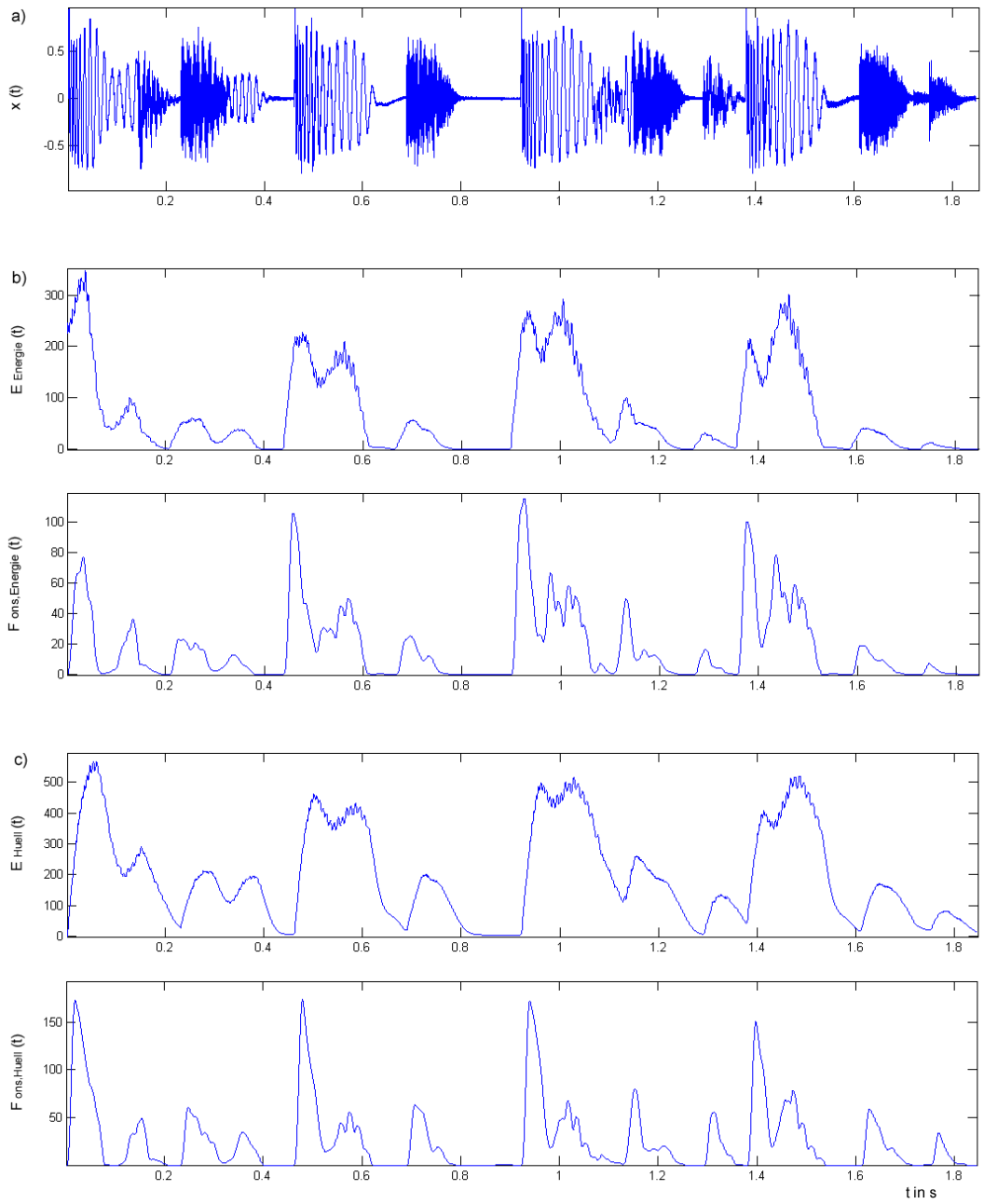

Abbildung 3.4: Ergebnisse der Algorithmen angewendet auf einen Schlagzeug-Loop: a)Zeitsignal, b)Energie-Integration und Onsetkurve, c)Hüllkurve und Onsetkurve

3.3.2 Stärken der Verfahren

Die Stärken der Verfahren können aus zwei Blickwinkeln betrachtet werden, zum einen, für welche Art von Musiksignalen sie besonders geeignet sind und zum anderen, welche Vorteile sie hinsichtlich Recheneffizienz bieten.

Im Bezug auf die Art des Musiksignals am Eingang sind die obigen Verfahren besonders gut für Schlagzeug-basierte Musik geeignet. Perkussive Schläge mit transienten Einschwingvorgängen zeichnen sich sehr deutlich im Hüllkurvenverlauf ab, so dass diese Onsets mit einer zeitlich recht hohen Genauigkeit extrahiert werden können. Schläge im Bassbereich kommen vor allem durch ihren hohen Energiegehalt gut zur Geltung. Da das Schlagzeug den Rhythmus bestimmt und die Verfahren am sensibelsten auf dessen Onsets reagieren, liefert der Einsatz in einem Beattracking-System bei Schlagzeug-basierter Musik sehr gute Ergebnisse.

Aber auch bei Musik ohne Schlagzeug können die Verfahren eingesetzt werden. So werden Onsets einstimmiger Stücke sehr gut erkannt, in denen Instrumente vorkommen, dessen Klangerzeugung markante Einschwingphasen aufweisen, wie zum Beispiel Anschläge von Tasten eines Klaviers oder Saiten einer Gitarre. Dabei ist jedoch stets vorausgesetzt, dass die Anschläge auch explizit genug gespielt werden.

Die Stärken im Hinblick der Signalverarbeitung liegen zum einen in der geringen Komplexität der Berechnung. Sowohl die Energie-Integration als auch die Tiefpassfilterung kann mit linearem Aufwand bewältigt werden, weshalb sich diese Algorithmen besonders gut für Echtzeitsysteme eignen. Zum anderen ist die gute zeitliche Auflösung zu nennen. Da im Gegensatz zu frequenzbasierten Verfahren (vergleiche Abschnitt 3.4) das Signal mit einer Schrittweite von einem Stützwert durchlaufen wird, ist lediglich die zeitliche Ungenauigkeit zu beachten, die aufgrund der gewählten Art der Signal-Integration entsteht. Dieser zeitliche Versatz ist anhand der Intervallgröße, bzw. bei der Tiefpassfilterung anhand der Länge und Form der Impulsantwort des Tiefpassfilters, in etwa abzuschätzen.

3.3.3 Bekannte Schwächen

Die obigen Verfahren extrahieren Onsets aus Hüllkurven oder Verläufen ähnlich der Hüllkurve des Gesamtsignals. Sie berücksichtigen dabei nicht die unterschiedlichen spektralen Eigenschaften der musikalischen Ereignisse. Verwendet man zum Beispiel die Energie-Integration (Gleichung 3.1) fällt auf, dass tiefe Frequenzen wesentlich höher ins Gewicht fallen, da die Energie zu tiefen Frequenzen zunimmt (vergleiche Abbildung 3.4). Wichtige Onsets von Klängen, deren Energieanteile in den höheren Frequenzen liegen, können daher von Schwingungen energetisch stärkerer Frequenzen verdeckt werden und somit aufgrund des Verfahrens verloren gehen.

Natürliche Musiksignale bestehen oft aus Gemischen von Klängen, die sich stark überlappen. So kommt es vor, dass ein Onset durch stationäre Phasen oder

Ausschwingvorgänge anderer musikalischer Ereignisse überlagert wird und in der Hüllkurve des Gesamtsignals untergeht. Abbildung 3.5 zeigt einen Ausschnitt eines Musikstücks, das durch Tiefpassfilterung in eine Hüllkurve und mittels Differenzenbildung in die Onsetkurve transformiert wurde. Weiterhin ist die Zeit t_{ons} eines auditiv als sehr dominant wahrzunehmenden HiHat-Onsets markiert. Bereits der Hüllkurven-Darstellung ist zu entnehmen, dass dieser Onset eine sehr schwache Ausprägung aufweist und beim Peakpicking vermutlich nicht als herausragendes Maximum bewertet wird. Eine gewichtige Onset-Zeit wird demnach nicht erkannt.

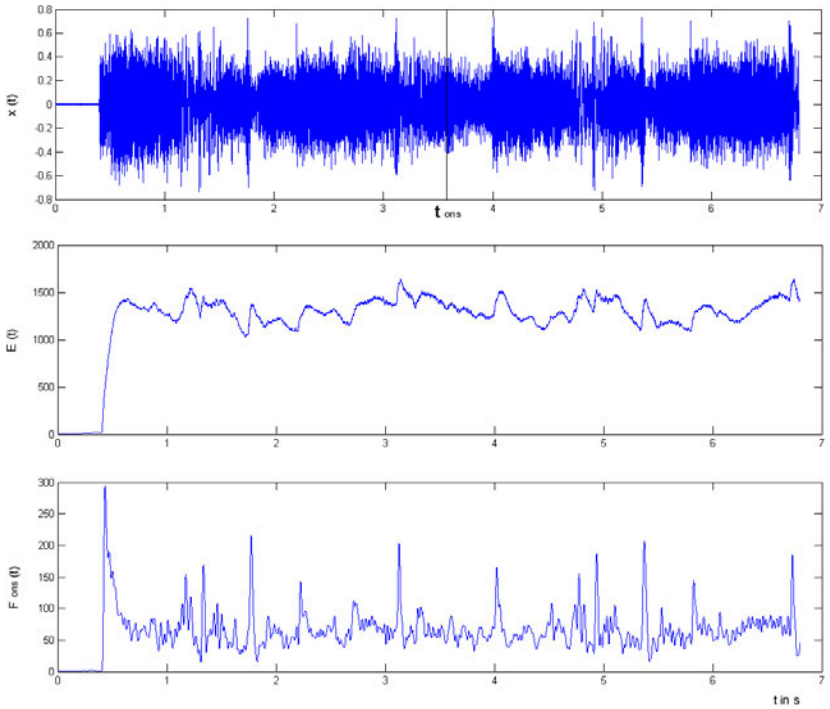

Abbildung 3.5: Ausschnitt aus Popmusik: Zeitsignal (oben), Hüllkurve (mitte), Differenzensignal (unten), auditiv dominante Onsetzeit t_{ons} (markiert)

Die entscheidende Schwäche dieser Ansätze liegt also im Grundgedanken. Es wird vorausgesetzt, dass sich der Eintritt eines musikalischen Ereignisses immer durch einen deutlichen Anstieg der Signalintensität äußert. Die Fähigkeit der Algorithmen hängt deshalb davon ab, ob am Beginn des Ereignisses ein transienter Anteil, wie zum Beispiel ein Anschlag, vorhanden ist. Natürlich ist dies oft auch der Fall. Bei einer Großzahl von Musikinstrumenten folgt der harmonische Anteil einem Anschlag, bedingt durch die Physik des Instrumentes oder der Spielweise. Diese Art des Einschwingens hat meist einen sehr rauschhaften Charakter, bei dem die Energie in etwa über das ganze Spektrum gleich verteilt ist. Spezielle spektrale Eigenschaften müssen dann nicht berücksichtigt werden und es genügt in der Hüllkurve des gesamten Signals zu suchen.

Es gibt aber auch eine Vielzahl an Gegenbeispielen. Genannt seien hier zum Beispiel ineinander weich übergehend gesungene Töne ähnlicher Intensität oder Instrumente, die in ihrer Klangerzeugung nicht von Anschlägen oder sonstigen transienten Vorgängen abhängen. Extreme Beispiele sind in der synthetischen Klangerzeugung zu finden, wo es mitunter möglich ist, Übergänge vollständig unkenntlich zu gestalten. Harmonische Änderungen, die nicht durch einen Anschlag oder Ähnliches angeregt werden, sind mit diesen Verfahren somit nicht auszumachen.

3.3.4 Optimierung

Ansätze zur Optimierung zielen darauf, den Problemen der Überlagerungs- und Verdeckungseffekten entgegenzutreten und die Frequenzauflösung zu verbessern. Dafür wird das Signal mittels einer Filterbank in mehrere Frequenzbänder aufgeteilt und jedes Band separat betrachtet.

E. Scheirer beschreibt dazu in [Sch97] eine psychoakustische Untersuchung zur Rhythmus-Erkennung, die zeigt, dass das Musiksignal auf gewisse Weise vereinfacht werden kann, ohne den wahrgenommenen rhythmischen Inhalt zu beeinflussen. Dazu wird das Gesamtsignal in Frequenzbänder geteilt und von jedem Band die Hüllkurve berechnet. Anschließend wird ein Rauschsignal in dieselben Bänder zerlegt und jedes Teilrauschsignal mit der jeweiligen erzeugten Hüllkurve des Originalsignals moduliert. Dem Hörer wird nun die Summe dieser modulierten Rauschsignale präsentiert und seine rhythmische Wahrnehmung untersucht. Die Ergebnisse besagen, dass der Rhythmus in gleicher Weise wahrgenommen wird wie der des Originals, sobald eine Aufteilung in mindestens vier Frequenzbänder vorgenommen wird. Benutzt man hingegen nur ein einziges Band, ist der Rhythmus des ursprüngliche Signals aus seiner vereinfachten Form nicht mehr erkennbar.

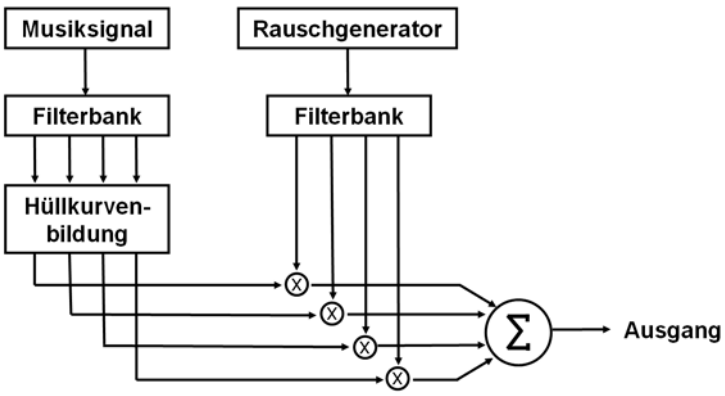

Abbildung 3.6: Aufbau für die psychoakustische Untersuchung der Rhythmus-Wahrnehmung nach Scheirer [Sch97]

A. Klapuri greift diesen Gedanken auf und erstellt in [Kla99] ein System zur automatischen Onset-Erkennung mit einer Filterbank bestehend aus insgesamt 21 Bandpassfiltern. Diese Filter teilen das Signal in 21 nicht überlappende Frequenzbänder in Anlehnung der kritischen Bandbreiten aus der Psychoakustik [ZF99]. Der Aufbau des Gesamtsystems ist in Abbildung 3.7 dargestellt. Für jedes Band wird die Hüllkurve durch Tiefpassfilterung gebildet und die Onset-Komponenten extrahiert. Für jede Onset-Komponente wird Zeitpunkt und Intensität festgelegt. Am Ende werden die Ergebnisse aus den einzelnen Bändern kombiniert und zu einem Vektor mit Onset-Zeiten zusammengefasst.

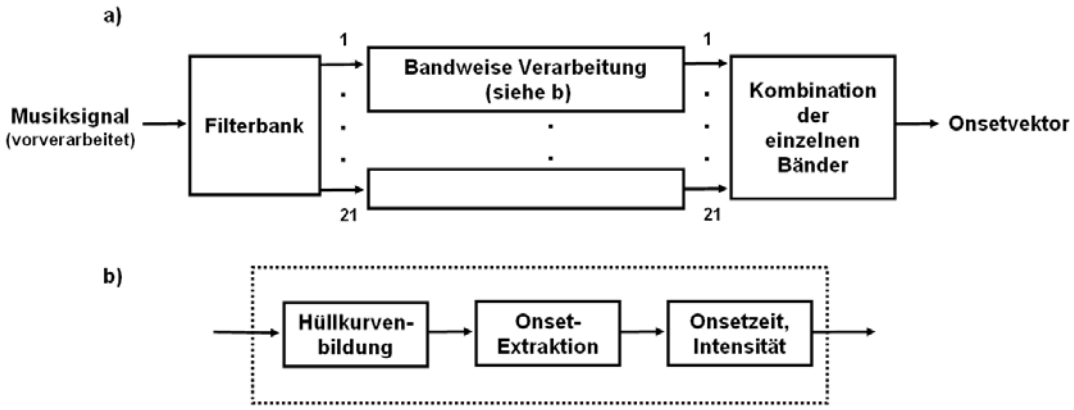

Abbildung 3.7: System zur automatischen Onset-Erkennung in [Kla99]: a) Überblick, b) Komponente im Detail

Für dieses System wird weiterhin eine Methode beschrieben, die Genauigkeit der erkannten Onset-Zeiten zu optimieren. Hierzu werden die Signale der einzelnen Frequenzbänder nach der Differenzenbildung betrachtet. Es stellt sich heraus, dass diese gut die Intensität der auftretenden Onsets widerspiegeln, die jeweiligen Maxima jedoch nicht präzise genug den Zeitpunkt des Auftretens markieren. Das äußert sich zum einen darin, dass Ereignisse in tiefen Frequenzen längere Einschwingphasen aufweisen als in hohen Lagen. Die Zeiten von Onsets, die sich in mehreren Bändern abzeichnen, sind somit bei der Kombination der Frequenzbänder schwer festzulegen, da das mit einem Onset assoziierte Maximum zwischen den Bändern verschoben ist. Zum anderen fällt auf, dass die Hüllkurve bei einem Onset nicht zwangsweise monoton ansteigt. Daher erhält man nach der Differenzenbildung oft mehrere lokale Maxima in der Nähe des Onsets (vergleiche Abbildung 3.8). Klapuri schlägt vor, anstelle der bisherigen absoluten Differenzenbildung eine *relative Differenzenfunktion* $W(t)$ einzuführen. Diese wird gebildet, indem die absolute Differenzenfunktion $D[n]$ mit der Funktion der Hüllkurve $E[n]$ ins Verhältnis gesetzt wird.

Für das Frequenzband k bedeutet dies:

$$W[n,k] = \frac{D[n,k]}{E[n,k]}. \tag{3.3}$$

Die Intensität eines Onsets wird durch Multiplikation des Wertes der absoluten Differenzenfunktion mit der Mittenfrequenz des Frequenzbandes k festgelegt. Dadurch sollen die Ausgaben der bandweisen Verarbeitung für die anschließende Zusammenfassung angepasst und höhere Frequenzen stärker bewertet werden.

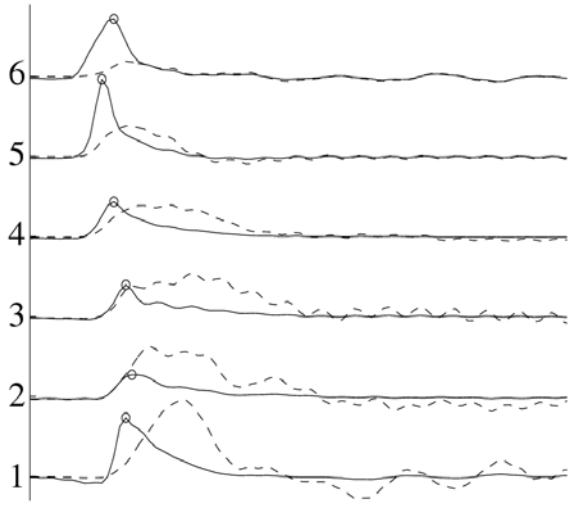

Abbildung 3.8: Onset eines Klaviertones: absolute Differenzenfunktion (gestrichelt), relative Differenzenfunktion (durchgehend) [Kla99]

3.4 Algorithmen im Frequenzbereich

Die Optimierungen aus dem vorigen Abschnitt basieren darauf, besser auf die spektralen Eigenschaften unterschiedlicher musikalischer Ereignisse einzugehen. Im diesem Sinne könnten diese Verfahren auch dem Frequenzbereich zugeordnet werden. Da zur Transformation in die Onsetkurven jedoch weiterhin ausschließlich Hüllkurven der Zeitsignale betrachtet werden, sind sie dem Zeitbereich zugeteilt. Algorithmen im Frequenzbereich betrachten in dieser Arbeit ausnahmslos das Spektrum des Musiksignals, der Grundgedanke ist jedoch derselbe: Onsets werden in markanten Änderungen eines Signals gesucht, mit dem Unterschied, dass das Signal diesmal den zeitlichen Verlauf des Spektrums darstellt und es sich deshalb um spektrale Änderungen handelt. Eine anschauliche Darstellung bietet hier das Spektrogramm. Abbildung 3.9 zeigt zum Beispiel das Spektrogramm des in Abschnitt 3.3.3 behandelten Signals. Die spektralen Änderungen sind deutlich erkennbar, sogar der im Hüllkurvenverlauf untergehende Onset sticht klar hervor (vergleiche hierzu Abbildung 3.5).

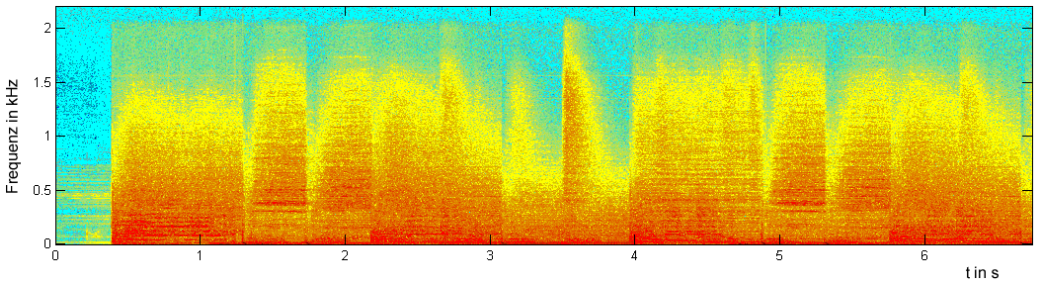

Abbildung 3.9: Spektrogramm eines Ausschnitts aus Popmusik

Für alle folgenden Verfahren wird das Zeitsignal mittels Kurzzeit-FFT (vergleiche 2.2.3.2) in das Spektrum transformiert. Wichtige Parameter sind hierbei FFT-Größe, Fensterung und Schrittweite, die sowohl Zeit- als auch Frequenzauflösung bestimmen. Den Verfahren gemeinsam ist der generelle Ablauf und, dass sie die Änderungen anhand einer bestimmten Distanz zwischen aufeinander folgenden FFT-Blöcken feststellen. Trägt man diese Distanz über der Zeit auf, erhält man eine Onsetkurve.

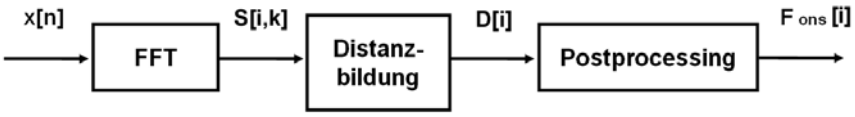

Abbildung 3.10: Gemeinsames Blockbild der Algorithmen im Frequenzbereich

Abbildung 3.10 zeigt wie auch das Blockbild in Abbildung 3.2 den Inhalt

der ersten Blackbox aus dem Grundschema der Onset-Erkennung (Abbildung 3.1). Das Postprocessing hat wiederum die Aufgabe, mittels Tiefpassfilterung den Signalverlauf zu glätten. Zur Bildung der Distanz gibt es nun verschiedene Ansätze, die im Folgenden beschrieben werden.

3.4.1 Distanzen im Betragsspektrum

Aus dem komplexen Ergebnis der FFT lässt sich sowohl ein Betrags- als auch ein Phasenspektrum berechnen (vergleiche Abschnitt 2.2.3.1). Die drei folgenden Distanzen betrachten jedoch nur das Betragsspektrum des Musiksignals.

3.4.1.1 Euklidische Distanz

Eine simple Variante stellt die Euklidische Distanz D_{EUC} dar. Von jedem Frequenz-Bin eines FFT-Blocks wird der Bin des Vorgänger-Blockes abgezogen und die Differenz quadriert. Anschließend werden die einzelnen Werte des Ergebnisvektors aufsummiert. Für jeden FFT-Block n ist also die Euklidische Distanz wie folgt zu berechnen:

$$D_{EUC}[n] = \sum_{k=0}^{\frac{N}{2}-1} (A[n,k] - A[n-1,k])^2. \tag{3.4}$$

N: FFT-Größe
$A[n,k]$: Betrag des Kurzeitspektrums des n-ten FFT-Blocks

3.4.1.2 MKL-Distanz

Diese Distanz wird in [HM03] beschrieben. Sie wird von der Kullback-Leibler Distanz abgeleitet und deshalb *Modified Kullbach-Leibler (MKL)* genannt. Es werden zwei aufeinander folgende FFT-Blöcke betrachtet und pro Frequenz-Bin eine Distanz $d[n]$ errechnet nach:

$$d[n,k] = log_2 \left(\frac{A[n,k]}{A[n-1,k])} \right). \tag{3.5}$$

Anschließend wird die Summe gebildet, jedoch mit der Bedingung verknüpft, nur die positiven Werte des Distanzvektors aufzusummieren. Damit wird erreicht, dass lediglich Anstiege der Intensität, also mögliche Onsets, erkannt werden.

$$D_{MKL}[n] = \sum_{k=0, d[k]>0}^{\frac{N}{2}-1} d[n,k]. \tag{3.6}$$

3.4.1.3 Distanz nach Goto

M. Goto und Y. Muraoka haben im Rahmen ihrer Forschungen eine Reihe von Artikeln zum Thema Beattracking veröffentlicht [GM95] [Got01]. Dabei wurde viel Wert auf die automatische Erkennung von Onsets aus dem Musiksignal gelegt. Ziel war es möglichst robuste Systeme zu entwerfen, sowohl für Schlagzeug-basierte Musik, als auch für Musiksignale ohne Schlagzeug. Für diese Algorithmen wurde auch eine eigene Distanz entwickelt, die das Betragsspektrum betrachtet, jedoch im Gegensatz zu den vorigen auch benachbarte der jeweiligen Frequenz-Bins zur Bildung der Distanz berücksichtigt.

Jede Teildistanz berechnet sich nach

$$d[n,k] = \begin{cases} max(A[n,k], A[n,k+1]), \ min(A[n,k], A[n,k+1]) > A_{maxVor} \\ 0 \ sonst. \end{cases} \tag{3.7}$$

$$A_{maxVor} = max(A[n-1,k-1], A[n-1,k], A[n-1,k+1])$$

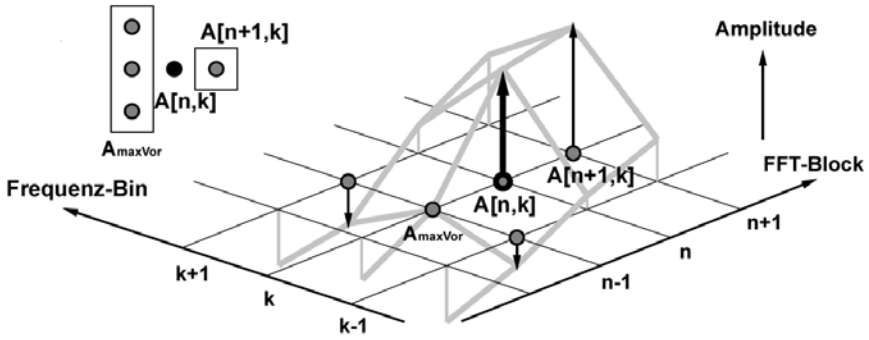

Abbildung 3.11: Distanz nach Goto [Got01]

Der Algorithmus beachtet wie die MKL-Distanz nur Anstiege der Signalintensität. Durch die Berücksichtigung der benachbarten Frequenz-Bins wird zusätzlich erreicht, dass positive Differenzen zwischen aufeinander folgenden Bins nur dann ins Gewicht fallen, wenn es sich insgesamt um einen Anstieg der Signalintensität zwischen den betrachteten FFT-Blöcken handelt. Die resultierende Gesamtdistanz ergibt sich aus

$$D_{GOTO}[n] = \sum_{k=0}^{\frac{N}{2}-1} d[n,k]. \tag{3.8}$$

3.4.2 Distanz in der Komplexen Ebene

Die bisher beschriebenen Verfahren im Frequenzbereich betrachten ausschließlich das Betragsspektrum, Phaseninfomationen werden nicht berücksichtigt. Unter-

suchungen in [DBDS03] zeigen, dass sich Onsets auch im Verlauf des Phasenspektrums abzeichnen. Die Phase einer Frequenz bzw. eines Frequenz-Bins ist in stationären Bereichen des Signals gut vorhersagbar (prädizierbar), das heißt, die Differenz der vorhergesagten Phase $\widetilde{\varphi}[n,k]$ des k-ten Bins eines FFT-Block n und der wahren Phase $\varphi[n,k]$ ist klein oder gleich Null. Hierbei berechnet sich $\widetilde{\varphi}[n,k]$ gemäß

$$\widetilde{\varphi}[n,k] = 2\varphi[n-1,k] - \varphi[n-2,k]. \tag{3.9}$$

$\widetilde{\varphi}$: unwrapped Phase[1]

Tritt hingegen ein neues Ereignis - ein Onset - auf, steigt diese Differenz deutlich an. Betrachtet man die Verteilungsdichtefunktion dieses Vorhersage-Fehlers aller k Frequenz-Bins, weist diese bei stationären Bereichen einen spitzen Verlauf mit einer hohen Konzentration um den Nullwert auf. Bei einem Onset flacht der Verlauf deutlich ab und die Funktion nimmt an Breite zu. Durch Auftragen dieser Breite über der Zeit, lässt sich somit eine Onsetkurve erstellen.

Basierend auf dem obigen Ansatz wird in [DBDS03] zur Berechnung einer Onsetkurve die Kombination der Informationen aus Betrags- und Phasenspektrum verwendet. In [DDS03] wird diese Kombination auf die Bildung einer Distanz in der komplexen Ebene übertragen. Betrachtet man den k-ten Frequenz-Bin als Vektor in der komplexen Ebene (vergleiche Abbildung 3.12), werden Änderungen im Signal durch die Euklidischen Distanz zwischen aktuellem Vektor und seinem Vorgänger $d_{KMPX}[n,k]$ ausgedrückt:

$$\begin{aligned} d_{KMPX}[n,k] = \; & ((\Re(X[n,k]) - \Re(X[n-1,k]))^2 + ... \\ & \Im(X[n,k]) - \Im(X[n-1,k]))^2)^{\frac{1}{2}}. \end{aligned} \tag{3.10}$$

Die Gesamtdistanz $D_{KMPX}[n]$ ergibt sich wiederum aus der Summe der Teildistanzen $d_{KMPX}[n,k]$.

Der Zusammenhang der Distanzbildung im Betragsspektrum und in der komplexen Ebene wird folgendermaßen veranschaulicht: Dreht man die Vektoren aus Abbildung 3.12 bis der Vektor von $X[n-1,k]$ direkt auf der reellen Achse liegt, kann $X[n,k]$ dargestellt werden als

$$X[n,k] = A[n,k]e^{j\varphi[n,k]}. \tag{3.11}$$

Gleichung 3.10 kann somit umformuliert werden zu

$$\begin{aligned} d_{KMPX}[n,k] = \; & ((A[n-1,k] - A[n,k]\cos(\varphi[n,k]))^2 + ... \\ & (A[n,k]\sin(\varphi[n,k]))^2)^{\frac{1}{2}}. \end{aligned} \tag{3.12}$$

[1] Die resultierenden Phasen werden durch so genanntes unwrapping (engl.) auf den Bereich $[-\pi,\pi]$ abgebildet.

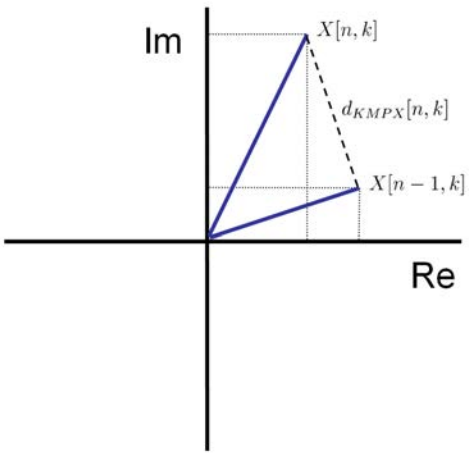

Abbildung 3.12: Distanz in der komplexen Ebene

Nach einem Erweiterungsschritt und anschließender Vereinfachung erhält man

$$d_{KMPX}[n,k] = \begin{array}{l} (A[n-1,k]^2 + A[n,k]^2 + ... \\ 2A[n-1,k]A[n,k]\cos(\varphi[n,k]))^{\frac{1}{2}}. \end{array} \tag{3.13}$$

Ist $\varphi[n,k] = 0$, wird $d_{KMPX}[n,k]$ zu $A[n,k] - A[n-1,k]$. Die Distanz in der komplexen Ebene kann somit interpretiert werden als die Distanz im Betragsspektrum gewichtet mit dem Fehler der Phasen-Prädiktion.

3.4.3 Merkmale frequenzbasierter Verfahren

Den zeitlichen Verlauf von Betrags- und Phasenspektrum eines Signals erhält man durch die Kurzzeit-FFT. Wie schon in den Grundlagen erörtert (vergleiche Abschnitt 2.2.3.2), ist es hierbei stets notwendig einen Kompromiss aus Zeitauflösung, Frequenzauflösung und Rechenzeit einzugehen. Für die Onset-Erkennung ist die zeitliche Auflösung besonders wichtig. Deshalb muss beachtet werden, die Schrittweite und die Fenstergröße möglichst klein zu halten. Das geht auf Kosten der Frequenzauflösung. So sind spektrale Änderungen in den tiefen Frequenzen zum Teil nicht mehr erkennbar, da sich dort die grobe Frequenzauflösung verstärkt bemerkbar macht. Ist die Frequenzauflösung zu gering, ist der Vorteil der guten zeitlichen Auflösung in tiefen Frequenzen nicht mehr nutzbar.

Abbildung 3.13 zeigt das Ergebnis durch Distanzenbildung in der komplexen Ebene für ein Musiksignal (Popmusik) mit der Abtastfrequenz 44.1 kHz aufgrund zweier unterschiedlicher Parametrisierungen:

- 256 Punkte FFT-Größe, 1,45 ms Schrittweite,
- 4096 Punkte FFT-Größe, 1,45 ms Schrittweite.

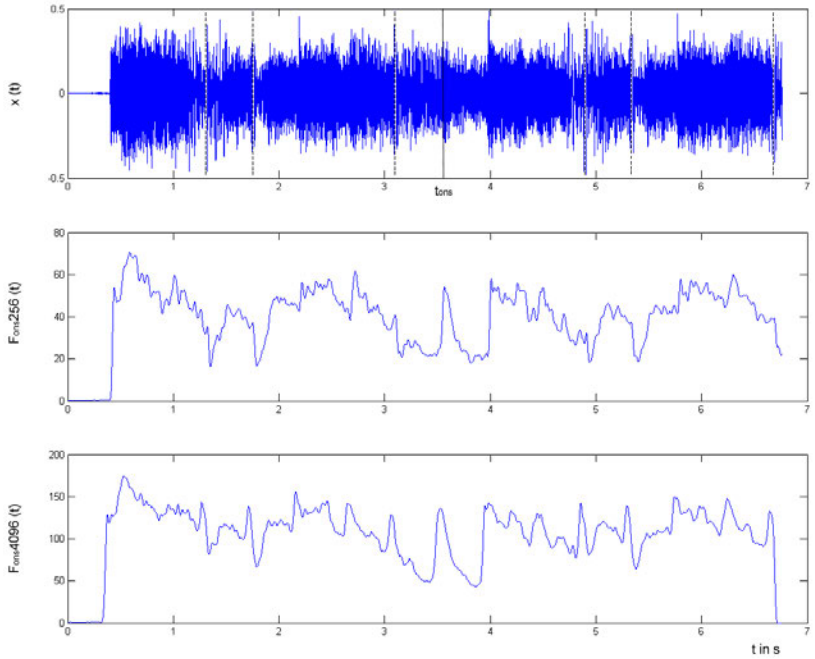

Abbildung 3.13: Zeitsignal (oben) und Onsetkurven durch Distanzbildung in der komplexen Ebene mit verschiedenen FFT-Größen: 256 (mitte), 4096 (unten), gestrichelt: Bassdrum-Schläge, durchgezogen: markanter Onset t_{ons}

Gestrichelt markiert sind jeweils Schläge einer Bassdrum. Bei einer 256-Punkte-FFT ist deutlich zu beobachten, dass diese Schläge aufgrund der niedrigen Frequenzauflösung im Bassbereich untergehen. Die Auflösung beträgt hier ca. 172 Hz, so dass der gesamte Bassbereich lediglich durch zwei Frequenz-Bins repräsentiert und damit unterbewertet wird. Im Vergleich dazu hat die 4096-Punkte-FFT eine Frequenzauflösung von ca. 10 Hz. Die Bassdrum-Schläge zeichnen sich in lokalen Maxima der Onsetkurve ab.

Zeitlich gesehen ist die kleinere FFT-Größe mit einer Auflösung von ca. 6 ms wesentlich genauer als die 4096-Punkte Variante mit 93 ms Auflösung. Hinzu kommt noch die Ungenauigkeit aufgrund der Schrittweite, die wie schon erwähnt möglich klein zu wählen ist und bei einer Abtastrate von 44,1 kHz und 64 Abtastwerten ca. 1,45 ms beträgt.

Der große Vorteil frequenzbasierter Verfahren liegt in der Unabhängigkeit von transienten Einschwingvorgängen in der Hüllkurve. So können bei einer einigermaßen guten Frequenzauflösung harmonische Wechsel erkannt werden, die sich nicht in Hüllkurvenverläufen abzeichnen. Geeignete Parametrisierungen für die Onset-Erkennung werden in Abschnitt 4.1 diskutiert. Auch deutliche spektrale Änderungen, die aufgrund von niedriger Energie von zeitbasierten Verfahren übersehen werden - so auch der Onset bei t_{ons} (Abbildungen 3.5 und 3.13 oben) - fallen durch klare Peaks in der Onsetkurve auf.

3.5 Ansatz mit Linearer Prädiktion

In diesem Abschnitt wird ein neuer Ansatz zur Optimierung bestehender Verfahren untersucht. Hierzu soll das Prinzip der linearen Prädiktion verwendet werden, das beispielsweise erfolgreich in der Sprachkodierung eingesetzt wird.

3.5.1 Idee

Ist ein Signal stationär und rauschfrei, ist der weitere Verlauf aufgrund vergangener Amplitudenwerte gut vorauszusagen. Plötzliche Änderungen äußern sich in einer Diskrepanz zwischen Vorhersage und Signalverlauf - einem Prädiktionsfehler. Diese Änderungen müssen sich nicht im Hüllkurvenverlauf des Zeitsignals abzeichnen, um einen deutlichen Anstieg bzw. einen Peak im Prädiktions-Fehlersignal zu verursachen. Abbildung 3.14 zeigt hierzu ein Beispiel durch eine Folge von vier Sinustönen verschiedener Frequenz und gleicher Amplitude.

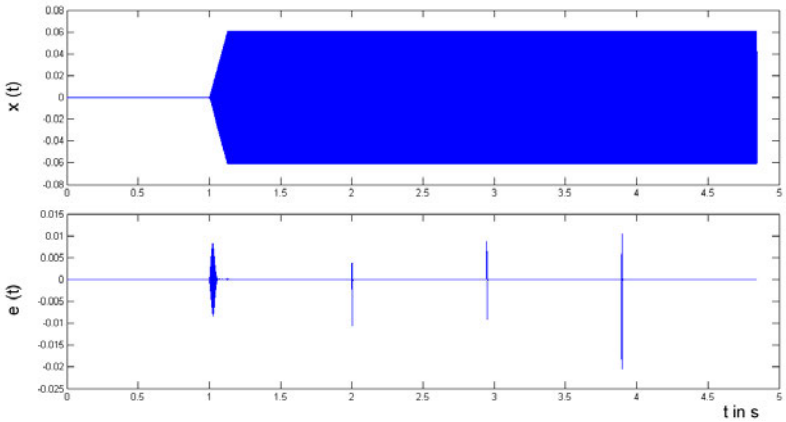

Abbildung 3.14: Zeitsignal und Prädiktionsfehler einer Folge von Sinustönen: 200 Hz, 250 Hz, 300 Hz, 400 Hz

Bei komplexeren Signalen werden sich diese Änderungen zwar nicht zwangsläufig in markanten Peaks äußern, könnten jedoch deutlicher im Prädikionsfehler zu erkennen sein als im ursprünglichen Eingangssignal. Die lineare Prädiktion kann somit als Vorverarbeitungsstufe verwendet werden, um Änderungen im Signalverlauf hervorzuheben bzw. zu verstärken.

3.5.2 Algorithmus

Prädiktion bedeutet, Vorhersagen über das Signal zu treffen, hier genauer, ein Signal $x[n]$ aus seinen vorangegangenen Werten $x[n-1], x[n-2], \ldots, x[n-M]$ zu bestimmen und den Vorhersagefehler festzuhalten (vergleiche Abschnitt 2.2.6).

Letzteres erinnert an die oben beschriebenen Verfahren in Zeit- und Frequenzbereich. Die Bildung von Differenzen und Distanzen ist im Grunde nichts anderes, als den aktuellen Stützwert durch vorhergehende Stützwerte zu prädizieren. Die Differenz aus aktuellem Stützwert $x[n]$ und seinem direkten Vorgänger $x[n-1]$ kann als Prädiktor erster Ordnung mit dem Koeffizienten $b_1 = 1$ dargestellt werden.

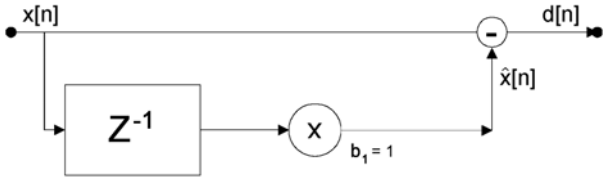

Abbildung 3.15: Differenz- bzw. Distanzbildung als Prädiktor erster Ordnung

Dies wird auf das Prinzip der linearen Prädiktion übertragen, wie in Abbildung 3.16 illustriert, auf ein System mit mehreren Prädiktor-Koeffizienten, der Auswertung des Fehlersignals und einer Adaptionsregel zur Anpassung der Koeffizienten.

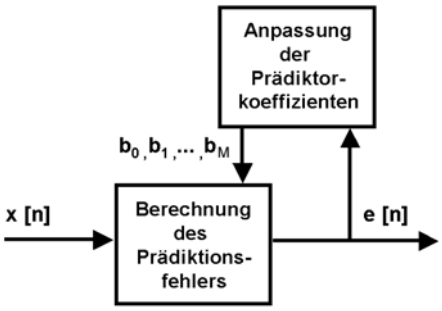

Abbildung 3.16: System zur Bildung der Onsetkurve mit linearer Prädiktion

Der Prädiktionsfehler wird für jeden Stützwert einzeln berechnet und danach die Prädiktorkoeffizienten im Hinblick auf Minimierung des Fehlers angepasst. Dazu wird ein Gradientenabstiegsverfahren verwendet [Hay96]. Dieses Verfahren führt gemittelt über der Zeit zu einer Minimierung des Prädiktionsfehlers. Die Adaption der Prädiktorkoeffizienten pro Iteration erfolgt mit:

$$b_l[n+1] = b_l[n] - \mu \cdot \Delta b_l. \tag{3.14}$$

Hier bei stellt Δb_l den Gradienten dar, der sich nach [Hay96] mit dem Abtastwert $x[n]$ und den aktuellen Prädiktionsfehler $e[n]$ folgendermaßen berechnet:

$$\Delta b_l = -2 \cdot e[n] \cdot x[n-l]. \tag{3.15}$$

Der Faktor μ aus Gleichung 3.14 ist ein konstanter Adaptionsfaktor, von dem Geschwindigkeit und Stabilität der Adaption abhängen. Für das Verfahren sind somit zwei Parameter festzulegen, welche die Qualität der Vorhersage bestimmen: Anzahl der Koeffizienten und Adaptionsfaktor.

Bei dem Algorithmus handelt sich hierbei um ein so genanntes Verfahren der kleinsten quadratischen Abweichung *(LMS)*[2]. Um die Stabilität der Adaption zu verbessern, wird der *NLMS* [3] verwendet, der den Gradienten auf die Energie des Eingangssignals normiert. Die Adaptionsregel lautet somit letztendlich:

$$b_l[n+1] = \frac{b_l[n] - \mu \cdot \Delta b_l}{\sum_{\forall n} x[n]^2 + c_{st}}. \tag{3.16}$$

c_{st}: Stabilisierungskonstante (häufig wird 1 gewählt)

3.5.3 Anwendbarkeit

Der Algorithmus zeigt Diskontinuitäten des Signal auf bzw. reagiert auf alles, was schlecht prädizierbar ist mit einem plötzlichen Anstieg des Prädiktionsfehlers. Schlecht prädizierbar ist Rauschen. So können transiente rauschartige Einschwingvorgänge, wie sie oft bei natürlichen Musikinstrumenten vorkommen, gut im Fehlersignal erkannt werden, besonders, wenn ihnen ein stationärer harmonischer Teil folgt. Besteht das Stück jedoch zum großen Teil aus rauschhaften Klängen oder werden gespielte Noten stark von rauschhaften Klängen überlagert, kann das Ergebnis sogar verschlechtert werden. Die Abbildungen 3.17 und 3.18 illustrieren Beispiele für eine gelungene Optimierung und eine starke Verschlechterung der Ergebnisse durch Anwendung der Vorverarbeitung mit linearer Prädiktion. Gezeigt sind Onsetkurven mit und ohne Vorverarbeitung, jeweils berechnet zum einen durch Hüllkurvenbildung im Zeitbereich (siehe Abschnitt 3.3.1), zum anderen durch Distanzbildung nach Goto im Frequenzbereich (siehe Abschnitt 3.4.1).

Die Untersuchungen ergeben, dass der Grad der Optimierung stark vom Eingangssignal abhängt. Dabei halten sich gute Ergebnisse und Verschlechterung ungefähr die Waage. Weiterhin sind für gute Ergebnisse eine hohe Anzahl an Prädiktorkoeffizienten zu wählen (>100), was einen großen Berechnungsaufwand bedeutet, da wie oben erwähnt die Adaption der Koeffizienten für jeden Abtastwert einzeln durchgeführt wird. In Hinblick auf das Ziel der Arbeit, einen robusten und effizienten Algorithmus zur Onset-Erkennung zu entwickeln, stellt somit dieses Verfahren keine Optimierung dar, wenn es wie beschrieben als Vorverarbeitungsschritt angewendet wird.

[2]engl. Least Mean Square
[3]Normalized LMS

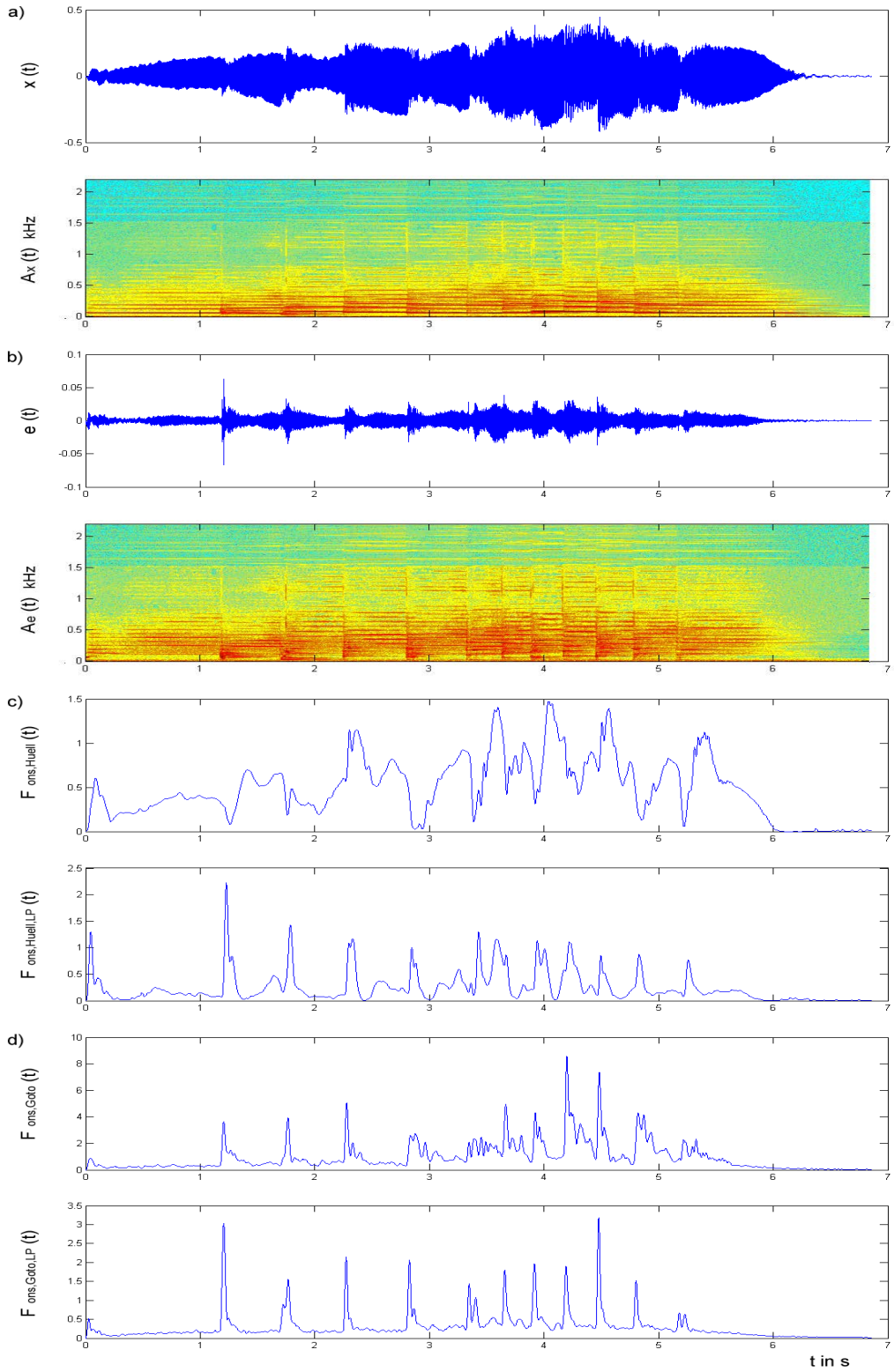

Abbildung 3.17: Gute Ergebnisse durch Vorverarbeitung mit linearer Prädiktion an einem Klarinettensolo: a) Originalsignal, b) Prädiktionsfehler, c) Onsetkurven im Zeit- und d) Frequenzbereich jeweils mit (oben) und ohne Vorverarbeitung (unten)

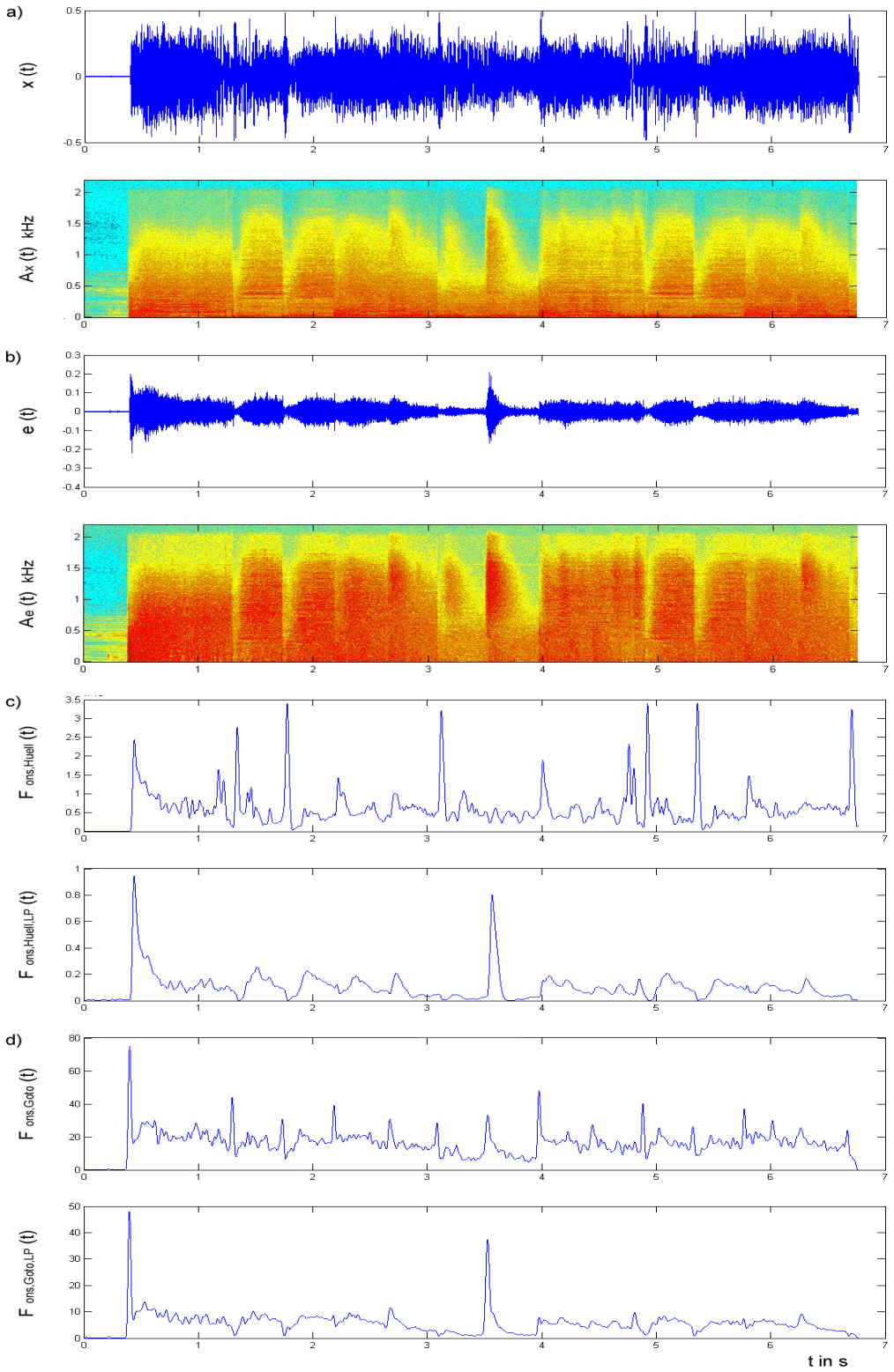

Abbildung 3.18: Verschlechterung der Ergebnisse durch Vorverarbeitung mit linearer Prädiktion an einem Popmusikausschnitt: a) Originalsignal, b) Prädiktionsfehler, c) Onsetkurven im Zeit- und d) Frequenzbereich jeweils mit (oben) und ohne Vorverarbeitung (unten)

3.6 Peakpicking

Wie am Anfang des Kapitels erwähnt, hat der Peakpicking-Algorithmus großen Einfluss auf die Qualität der Ergebnisse. Seine Aufgabe ist es, die herausstechendsten Maxima der Onsetkurve zu bestimmen, mit denen die Onsets des Musiksignals identifiziert werden. Selbst in stationären Phasen eines Klanges weisen die spektralen Verläufe natürlicher Musiksignale meist geringe Schwankungen auf, verursacht beispielsweise durch die physikalischen Gegebenheiten der Musikinstrumente oder durch die Spielweise. Die aus der Summe der Distanzbildungen resultierenden Onsetkurven der einzelnen Verfahren haben aus diesem Grund einen welligen Verlauf, dem ein stark rauschhaftes Signal überlagert ist. Trotz der im Postprocessing vorgenommenen Glättung der Onsetkurve, bleibt eine Welligkeit mit mehr oder weniger ausgeprägten Extremwerten erhalten (Abbildung 3.19). Das entscheidende Problem ist es, eine Schwelle zu definieren, um die aus diesen Schwankungen resultierenden lokalen Maxima herauszufiltern.

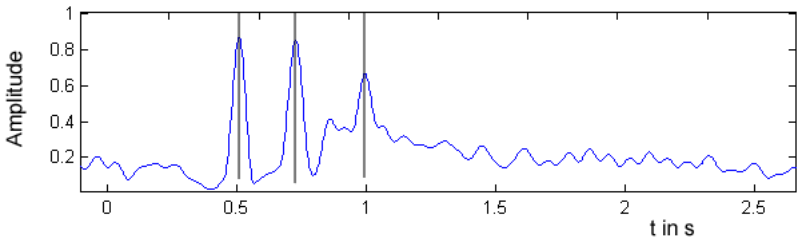

Abbildung 3.19: Auszug aus Onsetkurve mit starker Welligkeit. Neben den gekennzeichneten Onsets existieren viele weitere Maxima

Die in der Literatur zur automatischen Onset-Erkennung verwendeten Verfahren lassen sich aufgrund der Art und Weise den Schwellenwert festzulegen allgemein in statische und dynamische Verfahren unterteilen. Statisch bedeutet hier, dass es sich um konstante Schwellen handelt, die entweder auf Basis des gesamten Eingangssignals berechnet werden oder sehr große Blöcke betrachten, deren Grenzen nicht mehr im unmittelbaren Bezug des aktuell zu filternden Zeitpunktes der Onsetkurve liegen. In [HM03] wird der Schwellenwert beispielsweise durch den Mittelwert der Onsetkurve festgelegt (Abbildung 3.20). Es zeigt sich hier jedoch, dass durch dieses Verfahren zu viele falsche Onsets erkannt werden. Deshalb wird eine zweite Stufe eingesetzt, um diese Fehler wieder zu korrigieren. Eine weitere Methode betrachtet Blöcke von ca. fünf Sekunden Länge und wertet jeweils die Wahrscheinlichkeitsdichtefunktion (vergleiche Abschnitt 2.2.4) der auftretenden Maxima aus [DDS02]. Es wird zwar erwähnt, dass dieses Verfahren für beliebige Blockgrößen anwendbar ist, empfohlen werden jedoch Blockgößen mehrerer Sekunden. Für Näheres zu den beiden Verfahren sei hier auf die obigen Quellen verwiesen.

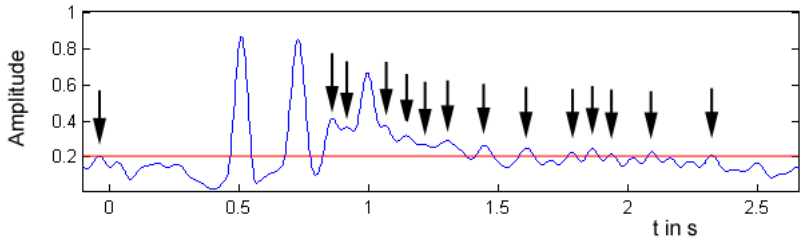

Abbildung 3.20: Statische Schwelle durch Mittelwertbildung. Falsch erkannte Onsets werden durch Pfeile markiert

Peakpicking-Algorithmen mit statisch festgelegten Schwellenwerten haben den Nachteil, dass sie sich schlecht an die Dynamik eines Musikstückes anpassen. Für eine robuste und sich auf das Eingangssignal hin optimierende Onset-Erkennung ist daher ein dynamisches Verfahren vorzuziehen. Diese berechnen den Schwellenwert für jeden Abtastwert aufgrund seiner unmittelbaren Umgebung. Hierzu wird um den zu betrachtenden Punkt n ein kleines Fenster gelegt und die jeweilige Schwelle $S[n]$ durch Mittelung der Werte innerhalb des Fensters berechnet:

$$S[n] = C_S \cdot \mu(w_n), \quad w_n \in [n - \frac{L}{2}, n + \frac{L}{2}]. \tag{3.17}$$

L: Fensterlänge
μ: Mittelwert

Die Konstante C_S ist ein notwendiger Gewichtungsfaktor und wird größer eins gewählt, um die Schwelle allgemein etwas anzuheben. Der genaue Wert wird hinsichtlich der Anzahl falsch erkannter Onsets heuristisch ermittelt.

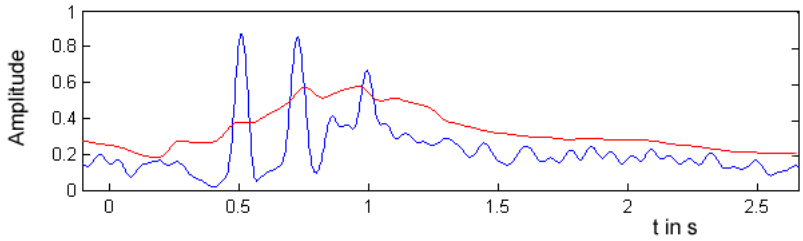

Abbildung 3.21: Dynamische Schwelle (moving-average) mit C=1.4

Dieser Algorithmus wird unter dem Stichwort *moving-average* in vielen Quellen verwendet [DDS03] [BS03] und stellt ein einfaches und naheliegendes Verfahren dar, sich an die allgemeine Dynamik des Musiksignals anzupassen. Partielle Schwankungen werden aufgrund der gewichteten Schwellenwertkurve unterdrückt, wohingegen steil ansteigende und ausgeprägte Maxima über diese Schwelle hinausstechen.

Kapitel 4

Entwicklung des optimierten Algorithmus

Im Folgenden wird auf der Basis der Untersuchungen des vorigen Kapitels ein optimierter Algorithmus zur automatischen Erkennung von Onsets für Beattracking-Systeme entwickelt. Ein maßgebliches Optimierungskriterium ist hierbei, dass das System unabhängig von der Art des Musiksignals arbeitet und sich somit automatisch an dieses Signal anpasst.

Hierzu werden die untersuchten Verfahren zur automatischen Onset-Erkennung im Hinblick auf die Zielsetzung dieser Arbeit (siehe Abschnitt 1.2) verglichen und der oder die günstigsten Ansätze als Basisalgorithmus ausgewählt. Dieser bildet die Ausgangslage für die schrittweise Erweiterung und Optimierung, an deren Ende ein robustes System erwartet wird, das aus einem unbekannten Musiksignal folgende für das Beattracking wichtige Informationen liefert: Den rhythmisch relevanten Onset-Zeitpunkt, einen allgemeinen Intensitätswert und einen Wert, der einen Hinweis über die Relevanz im Hinblick des Taktes liefert. Näheres zu dem Ausgabeformat folgt in Abschnitt 4.4.1.

4.1 Auswahl des Basisalgorithmus

4.1.1 Zeit- versus Frequenzbereich

Zunächst ist eine Entscheidung zwischen Zeit- oder Frequenzbereich zu treffen. Im Kapitel 3 sind dazu die allgemeinen Fähigkeiten bzw. Stärken und Schwächen der jeweiligen Betrachtungsweise des Musiksignals aufgeführt worden. Frequenzbasierte Verfahren berücksichtigen die spektralen Eigenschaften der Signale und sind im Gegensatz zu zeitbasierten unabhängig von markanten Spitzen im Hüllkurvenverlauf. Für die automatische Onset-Erkennung sind besonders komplexe Musiksignale mit stark überlappenden Klangereignissen problematisch. Hier ist es nicht ausreichend, lediglich die Informationen der Hüllkurvenverläufe auszuwer-

ten. Da die Zielsetzung ein von der Komplexität des Musiksignals unabhängiges System fordert, sind zeitbereichsbasierte Verfahren nicht robust genug. Aus diesem Grund wird zur Erkennung der Onsets in dieser Arbeit ein frequenzbasiertes Verfahren gewählt.

4.1.2 Vergleich der Verfahren im Frequenzbereich

In Abschnitt 3.4 sind mehrere Verfahren vorgestellt worden, die Onsets in markanten spektralen Änderungen auszumachen. Hierzu werden meist die Verläufe der Betragsspektren, aber auch Phaseninformationen ausgewertet. Es ist wenig anschaulich und aufschlussreich, die Verfahren auf rein theoretisch mathematische Weise zu vergleichen, da sie sich von der Grundidee zum Teil sehr ähnlich sind. Weiterhin ist es schwer einzuschätzen, ob sich die Betrachtung weiterer Informationen wie beispielsweise die des Phasenspektrums in jedem Fall positiv auf die Ergebnisse auswirkt. Zum einen sind Phaseninformationen besonders bei komplexen Musiksignalen nicht immer aussagekräftig, zum anderen stellt sich die Frage, ob diese Informationen für die Wahrnehmung von Onsets überhaupt relevant sind.

Aus diesem Grund werden die Verfahren anhand konkreter Ergebnisse - genauer anhand mit MATLAB berechneter Onsetkurven unterschiedlicher Musiksignale - verglichen. Die Testsignale sind eine Teilmenge der später in Kapitel 5 zur Evaluierung verwendeten Testdatenbank und stellen eine Auswahl als für die Onset-Erkennung problematisch und unproblematisch erachteter Musiksignale dar. Die Tests werden mit einer FFT-Größe von 1024 Abtastwerten, hanning-gefenstert und mit einer Schrittweite von 64 Abtastwerten durchgeführt. Im Folgenden werden die Ergebnisse zusammengefasst dargeboten.

4.1.2.1 Zu unproblematischen Testsignalen

Als unproblematisch für die Onset-Erkennung werden zum einen monophone Signale erachtet, dessen Klangereignisse deutliche Anschläge aufweisen und sich nicht überlappen, zum Beispiel hintereinander laut gespielte Klaviertöne. Des Weiteren sollten perkussive Elemente wie beispielsweise bei Schlagzeug-basierter Popmusik gut erkannt werden. Die Abbildungen 4.1 und 4.2 zeigen den Vergleich der in Abschnitt 3.4 vorgestellten frequenzbasierten Verfahren anhand der resultierenden Onsetkurven. Zur Orientierung sind die Onset-Zeitpunkte in der Darstellung des Zeitsignals markiert.

Im Vergleich der Verfahren schneiden hier die MKL-Distanz (Abschnitt 3.4.1.2) und die Distzanz nach Goto (Abschnitt 3.4.1.3) am besten ab. Beide weisen klare mit den Onsets korrespondierende Spitzen auf. Bei der Euklidischen Distanz sind die Maxima weniger eindeutig. So fällt beispielsweise auf, dass sich häufig mehrere lokale Maxima um den Zeitpunkt des Onsets befinden (siehe Abbildung 4.1). Je nach Definition der Schwelle kann dies zu Fehlerkennungen führen.

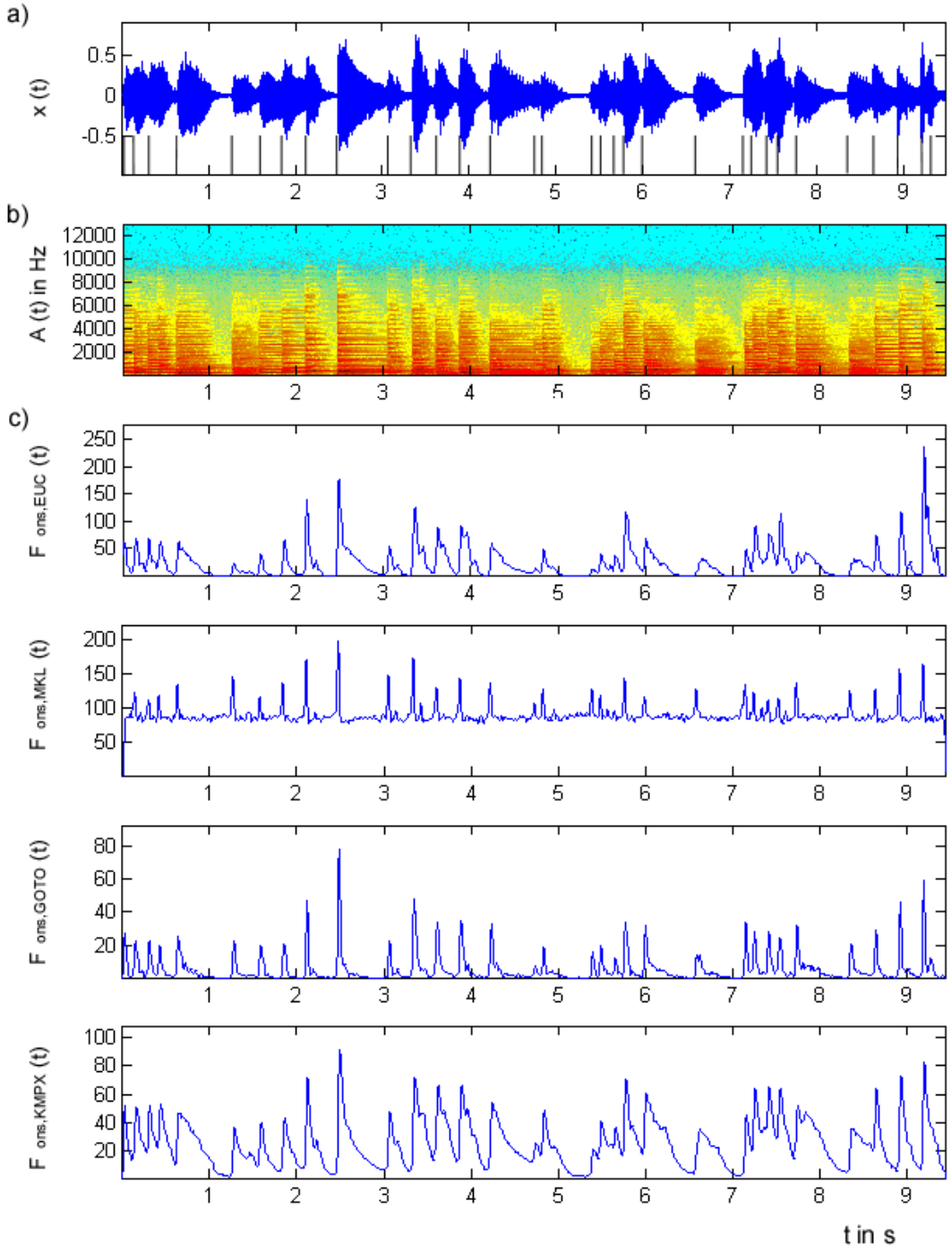

Abbildung 4.1: Ergebnisse einer Klaviertonfolge. a) Zeitsignal mit markierten Onset-Zeiten, b) Spektrogramm, c) Onsetkurven verschiedener Verfahren

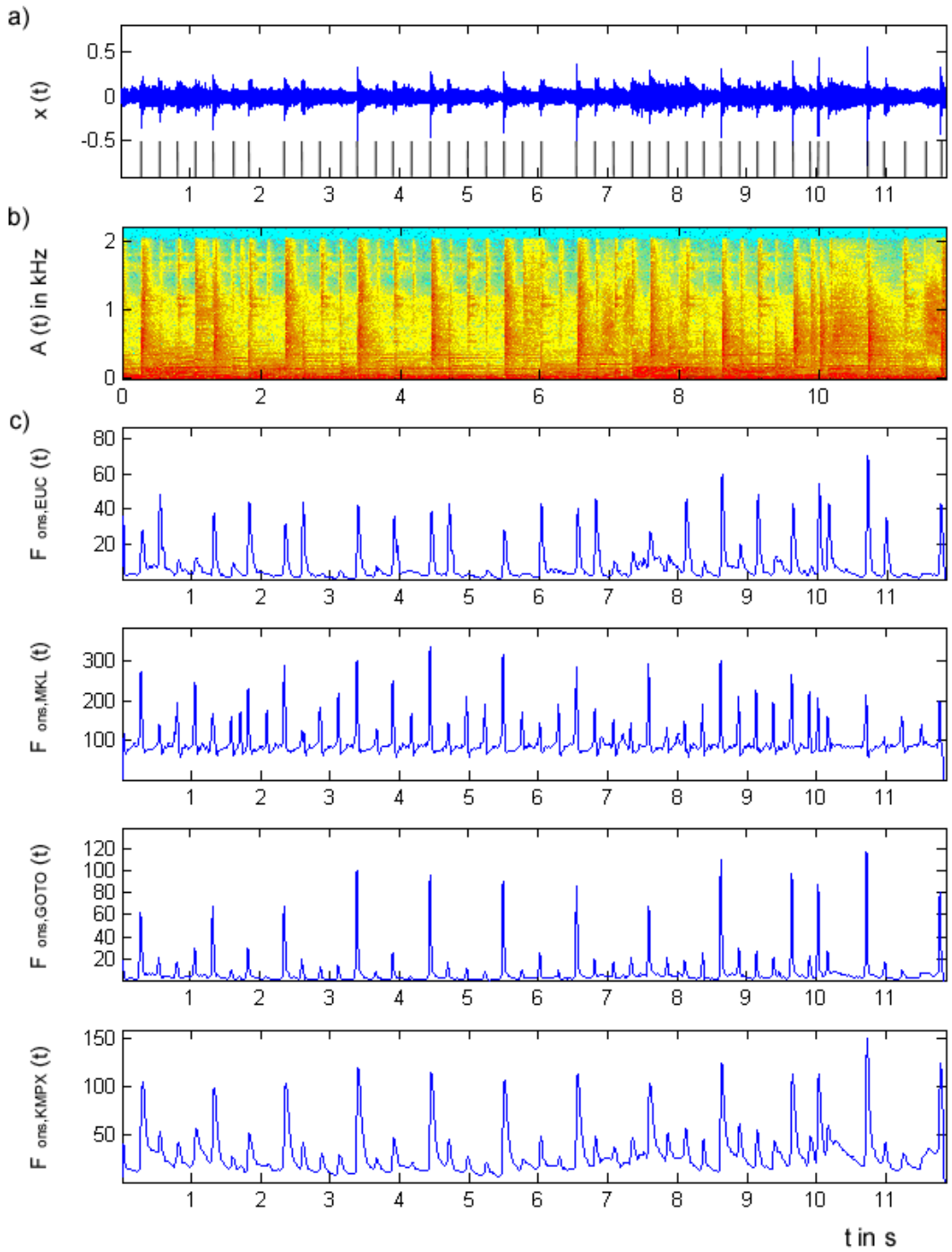

Abbildung 4.2: Ergebnisse eines Auszugs von schlagzeugbasierter Popmusik. a) Zeitsignal mit markierten Onset-Zeiten, b) Spektrogramm, c) Onsetkurven verschiedener Verfahren

Das Verfahren in der komplexen Ebene ist dem der Euklidischen Distanz wie schon in Abschnitt 3.4.2 beschrieben sehr ähnlich. Die zusätzliche Betrachtung der Phase wirkt sich positiv auf die Ausprägung der Peaks aus. Verglichen mit den zuvor betrachteten Verfahren, weisen die Peaks der komplexen Distanz jedoch einen weniger steilen Anstieg und einen deutlich flacheren Abfall auf. Dies entsteht durch unterschiedliche Summierung der Differenzen zwischen den zeitlich aufeinander folgenden Frequenz-Bins. Berücksichtigen MKL- und Goto-Distanz lediglich positive Differenzen, werden bei den anderen beiden Verfahren durch die Berechnung der Euklidischen Abstände (in der komplexen Ebene bzw. im Betragsspektrum) sämtliche Differenzbeträge aufsummiert. Dies kann sich nachteilig auf das Ergebnis auswirken, indem die Schwellen aufgrund des höheren Integrals der Onsetkurve zu stark gewichtet und somit schwächere - jedoch im Bezug zu einem Onset stehende Maxima - durch das Peakpicking unterdrückt werden. Abbildung 4.3 zeigt beispielhaft den unterschiedlichen Anstieg/Abfall eines Maximums von Goto- und komplexer Distanz und die Auswirkungen auf die Gewichtung der Schwelle eines dynamischen Peakpicking-Algorithmus.

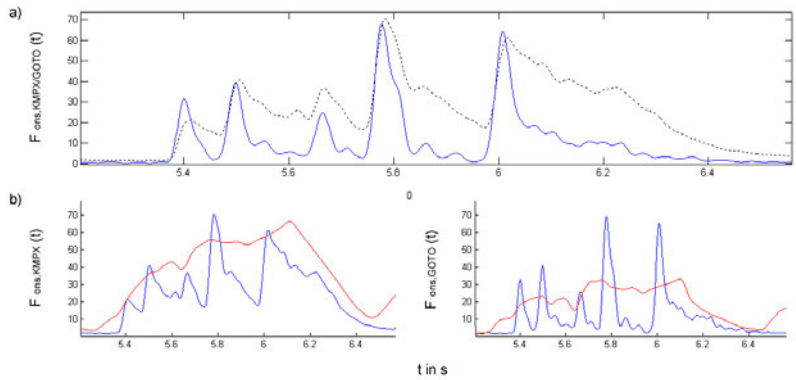

Abbildung 4.3: a) Unterschiedlicher Anstieg/Abfall bei komplexer und Goto-Distanz, b) Auswirkungen auf die dynamische Peakpicking-Schwelle

4.1.2.2 Zu problematischen Testsignalen

Mit steigender Komplexität des Musiksignals wird es für die automatische Onset-Erkennung zunehmend schwieriger, die einzelnen Klangereignisse im Signal auszumachen. Die Ereignisse können sich überlagern, so dass ein Einsatzpunkt durch stationäre oder ausschwingende Phasen höherer Energie maskiert werden oder Übergänge zu neuen Ereignissen durch die starke Vermischung der spektralen Komponenten verwischen. Typische Reaktionen der einzelnen Verfahren auf diese Testsignale zeigen die folgenden Abbildungen.

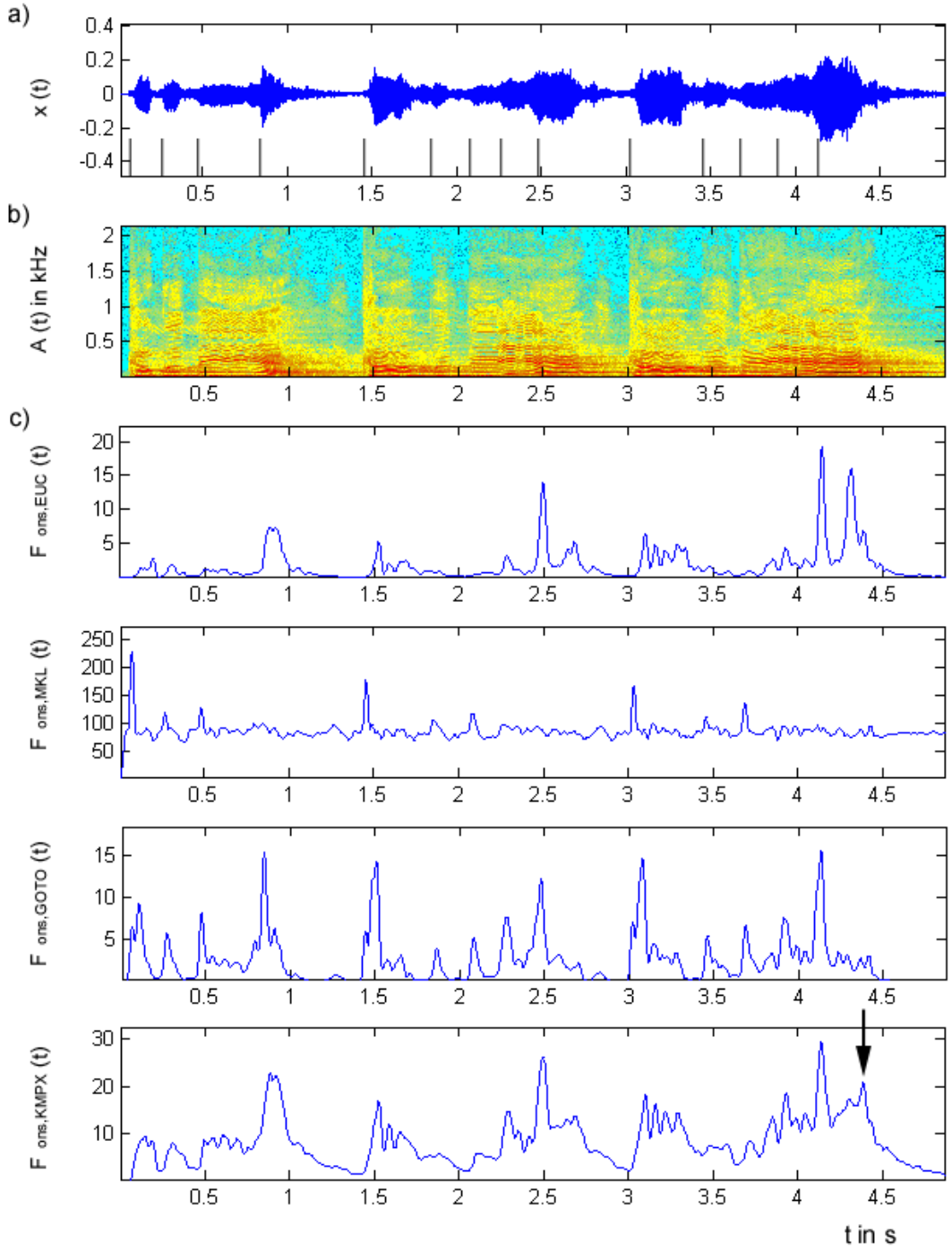

Abbildung 4.4: Ergebnisse eines stark verhallten Gesangssignals. a) Zeitsignal mit markierten Onset-Zeiten, b) Spektrogramm, c) Onsetkurven verschiedener Verfahren

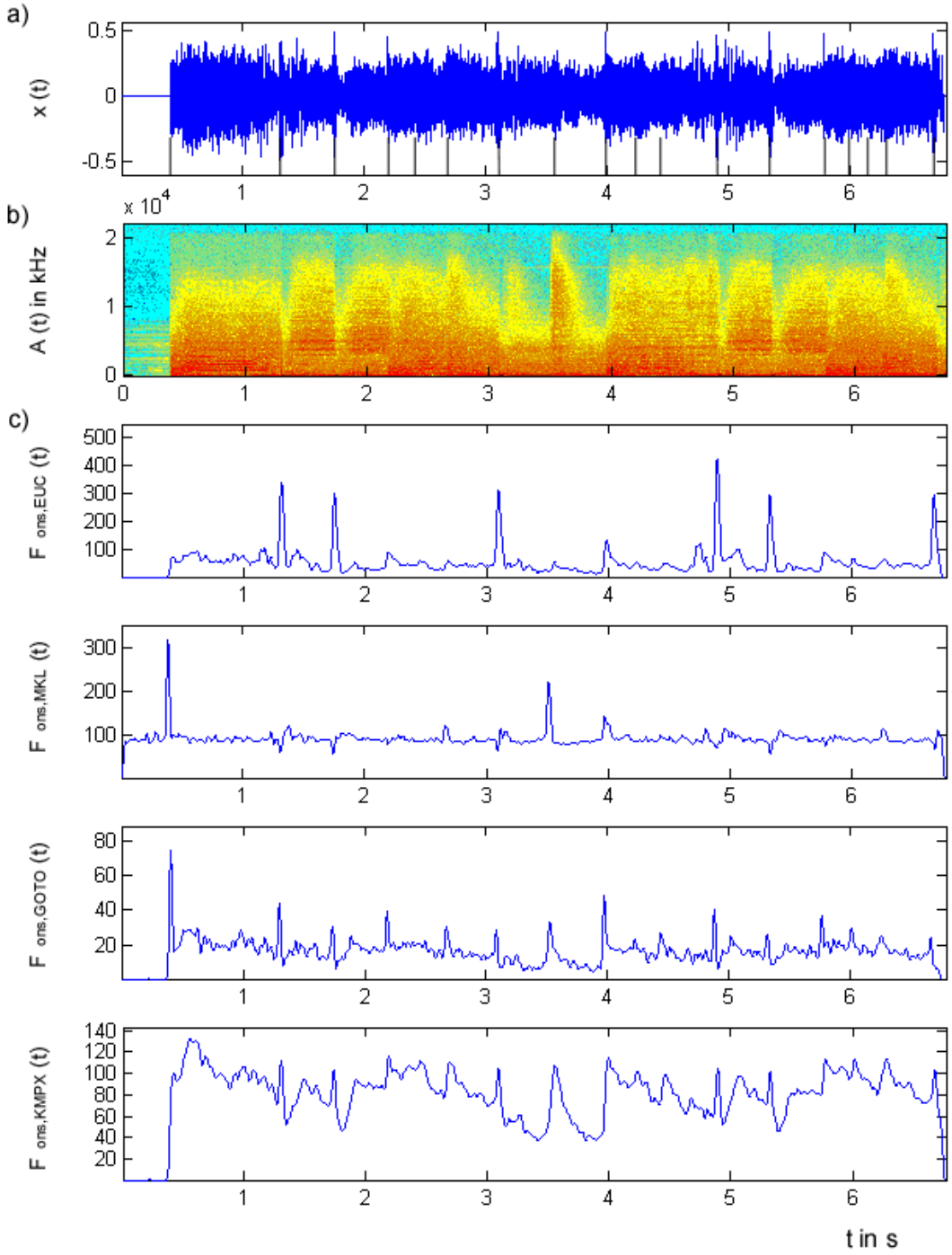

Abbildung 4.5: Ergebnisse eines Auszugs aus komplexer Popmusik. a) Zeitsignal mit markierten Onset-Zeiten, b) Spektrogramm, c) Onsetkurven verschiedener Verfahren

In den Abbildungen 4.4 und 4.5 sind die Auswirkungen aufgrund der komplexen bzw. Euklidischen Distanz noch deutlicher zu erkennen. In Abbildung 4.4 fällt in diesem Zusammenhang ein weiterer Nachteil des komplexen Verfahrens auf. Betrachtet man die Onsetkurve der komplexen Distanz, so befindet sich dort ein Maximum (durch Pfeil markiert), das nicht auf den Beginn, sondern auf das Ende eines Klangereignisses fällt. Der Abfall der spektralen Energie, der sich in einer negativen Summe der Distanzen zwischen den zeitlich aufeinander folgenden Frequenz-Bins äußert (vergleiche Abbildung 4.6), wird durch das Verfahren als positive Distanz gewertet und dadurch fälschlich für einen Onset gehalten.

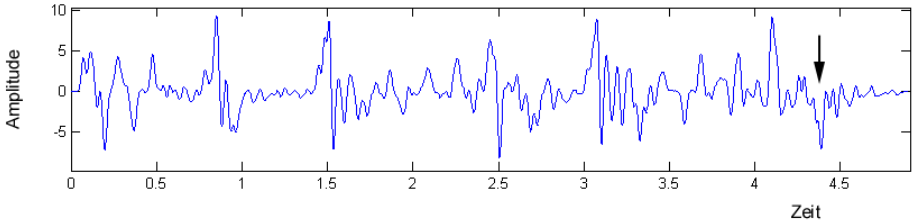

Abbildung 4.6: Verlauf der Summe der Distanzen im Betragsspektrum

Eine weitere wichtige Erkenntnis aufgrund der Untersuchung der problematischen Musiksignale ist, dass das MKL-Verfahren bei diesen Testsignalen sehr schlecht abschneidet, dessen Fähigkeiten somit stark von der Komplexität des Eingangssignals abhängen. Im Sinne der Zielsetzung kommt deshalb dieses Verfahren als Basisverfahren für das System dieser Arbeit nicht in Frage.

Der im Durchschnitt robusteste Ansatz ist die Distanzbildung nach Goto (Abschnitt 3.4.1.3). Vor allem gegenüber den problematischen Testsignalen weisen die Maxima seiner Onsetkurven eine hohe Übereinstimmung mit den Onsets des Musiksignals auf. Für diese Arbeit wird dieses Verfahren als Basisalgorithmus der nachfolgenden Optimierung ausgewählt.

4.1.3 Struktur des Basisalgorithmus

Für das Verfahren nach Goto hat sich die zum Vergleich der Verfahren voreingestellte Parametrisierung der FFT als günstigster Kompromiss erwiesen. Eine kleine Optimierung erbrachte noch die Verwendung eines Blackman- anstelle des Hanningfensters. In der Arbeit wir von einer Abtastrate von 44.1 kHz ausgegangen. Die Parametrisierung des Systems wird im Allgemeinen jedoch so konzipiert, dass die Parameter später noch in Abhängigkeit von der Abtastrate angepasst werden können. Bevor das Zeitsignal in das Betragsspektrum umgewandelt wird, erfolgt ein kurzes Preprocessing. Hier werden Stereosignale nach mono konvertiert und von dem Eingangssignal der Mittelwert abgezogen, um möglichst die Gleichanteile zu entfernen. Als grundlegenden Peakpicking-Algorithmus wird das

dynamische moving-average-Verfahren ausgewählt. Das Blockbild des Basissystems wird in Abbildung 4.7 dargestellt.

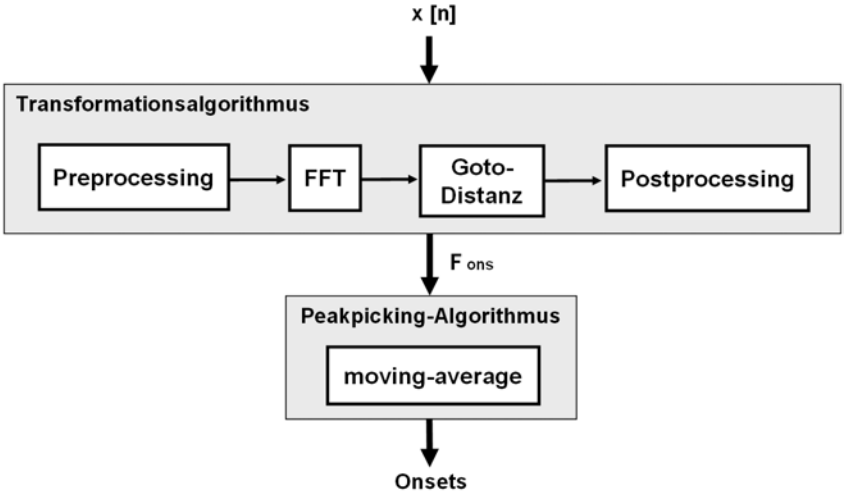

Abbildung 4.7: Blockbild des Basisalgorithmus

4.2 Onset-Erkennung in Frequenzbändern

Der Kompromiss aus Zeit- und Frequenzauflösung sowie der Rechenperformanz hat zur Folge, dass tiefere Frequenzen im Betragsspektrum zum Teil unterrepräsentiert werden und sich die spektralen Änderungen dort sehr undeutlich in den Onsetkurven abzeichnen. Für das gewählte Basisverfahren wird eine Frequenzauflösung von ca. 43 Hz festgelegt. Diese erweist sich in den meisten Fällen für eine Onset-Erkennung als ausreichend. Um die Robustheit zu steigern, soll jedoch genauer auf unterschiedliche spektrale Bereiche eingegangen - die Frequenzauflösung sozusagen erhöht werden, ohne die FFT-Größe zu verändern. In einem ersten Optimierungschritt wird das Musiksignal deshalb in einzelne Frequenzbänder aufgeteilt und Transformation in eine Onsetkurve mit anschließendem Peakpicking für jedes Band separat durchgeführt. Für das FFT-basierte Verfahren bedeutet dies, den Vektor der Frequenz-Bins in einzelne Bereiche einzuteilen und die Distanzen entlang der Zeit jeweils separat für diese Bereiche zu berechnen.

4.2.1 Das Problem der Bandaufteilung

Das entscheidende Problem ist die Festlegung der Bandanzahl und Bangrenzen. In der Literatur existieren sowohl zeitbasierte Verfahren [Kla99] [MM02] als auch frequenzbasierte Ansätze [Kla03] [Got01], die zur Onset-Erkennung das Distanzmaß in mehreren Frequenzbändern betrachten. Viele teilen das Signal in drei oder vier Bänder mit festgesetzten Bandgrenzen gemäß einer groben Einteilung nach

Höhen, Mitten und Tiefen auf. Dies war auch in dieser Arbeit der erste Ansatz, der allerdings zu unbefriedigenden Ergebnissen führte.

Wie in den Grundlagen zum Onset-Begriff (Abschnitt 2.1) diskutiert, ist es hilfreich, sich zur Erkennung der wahrgenommenen Onset-Zeiten an der menschlichen Empfindung zu orientieren und diese zu modellieren. Der in Abschnitt 3.3.4 vorgestellte Ansatz von Klapuri [Kla99] orientiert sich an der Psychoakustik und teilt das Signal in 21 Bänder gemäß den kritischen Bandbreiten des menschlichen Gehörs (vergleiche [ZF99]) ein. Dadurch wird angestrebt, sich der menschlichen Wahrnehmung von Onsets anzunähern. In [Reu95] werden Einschwingvorgänge aus psychoakustischer Sicht folgendermaßen beschrieben:

> „Gerade bei der Wahrnehmung von kurzzeitigen Vorgängen zeigt sich, dass das Ohr seine Verarbeitungsweise an die jeweiligen Wahrnehmungsverhältnisse angleicht, um ein Optimum an Informationen aus dem Schallsignal zu gewinnen: Unterliegt der Klang schnellen Veränderungen (wie z.B. im Einschwingvorgang), so vergrößert sich die zeitliche Auflösung des Gehörs, worunter dann bedingt durch den Aufbau der Frequenzgruppenbreiten die Auflösung im Frequenzbereich leidet. Umgekehrt verhält sich dies bei der Wahrnehmung von quasistationären Zuständen, wo die zeitliche Auflösung geringer ist, der Klang über größere Zeitabschnitte integriert wird, die Frequenzauflösung aber sehr genau wird.“

Demnach stellt sich die Frage, ob für die Onset-Wahrnehmung die Frequenzauflösung von Klapuri's Bandeinteilung nicht zu fein gewählt ist und die starre Festlegung der Bangrenzen den Wahrnehmungsverlauf adäquat nachbildet. Hierzu soll die signalabhängige Anpassung der Auflösung des Gehörs bei der Wahrnehmung von Klangereignissen betrachtet werden. Nach [Reu95] kann man diese durch drei Phasen beschreiben. In der ersten Phase von ca. 10 ms werden die Frequenzgruppenbreiten aufgebaut, wobei ein Übergang vom breitbandigen zum schmalbandigen Hörverhalten stattfindet. Die Mittenfrequenzen der kritischen Bandbreiten richten sich hierbei nach den spektralen Eigenschaften der gehörten Klangereignisse. Danach erfolgt eine deutliche Zunahme der Frequenzempfindlichkeit, während die zeitliche Auflösung abnimmt. Diese Phase liegt im Bereich von 50-55 ms und bildet in etwa die Schwelle zur Erkennung der Tonhöhe. Diese steht in Beziehung zur der Verwischungsschwelle, unter der diskrete Klangereignisse als Gesamtereignis wahrgenommen werden (vergleiche Abschnitt 2.1.2). In der letzten Phase (ca. 250 ms) erreicht das Gehör seinen „eingeschwungenen“ Zustand, ab dem keine weitere Zunahme der Empfindlichkeit mehr stattfindet.

Einschwingvorgänge natürlicher Instrumente sind meist bedeutend kürzer als die Zeit, nach der das Gehör in voller Empfindlichkeit ein Klangereignis wahrnimmt. Die Frequenzgruppenbreiten sind zwar bereits aufgebaut, die Empfindlichkeit und Frequenzselektivität des Gehörs bilden sich jedoch erst mit der weiteren Zeit und in Abhängigkeit des Signals aus.

Klapuri's Ansatz führt deshalb nicht zwangsläufig zu guten Ergebnissen, da zum einen die nachfolgende Bandzusammenfassung durch die hohe Anzahl an Bändern erschwert wird und zum anderen die gleich gewichtete Auswertung aller Bänder zu Fehlerkennungen führen kann. Steigt beispielsweise die Amplitude in einem Frequenzband, auf das sich das Gehör in den genannten Phasen nicht optimiert hat, kann dieser Anstieg dennoch als Onset gewertet werden. Abbildung 4.8 zeigt ein Beispiel anhand der Onsetkurven der beiden tiefsten Bänder (nach Klapuri) eines Gesangssolos. Die Kurven der tiefsten beiden Bänder zeigen lokale Maxima, die nicht mit Onsets korrespondieren. Einige fallen sogar auf das Ereignisende. Ein automatisches Beattracking-System könnte durch diese Fehlerkennungen stark irritiert werden.

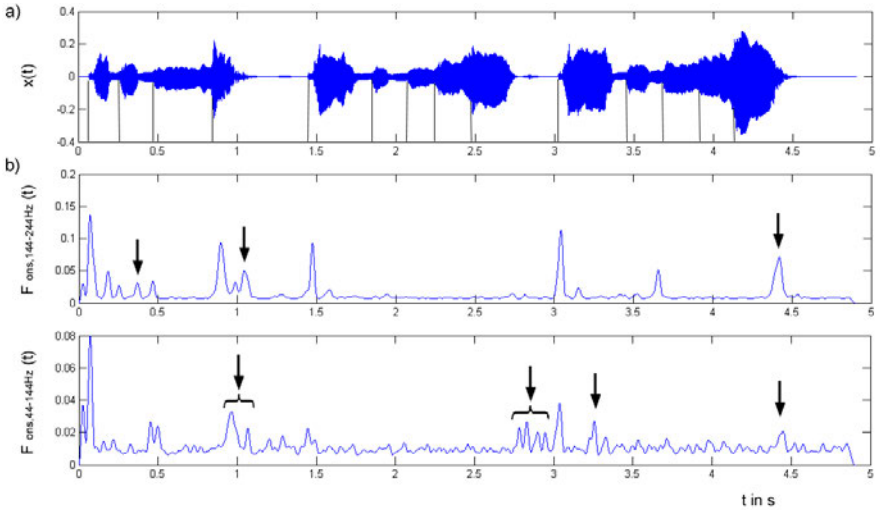

Abbildung 4.8: Fehlerkennung von Onsets in einem Gesangssolo aufgrund Klapuri's Frequenzbandeinteilung: a) Spektrogramm des Signals, b) Onsetkurven der beiden tiefsten Bänder, mögliche Fehlerkennungen (Pfeile)

Die Bandaufteilung in dieser Arbeit soll deshalb nicht a priori festgelegt sein, sondern sich an das jeweilige Eingangssignal anpassen. Es wird weiterhin eine breitbandigere Verarbeitung der musikalischen Ereignisse modelliert, in Orientierung daran, dass die Empfindlichkeit der schmalbandigen Verarbeitung im Gehör erst dann ausgeprägt ist, wenn das Ereignis bereits seine quasistationäre Phase erreicht hat. An dieser Stelle jedoch stößt das wahrnehmungsorientierte Modell dieser Arbeit an seine Grenzen, denn es ist weder in der Literatur beschrieben, wie empfindlich genau die Frequenzauflösung im Gehör zum Zeipunkt eines Onsets eingestellt ist, noch erlaubt der Rahmen dieser Arbeit eine eigene psychoakustische Untersuchung. Die Einteilung in die Frequenzbänder in Anpassung an das Eingangssignal orientiert sich aus diesem Grund an den spektralen Eigenschaften des Musiksignals selbst.

4.2.2 Bandaufteilung auf Basis von Feature-Extraktion

Das Music Information Retrieval basiert auf der Extraktion bestimmter Merkmale (engl. features) des Musiksignals. Die Genre-Klassifikation beispielsweise versucht, das Musiksignal anhand der charakteristischen Ausprägung bestimmter Features einem Genre zuzuordnen. Für die Bandaufteilung sollen die Parameter Bandanzahl und Bandgrenzen aufgrund von Features aus dem Eingangssignal eingestellt werden.

Im Frequenzbereich existieren eine Vielzahl von Features, die charakteristische Eigenschaften des Signals beschreiben. Zur Festlegung der Bandgrenzen soll ein Feature verwendet werden, das eine Aussage über die Verteilung der spektralen Energie des Signals liefert. Hierzu wird der *spektrale Zentroid (engl. spectral centroid (sc))* ausgewählt, der den Mittelpunkt der spektralen Verteilung innerhalb eines FFT-Blocks berechnet. Gelegentlich wird dieser auch als Balance-Punkt bezeichnet. Die Berechnungsformel lautet folgendermaßen:

$$sc = \frac{\sum_{k=0}^{\frac{N}{2}-1} k \cdot X[k]}{\sum_{k=0}^{\frac{N}{2}-1} X[k]}. \tag{4.1}$$

N: FFT-Größe
k: Index des Frequenz-Bins

Den spektralen Zentroid eines zeitlichen Signalabschnitts msc erhält man durch den Mittelwert der Zentroide der aufeinander folgenden FFT-Blöcke

$$msc = \mu(sc_i), \quad sc_i \in [0, L]. \tag{4.2}$$

L: Länge des Signalsabschnitts
i: Index des FFT-Blocks

Für den Algorithmus zur Berechnung der Bandgrenzen wird der Begriff des spektralen Zentroids erweitert, so dass der msc nicht nur für den gesamten Frequenzbereich, sondern auch bis zu einer bestimmten Grenze innerhalb des Frequenz-Bin-Vektors berechnet werden kann. Die Formel aus Gleichung 4.1 ändert sich dabei lediglich durch die Anpassung der oberen Grenze der Summen an die jeweilige festgesetzte Bereichsgrenze. Die Berechnung der einzelnen Bandgrenzen erfolgt iterativ, indem zunächst der gesamte Frequenz-Bin-Vektor als Bereich festgelegt wird. Der resultierende msc legt dann die Bereichsgrenze der folgenden Iteration fest usw. Der Algorithmus endet an einer Grenze, die durch eine minimale Anzahl den Bereich repräsentierender Frequenz-Bins festgelegt ist.

Diese Einteilung der Bandgrenzen stellt eine gute Anpassung an die spektralen Eigenschaften des Eingangssignals dar und hat weiterhin den Effekt, dass die Energieverhältnisse zwischen den einzelnen Bändern in etwa ausgeglichen sind. Bei weiterer Untersuchung fällt jedoch auf, dass es sich mitunter positiv auf die Ergebnisse auswirkt, einzelne Bänder wieder zusammenzufassen. Besonders bei

Solostimmen mit starken harmonischen Anteilen erzeugt die bandweise Betrachtung zum Teil redundante Information. Nach Zusammenfassung dieser redundanten Bänder erhält man im Gesamtband mitunter sogar klarere Ergebnisse. Aus diesem Grund wird der obige Algorithmus insofern erweitert, dass in jedem Iterationsschritt eine Entscheidung gefällt wird, ob der betrachtete Frequenzbereich durch die berechnete Bandgrenze aufgeteilt wird. Fällt die Entscheidung negativ aus, so stellt der aktuell errechnete *msc* dennoch die Grenze des nächsten zu betrachtenden Bereichs dar.

Als Entscheidungskriterium dient ein weiteres spektrales Feature, das eine ungefähre Aussage über die Anzahl der Stimmen im Musikstück liefert. Man muss hierbei feststellen, dass es kein einzelnes Feature gibt, das eine derartige Aussage eindeutig treffen kann, da es sich keineswegs um ein triviales Problem handelt. In der Literatur wird die *spektrale Glätte (spectral smoothness (ss))* als guter Indikator für die Anzahl der an einem komplexen Musikstück beteiligten Stimmen angesehen [Kla01]. Dieses Feature ist psychoakustisch motiviert und wird auf die Organisationsstrategie des auditiven Systems [Bre90] zurückgeführt, ähnliche spektrale Reize zusammenzufassen und als ein Gesamtereignis zu betrachten [Rie03]. Die Untersuchung dieses Features hinsichtlich der Anzahl der Stimmen ergeben jedoch wenig klare Aussagen. Dem Verlauf der spectral smoothness oder seinem Mittelwert über der Zeit können keine regelmäßigen Reaktionen auf komplexe oder einstimmige Musikstücke entnommen werden (vergleiche Abbildung 4.9 a)).

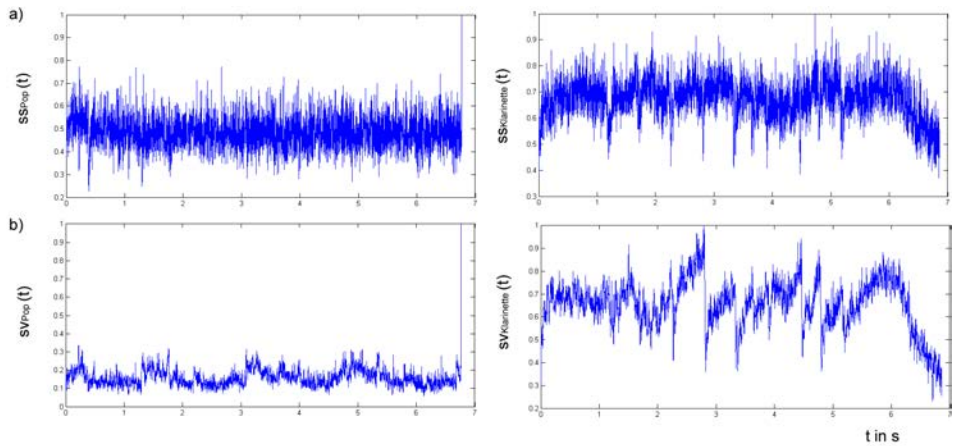

Abbildung 4.9: Vergleich zwischen a) spectral smoothness und b) spectral variance an komplexer Popmusik (links) und Klarinettensolo (rechts)

Eine deutlichere Aussage ist einem anderen Feature, der *spekralen Varianz (engl. spectral variance (sv))*, zu entnehmen. Besteht ein Signal hauptsächlich aus harmonischen Klangereignissen und sind deren Verläufe im Betragsspektrum des Musiksignals deutlich auszumachen, weist die spektrale Varianz im zeitlichen Mittel einen sehr hohen Wert auf. Sind die Klangereignisse jedoch stark vermischt

oder rauschhaft, wird ein niedriger Wert verzeichnet (siehe Abbildung 4.9 b)). Die spektrale Varianz ist im Vergleich zur spektralen Glätte der bessere Indikator für die Komplexität, bzw. für die Anzahl der Stimmen des Musikstückes.

Sicherlich haben die Klänge von Solostimmen nicht immer stark ausgeprägte harmonische Anteile. In diesen Fällen wird das Signal fälschlicherweise für komplex erachtet. Im Mittel ergibt dieses Verfahren jedoch gute Ergebnisse, so dass die spektrale Varianz als Kriterium für die Entscheidung über die Bandaufteilung des betrachteten Frequenzbereichs herangezogen wird. Hierbei erfolgt die Aufteilung immer dann, wenn die mittlere Varianz kleiner als ein festgelegter Grenzwert (heuristisch ermittelt) ausfällt.

Mit dem spektralen Zentroid als Feature zur Ermittlung der Bandgrenzen und der spektralen Varianz zur Festlegung der Bandanzahl hat der Algorithmus folgenden Ablauf (dargestellt in Pseudocode):

```
grenze = FFT-Größe/2;
bandgrenzen = {}
while grenze > limitgrenze
  grenze = mean spectral centroid(A{1:grenze})
  varianz = mean spectral variance(A{1:grenze})
  if varianz < limitvarianz
    bandgrenzen = bandgrenzen U {grenze}
  endif
endwhile
```

4.2.3 Optimiertes Peakpicking

Für einen robusten Algorithmus ist eine dynamische Anpassung an das Eingangssignal notwendig. Das gilt auch für das Peakpicking. So kommen statische Verfahren für diese Arbeit nicht in Frage. Der hier verwendete Peakpicking-Algorithmus basiert auf dem moving-average-Verfahren (zu den Verfahren siehe Abschnitt 3.6).

Als problematisch erweist sich, einen für die Festlegung des konstanten Gewichtungsfaktors C_S (vergleiche Gleichung 3.17) geeigneten Kompromiss zu finden. Trotz der Verwendung des Goto-Verfahrens, weist das Signal der Onsetkurve zum Teil für längere Zeit hohe Amplituden auf. In diesem Fall wird ab einem bestimmten C_S die Schwelle aufgrund der gewichteten Integration der Amplituden zu hoch angesetzt. Bei durchschnittlich niedrigen Amplituden verursacht eine für den vorigen Fall optimierte Schwelle eine hohe Anzahl an Fehlerkennungen. Aus diesem Grund soll sich der Gewichtungsfaktor C_S des moving-average-Verfahrens ebenfalls dynamisch an das Signal anpassen.

Der Verlauf des optimalen Gewichtungsfaktors C_S für hohe und niedrige Amplituden der Onsetkurve lässt sich gut durch eine Sigmoidfunktion (siehe Abbildung 4.10) modellieren. Der jeweilige Faktor $C_S[n]$ berechnet sich somit aus der

Peakpicking-Schwelle $\mu(w_n)$ folgendermaßen:

$$C_S[n] = \frac{1}{1 + e^{\mu(w_n)}} + u. \tag{4.3}$$

u: Verschiebungsfaktor (orientiert an den für hohe und niedrige Amplituden optimalen Gewichtungsfaktoren)

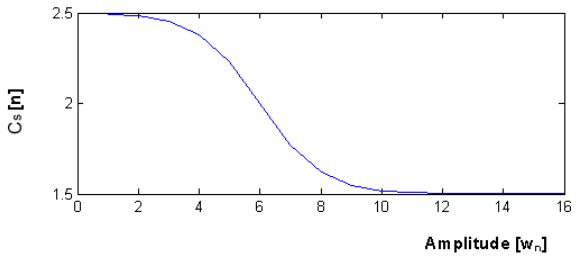

Abbildung 4.10: Sigmoide Gewichtungsfunktion

Die endgültige Schwellenwertfunktion des dynamischen Peakpicking-Algorithmus lautet im Vergleich zu 3.17 somit

$$S[n] = C_S[n] \cdot \mu(w_n), \quad w_n \in [n - \frac{L}{2}, n + \frac{L}{2}]. \tag{4.4}$$

4.2.4 Intensität des Onsets

Mit der Verwendung der Schwellenwertfunktion werden pro Frequenzband die Zeiten der herausstechendsten Maxima der Onsetkurven extrahiert. Der nächste Schritt besteht darin, pro Band die Stärke der Onsets festzulegen, um die Ausgangslage für die nachfolgende Bandzusammenfassung sowie der Bestimmung der endgültigen Intensität zu schaffen. Die Berechnung der Intensität $I_{ons}[j]$ des extrahierten Onsets an der Stelle j berücksichtigt zwei Größen, die absolute Amplitude der Onsetkurve $F_{ons}[j]$ und die Differenz $\delta_{max,min}[j]$ aus $F_{ons}[j]$ und dem vorhergehenden Minimum in der Onsetkurve $PrevMin_{F_{ons}[j]})$:

$$I_{ons}[j] = \frac{(C_I \cdot F_{ons}[j]) + \delta_{max,min}[j]}{2}. \tag{4.5}$$

$\delta_{max,min}[j] = F_{ons}[j] - PrevMin_{(F_{ons}[j])}$

Der Faktor C_I gewichtet hierbei die absolute Amplitude der Onsetkurve etwas höher. Abbildung 4.11 zeigt die Ergebnisse an einem Gesangssolo. Hierbei wurde das Signal aufgrund des Verfahrens in zwei Bänder aufgeteilt. Die unteren beiden Diagramme zeigen die berechneten Intensitäten, darüber sind die Onsetkurven der Bänder samt Peakpicking-Schwelle und $\delta_{max,min}[j]$ in Form von Dirac-Impulsen dargestellt.

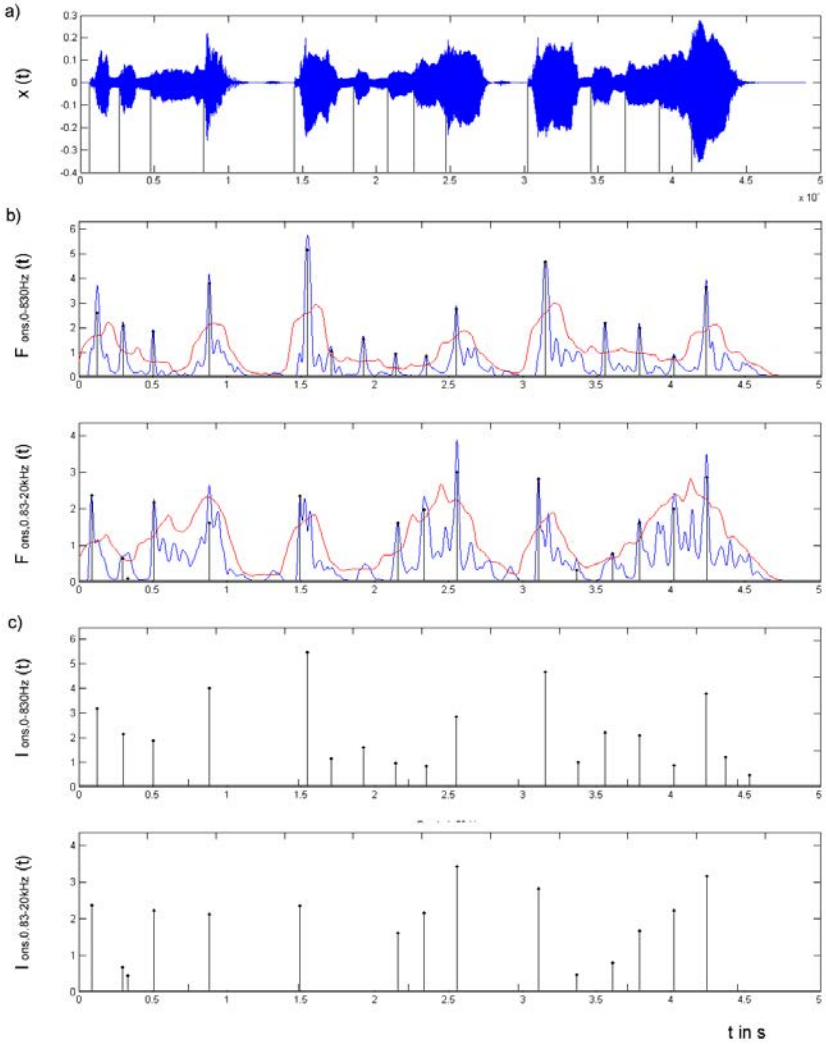

Abbildung 4.11: Ergebnisse an einem Gesangssolo. a) Zeitsignal, b) Onsetkurven mit Peakpicking-Schwelle und $\delta_{max,min}[j]$, c) Verlauf der Onset-Intensitäten

4.2.5 Zusammenfassung der Frequenzbänder

Nach der Berechnung der Onset-Zeiten und ihrer Intensitäten in jedem Frequenzband, müssen die Informationen der Teilbänder wieder zusammengefasst werden. Hierzu ist es notwendig, eine Strategie zu entwickeln, um zum einen die zeitliche Toleranz festzulegen, in der Onsets der einzelnen Bänder als zusammengehörig eingestuft werden und zum anderen, um die Intensitäten der zusammengehörigen Onsets geeignet zusammenzufassen. Initialer Schritt ist die Addition der Intensitätsfolgen der Frequenzbänder.

Für die weitere Zusammenfassung orientiert sich der Algorithmus an den Betrachtungen zur zeitlichen Auflösung des Gehörs in Abschnitt 2.1.2. So werden

die erkannten Onsets ausgehend vom größten Intensitätswert der addierten Bänder der Größe nach abgearbeitet und alle im Bereich von 50 ms vor- und nach dem aktuellen Zeitpunkt befindlichen Onset-Intensitäten diesem Wert zugeordnet (Clustering). Der endgültige Intensitätswert ergibt sich aus der Summe der geclusterten Teilintensitäten. Abbildung 4.12 zeigt die Addition der Frequenzbänder des Beispielsignals aus Abbildung 4.11 sowie das Ergebnis nach Clustering und Summierung.

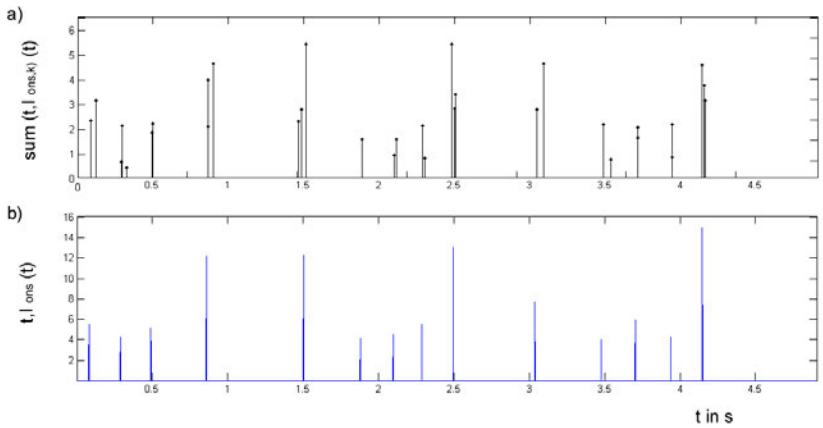

Abbildung 4.12: Verlauf der Onset-Intensitäten eines Gesangsolos: a) vor und b) nach Clusterung und Zusammenfassung

Direkt nach der Zusammenfassung werden in einem zweiten Schritt (Selektion) ausgehend vom aktuellen Onset alle im Bereich von 120 ms nachfolgenden Intensitäten, deren Wert unter einer definierten Schwelle (Abbildung 4.13) liegt, gelöscht. Die Dauer dieses Schwellenwertfensters richtet sich hierbei nach Zwicker's *Model of Rhythm* [ZF99]. Diesen Vorgang kann man auch als zweite Stufe des Peakpicking-Algorithmus auffassen. Die Schwellenwertfunktion hat bis zum Erreichen der Verwischungsschwelle (vergleiche Abschnitt 2.1.2) und etwas darüber hinaus die Form eines Rechteckfensters mit der Amplitude der Onset-Intensität und schwingt dann wie ein Hanningfenster bis zum Nullpunkt aus.

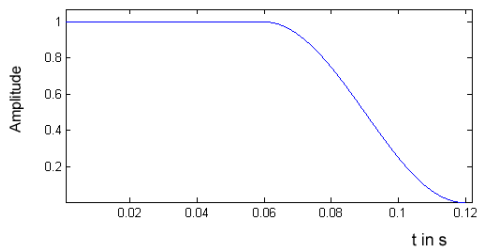

Abbildung 4.13: Schwelle für die Selektion

4.2.6 Erweiterte Struktur des Systems

Mit der automatischen Aufteilung des Signals in einzelne Frequenzbänder, des hinsichtlich der Signalamplitude optimierten Peakpicking-Algorithmus und der anschließenden Zusammenfassung der Bänder mit Hilfe psychoakustisch orientierter Schwellen, weist das Gesamtsystem nun die in Abbildung 4.14 dargestellte Architektur auf. Als Ausgabe erhält man einen Vektor $Ons[j]$ mit den Zeitpunkten $t_{ons}[j]$ und der jeweiligen Intensität $I_{ons}[j]$ der erkannten Onsets.

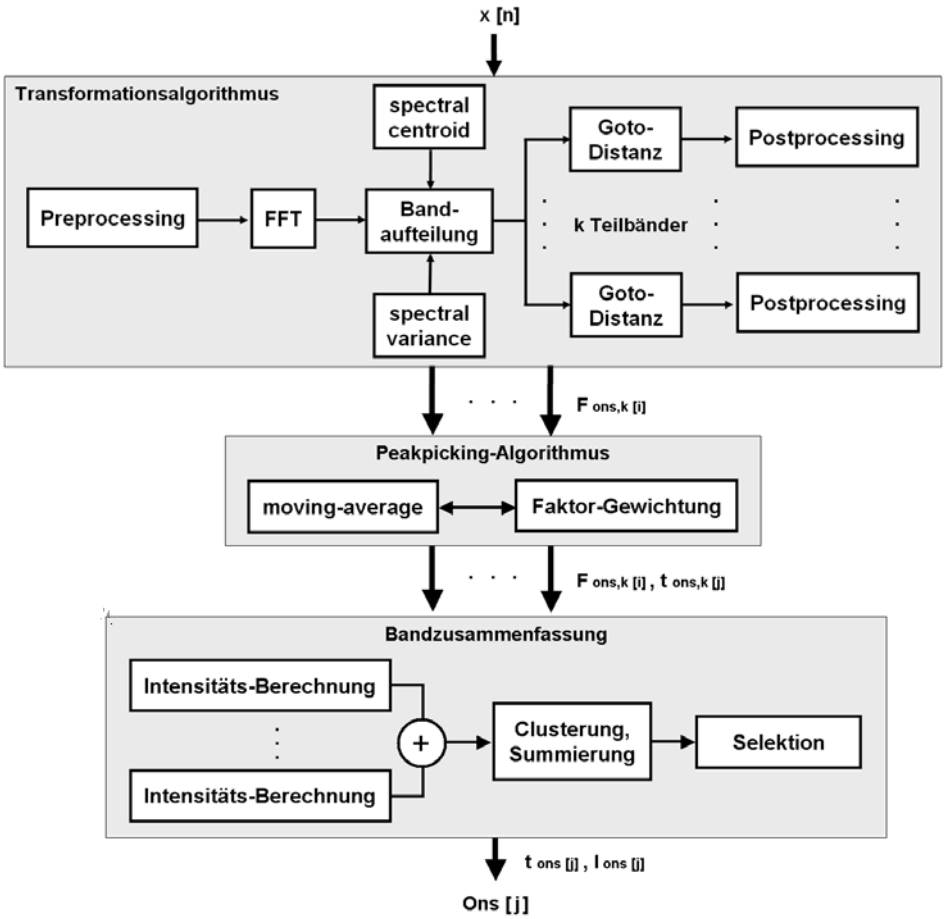

Abbildung 4.14: Blockbild des Systems nach den durchgeführten Optimierungsschritten

4.3 Onset-Klassifizierung zur Takterkennung

Im Folgenden soll aus dem Onset-Vektor, bestehend aus den Zeit- und Intensitätsinformationen der erkannten Onsets, eine weitere Information extrahiert werden, die jeden Onset hinsichtlich seiner Relevanz im rhythmischen Gefüge einstuft. Das Ziel ist, einen Hinweis auf die schwerste Zählzeit des Taktes - der „Eins“ - zu erhalten, so dass diese Information einem Beattracking-System zur Berechnung der Taktart von Nutzen sein kann.

4.3.1 Schwere Zählzeiten

Durch den Takt eines Musikstückes entsteht die Einteilung des Beats in betonte und unbetonte Schläge. Die betonten Schläge werden auch als schwere- oder Hauptzählzeiten bezeichnet. Bei einem 4/4-Takt beispielsweise stellen die schweren Zählzeiten die „Eins“ und die „Drei“ des Beats dar, wobei die „Eins“ in der Regel die größte Betonung aufweist. Die unbetonten Schläge werden auch als Nebenzählzeiten bezeichnet. Die absichtliche Abweichung von diesem Betonungsschema nennt man *Synkopieren*. In Abbildung 4.15 werden die Zählzeiten anhand eines Auszugs aus Popmusik (4/4-Takt) dargestellt. Markiert sind jeweils die schweren Zählzeiten t_{eins} und t_{drei}.

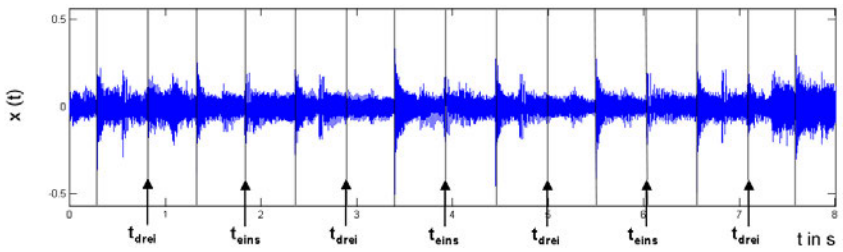

Abbildung 4.15: Zählzeiten in einem Auszug aus Popmusik: Schwere Zählzeiten durch Pfeile markiert

4.3.2 Korrelationsmethode

Die Einstufung der Onsets hinsichtlich ihrer Schwere bezüglich des Taktes erfolgt in zwei Arbeitsschritten. Im ersten wird das Onsets-Signal $Ons(t)$ auf Periodizität mit Hilfe der Autokorrelation untersucht. Um Informationen über den Takt zu erhalten ist es notwendig, das Signal in einem größeren zeitlichen Abschnitt zu betrachten. Die erforderliche Größe dieses Intervalls lässt sich durch die vereinfachende Annahme abschätzen, dass Musikstücke in der Regel keine langsameren Tempi als 60 BPM[1] aufweisen. Geht man weiterhin von einem 4/4-Takt aus, so

[1] engl. Beats Per Minute

ergibt sich eine maximale Taktlänge von vier Sekunden. Dieses Intervall wird für eine modifizierte Autokorrelation *(mAKF)* verwendet, bei der der aktuelle Block i nicht nur entlang sich selbst, sondern weiterhin über die Werte des darauf folgenden Blocks $i+1$ verschoben wird. Der Sinn steckt darin, den für die Autokorrelation typischen abfallenden Verlauf zu kompensieren, um sich eine anschließende Trendkorrektur des Ergebnisses zu ersparen. Es handelt sich also genau genommen um eine Kreuzkorrelation zwischen dem durch Anfügen einer gleichlangen Nullfolge zp erweiterten aktuellen Signalblock i und dem Gesamtsignal der beiden aufeinander folgenden Blöcke i und $i+1$. Der Lage des absoluten Maximums des relevanten Bereichs des Ergebnisses (Bereich ab Nullpunkt bis zur Länge des aktuellen Blocks i) entnimmt man die Dauer der Periodizität des Onset-Signals. Abbildung 4.16 zeigt die ersten acht Sekunden des Onset-Signals eines Auszugs aus Popmusik und das Ergebnis des ersten Arbeitsschrittes $mAKF_{(i,zp),(i,i+1)}(t)$. Hierbei sind weiterhin die Grenzen der verwendeten Blöcke und des relevanten Bereichs der $mAKF$ sowie das Maximum $t_{max,T}$ für die Bestimmung der Periodendauer T eingezeichnet.

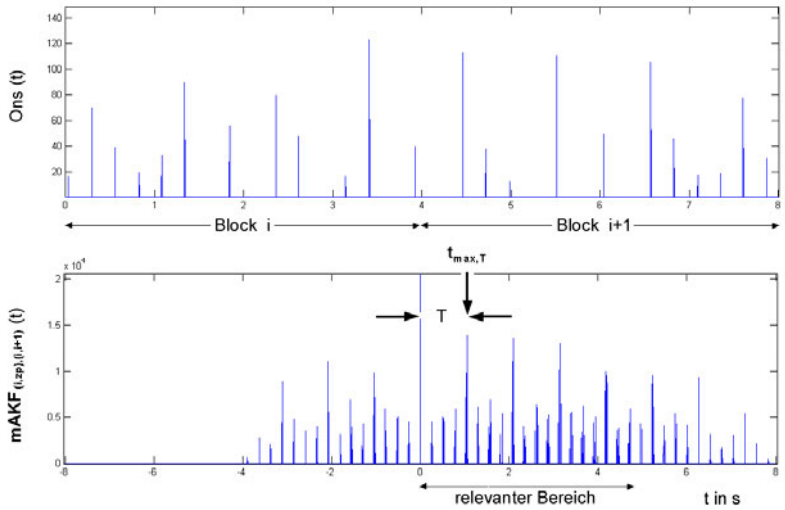

Abbildung 4.16: Ergebnisse des ersten Arbeitsschritts der Korrelationsmethode

Im zweiten Arbeitsschritt soll die Lage der betonten Onsets ermittelt werden. Hierzu wird zunächst ein Deltakamm $\delta_T(t)$ der Länge zweier Blöcke generiert, dessen Impulse im äquidistanten Abstand der im vorigen Arbeitsschritt ermittelten Periodendauer T liegen. Mit diesem Signal und dem Onset-Signal der beiden Blöcke i und $i+1$ führt man anschließend eine Kreuzkorrelation durch, in deren Ergebnis wiederum nur der bestimmte Bereich ab dem Nullpunkt bis zur Blocklänge für den aktuell betrachteten Block i relevant ist. Extrahiert man diesen Teil und normiert ihn auf die Anzahl der Impulse des Deltakamms, erhält man eine Funktion, deren Peaks die Zeitpunkte der periodisch betonten Onsets anzeigen. Nach einer Glättung kann diese Funktion für jeden Onset zur Einstufung hin-

sichtlich der Schwere im Takt verwendet werden. Abbildung 4.17 illustriert die einzelnen Ergebnisse.

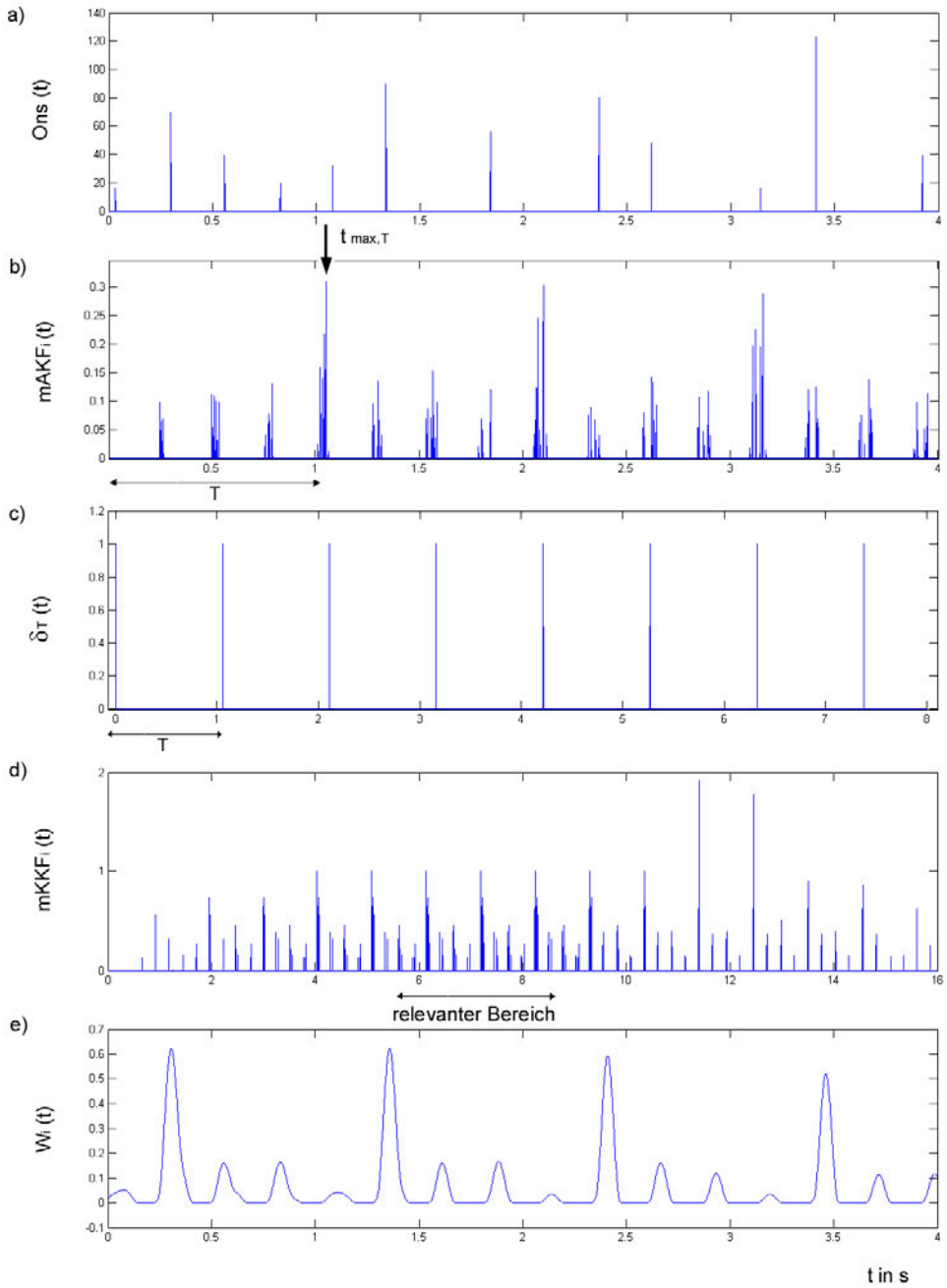

Abbildung 4.17: Ergebnisse der Korrelationsmethode an einem Auszug aus Popmusik: a) Onsetkurve, b) relevanter Bereich der mAKF, c) generierter Deltakamm, d) mKKF, e) Wahrscheinlichkeit für schwere Zählzeit

4.4 Gesamtsystem

Durch die Erweiterung mit der Stufe zur Klassifikation hinsichtlich der Schwere im Takt, hat das endgültige System die Struktur aus Abbildung 4.18. Das Datenformat der Ausgabe für ein Beattracking-System beschreibt Abschnitt 4.4.1.

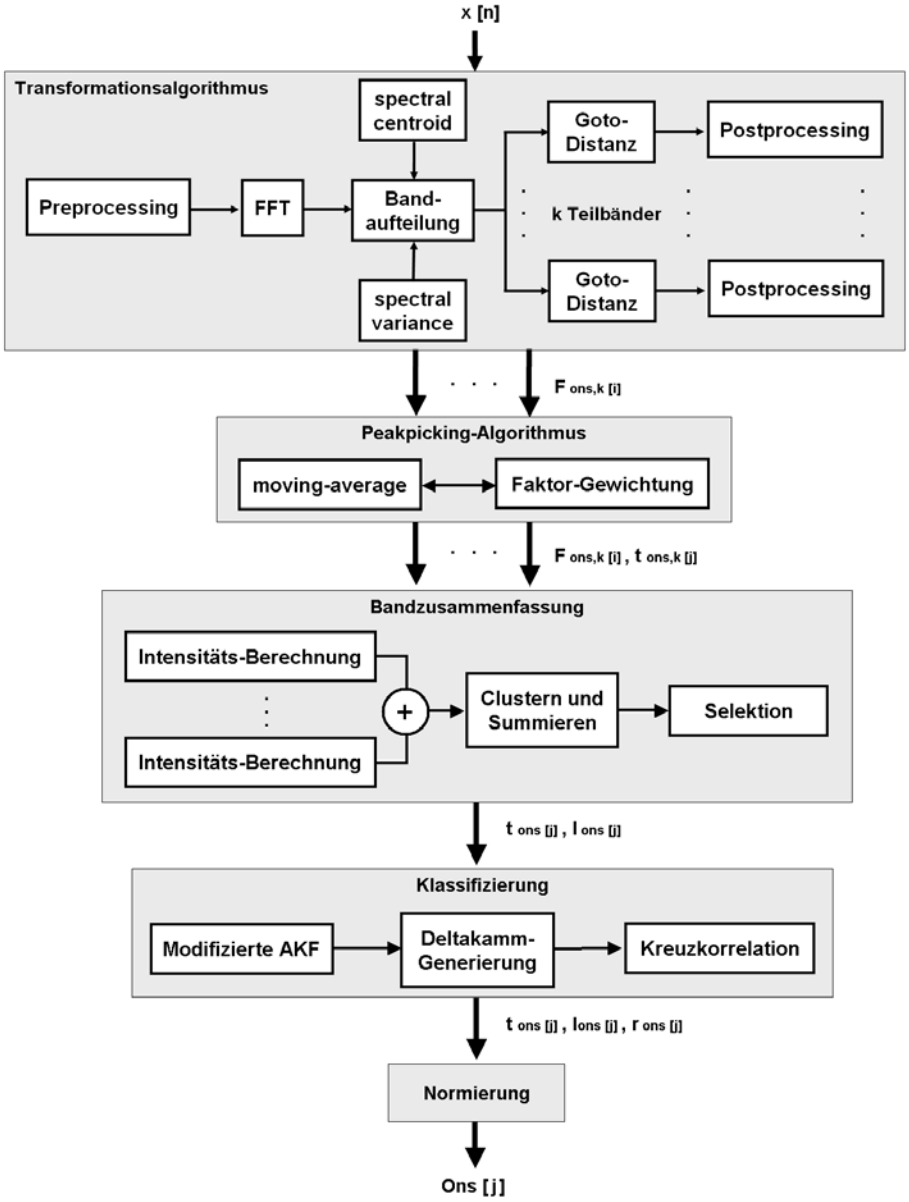

Abbildung 4.18: Blockbild des Algorithmus für die automatische Onset-Erkennung zur Steuerung von Beattracking-Systemen

4.4.1 Datenformat

Das Gesamtsystem soll folgende für das Beattracking relevante Informationen liefern:

1. Zeitpunkt des Onsets in Samples für die höchstmögliche zeitliche Auflösung. Der automatischen Onset-Erkennung wird zur Initialisierung die Abtastrate übergeben.

2. Onset-Intensität. Dieser Wert liegt im Intervall von [0,1]. Dazu werden die Intensitäten des Onset-Vektors auf das absolute Maximum normiert. Ein Beattracking-System kann den Intensitätswert auch als Wahrscheinlichkeit interpretieren, ob es sich bei dem angegebenen Zeitpunkt überhaupt um einen Onset handelt. Dadurch können eventuelle Fehlerkennungen ausgeschlossen werden.

3. Wahrscheinlichkeit, dass es sich um eine schwere Zählzeit handelt. Der Wert liegt ebenfalls innerhalb [0,1]. Das Verfahren aus Abschnitt 4.3.2 liefert lediglich einen Hinweis auf die Relevanz des Onsets für den Takt. Es handelt sich somit nur um eine grobe Abschätzung der „Eins“. Ein Beattracking-System sollte dies berücksichtigen.

Kapitel 5

Evaluierung

Dieses Kapitel beschreibt die Evaluierung des entwickelten Systems in zwei Teilen. Der erste befasst sich mit der Evaluierungsmethodik. Hierbei wird die allgemeine Vorgehensweise für diese Evaluierung erörtert, um im Anschluss die Entwicklung der Evaluierungsmethode zu beschreiben. Im zweiten Teil wird diese Methode angewendet und das System mit Hilfe einer selbst erstellten Testdatenbank getestet. Die Resultate werden mit denen anderer vorgestellter Verfahren verglichen. Zielsetzung ist zu überprüfen, ob das System die geforderte Maßgabe erfüllt, robust gegenüber dem Musiksignal die für ein Beattracking-System wichtigen Onset-Informationen zu liefern.

5.1 Methodik zur Evaluierung

Die Evaluierung bildet meist den Abschluss der Entwicklung - oder einer Entwicklungsstufe - eines Systems und trifft eine wichtige Aussage über dessen Qualität. Häufig werden aber lediglich die Ergebnisse bestimmter Güteklassifikatoren oder Fehlermaße präsentiert und basierend darauf das System beurteilt. Die angewandte Methode wird nicht beschrieben, so dass die Ergebnisse zum Teil nicht sehr aussagekräftig sind [DBDS03] [HM03] [MM02].

Eine wichtige Information ist die Beschreibung der Evaluierungsmethode. Um eine Evaluierungsmethode zu entwickeln, muss das Ziel der Evaluierung festgesetzt werden. Ziel dieser Evaluierung ist, die Robustheit des Systems gegenüber den Musiksignalen zu testen und mit bestehenden Verfahren zu vergleichen. Ausgehend von dieser Zielsetzung werden zunächst allgemeine Schwierigkeiten zur Evaluierung des Systems erörtert, die aus dem vom System zu lösenden Problem her entstehen. So ist es für den Test einer automatischen Onset-Erkennung für das Beattracking beispielsweise problematisch, geeignete Referenz-Onsets zu generieren, da der genaue Zeitpunkt des Onsets in diesem Kontext nicht klar definiert ist (vergleiche Abschnitt 2.1.1).

Ist das Ziel bestimmt und sind die allgemeinen Probleme eingeschätzt, wird

anschließend auf dieser Basis eine sinnvolle Anzahl von Evaluierungskriterien definiert, die das System aus mehreren Blickwinkeln testen. Dabei ist sowohl die einzelne Aussagekraft des Kriteriums zu bewerten als auch, ob die Zusammenfassung der Ergebnisse aus den einzelnen Kriterien die Gesamtfähigkeit des Systems widerspiegelt. Es wird untersucht, welche Kriterien eine gewichtigere Rolle spielen.

Der nächste Schritt ist die Auswahl des Testmaterials. Tests haben stets einen Stichproben-Charakter, einen allumfassenden Test - das heißt in diesem Fall einen Test, der alle möglichen Musiksignale berücksichtigt - gibt es nicht. So stellt die Auswahl keineswegs ein triviales Problem dar, denn es ist aus der riesigen Menge an Musikdaten eine beschränkte Teilmenge an Testsignalen zu bestimmen, unter deren Verwendung das angestrebte Testziel erreicht werden kann. Bei der Testsignalmenge zur Evaluierung des in dieser Arbeit entwickelten Systems muss beispielsweise beachtet werden, dass möglichst die gängigen Musikgenres vertreten sind und sowohl einstimmige als auch mehrstimmige und komplexe Musiksignale vorkommen, da das Testziel den Nachweis der Robustheit des Systems gegenüber den Musiksignalen fordert.

Ist die Evaluierungsmethode samt Testmaterial entwickelt und beschrieben, können die Tests durchgeführt und die Ergebnisse präsentiert werden. Da die Aussagekraft einiger Evaluierungskriterien häufig erst während oder nach Testdurchführung und Auswertung besser einzuschätzen ist, ist ein abschließendes Fazit über die angewandte Evaluierungsmethode sinnvoll.

5.2 Entwicklung der Evaluierungsmethode

Im Folgenden wird in Anlehnung an die Evaluierungsmethodik die Evaluierungsmethode für das in dieser Arbeit entwickelte System im Detail beschrieben. Die Methode basiert im Allgemeinen darauf, die automatisch erkannten Onsets mit festgelegten Onset-Zeiten einer Referenz zu vergleichen und bestimmte Evaluierungskriterien auszuwerten. Anhand der Ergebnisse können die Fähigkeiten des Systems mit den in der Literatur existierenden Verfahren verglichen werden.

5.2.1 Ziel dieser Evaluierung

Für die Aufstellung der Evaluierungskriterien ist das Hauptziel der Evaluierung maßgeblich. In dieser Arbeit ist das Hauptziel, den Nachweis über die Robustheit des Systems gegenüber den Musiksignalen am Eingang zu liefern. Die Fähigkeiten des Systems zur Onset-Erkennung werden außerhalb eines Beattracking-Systems bewertet, da die Ergebnisse ansonsten zu stark von Qualität und Arbeitsweise des Beattrackers abhängen. Die Beurteilung basiert somit auf dem Vergleich mit den festgesetzten Referenz-Onsets. Des Weiteren bleiben Faktoren wie Rechenperformanz und Echtzeitfähigkeit hinsichtlich eines Softwaresystems unbeachtet.

5.2.2 Allgemeine Probleme zur Evaluierung von Systemen zur automatischen Onset-Erkennung

Für Verfahren aus dem Bereich des Music Information Retrieval - wie der automatischen Onset-Erkennung - ist es allgemein sehr problematisch, die Fähigkeiten verschiedener Ansätze objektiv zu bewerten oder untereinander zu vergleichen. Der Grund liegt darin, dass es für das MIR nach [Dow03]

1. keine standardisierte Sammlung von Musiksignalen gibt, mit der die Verfahren getestet werden können,

2. keine standardisierte Menge an Evaluierungskriterien vorhanden ist und

3. keine standardisierte Evaluierungsmethode vorgeschrieben ist.

Im Gegensatz dazu existieren derartige Standards für textbasiertes Information Retrieval schon seit über einem Jahrzehnt. Aus diesem Grund hat sich um J. Stephen Downie eine Gruppe von Leuten gebildet, die sich mit der Entwicklung von Standards zur Evaluierung von MIR-Systemen auseinandersetzen (für Näheres siehe `http://music-ir.org`).

Für die automatische Onset-Erkennung ergeben sich hieraus im Konkreten weitere Probleme. So existieren keine vorgefertigten Signale mit Referenz-Onsets. Das erschwert enorm den Soll-Istwert-Vergleich beim Test, da entweder durch die manuelle Festlegung der Referenz-Onsets (siehe Abschnitt 5.2.4) oder durch die fehlende exakte Definition der für das Beattracking relevanten Onset-Zeit (vergleiche Abschnitt 2.1.1) Ungenauigkeiten entstehen, die bei der Auswertung der Ergebnisse zusätzlich berücksichtigt werden müssen.

5.2.3 Evaluierungskriterien

Konkretes Testobjekt stellen die Ausgaben des Systems - im Sinne der Zielsetzung genauer genommen nur die jeweiligen Zeitinformationen des Onset-Vektors dar. Die Intensitäten und die Wahrscheinlichkeiten für die Schwere hinsichtlich des Taktes sind natürlich weiterhin wichtige Informationen, deren Güte jedoch ohne die Auswertung mit Hilfe eines Beattracking-Systems nicht einzuschätzen ist.

Im Folgenden werden die einzelnen Kriterien beschrieben und deren Aussagekraft im Hinblick auf die Gesamtbeurteilung des System kurz eingeschätzt.

5.2.3.1 Quote der richtig erkannten Onsets (QRE)

Ein einfaches Maß zur Bewertung der Fähigkeit des Systems, sämtliche Onsets im Signal zu finden, orientiert sich an [CDK00]. Hier wird ein Beattracking-System getestet, indem die Anzahl der richtig gesetzten Beats mit der absoluten Anzahl der Beats ins Verhältnis gesetzt wird. Betrachtet man anstelle der Beats die in

dem Signal vorkommenden Onsets, berechnet sich die *Quote der richtig erkannten Onsets* (QRE_{Ons}) mit

$$QRE_{Ons} = \frac{Anz_{match}(Ons)}{Anz_{total}(Ons)} \cdot 100. \tag{5.1}$$

$Anz_{match}(Ons)$: Anzahl der richtig erkannten Onsets
$Anz_{total}(Ons)$: absolute Anzahl der Onsets in der Referenz

Ein Onset gilt als richtig erkannt, wenn er sich innerhalb eines Rechteckfensters von 50 ms (vergleiche hierzu Abschnitt 2.1.2) um den jeweilig betrachteten Referenz-Onset befindet. Dieses Maß ist lediglich eine grobe Einschätzung der Fähigkeiten, da Fehlerkennungen hier nicht berücksichtigt werden. Das Resultat liegt im Bereich von 0 bis 100, so dass die QRE auch in Prozent angegeben werden kann. Für dieses Maß und die beiden folgenden Kriterien gilt, dass je größer die Werte des Maßes ausfallen, das System besser in der Lage ist, Onsets korrekt zu erkennen.

5.2.3.2 Correct Detection Rate (CDR)

In [RLM03] wird für die automatische Onset-Erkennung ein weiteres Maß für die Erkennungsfähigkeit definiert. Die *Correct Detection Rate* (CDR_{Ons}) berücksichtigt im Gegensatz zur QRE_{Ons} auch die Fehlerkennungen und berechnet sich folgendermaßen:

$$CDR_{Ons} = \frac{Anz_{total}(Ons) - Anz_{missing}(Ons) - Anz_{spurious}(Ons)}{Anz_{maxtotal}(Ons)} \cdot 100. \tag{5.2}$$

$Anz_{missing}(Ons)$: Anzahl der nicht erkannten Onsets
$Anz_{spurious}(Ons)$: Anzahl der Fehlerkennungen
$Anz_{maxtotal}(Ons)$: Maximum aus Anzahl der Referenz- und erkannten Onsets

Der Nachteil an diesem Verfahren liegt darin, das für CDR_{Ons} im Extremfall negative Werte errechnet werden, wenn die Summe aus $Anz_{missing}(Ons)$ und $Anz_{spurious}(Ons)$ größer als $Anz_{total}(Ons)$ ist. Die CDR kann ebenfalls maximal den Wert 100 annehmen, wobei die Angabe in Prozent wegen möglicher negativer Werte nicht zu empfehlen ist.

5.2.3.3 Onset-Erkennungsmaß (OEM)

Aufgrund des Nachteils der CDR_{Ons} wird für diese Arbeit ein eigenes Erkennungsmaß definiert. Um negative Werte zu vermeiden wird das Erkennungsmaß OEM_{Ons} berechnet mit

$$OEM_{Ons} = \frac{Anz_{match}(Ons)}{Anz_{match}(Ons) + Anz_{missing}(Ons) + Anz_{spurious}(Ons)} \cdot 100. \tag{5.3}$$

Von diesem Maß wird sich erhofft, die Erkennungsfähigkeit des Systems möglichst kritisch zu beurteilen. Der Wertebereich liegt zwischen 0 und 100.

5.2.3.4 Statistische Streuungsmaße

Für die Abschätzung der zeitlichen Genauigkeit des Verfahrens werden statistische Streuungsmaße verwendet. Berechnet werden Mittelwert, Standardabweichung sowie Kurtosis, die jeweils verschiedene Aussagen über die zeitliche Abweichung der erkannten Onset-Zeit zu dem Referenz-Zeitpunkt treffen.

Mittelwert. Dem Mittelwert (siehe Gleichung 2.14 in Abschnitt 2.2.4.2) kommt für die Bewertung der zeitlichen Genauigkeit die größte Gewichtung zu. Er trifft sowohl die Aussage darüber, wie stark die erkannten Onsets im Mittel von der Referenz abweichen als auch, ob die Onsets eher zu früh oder zu spät erkannt werden. Der Mittelwert kann als erste Näherung einer konstanten Abweichung zur Kompensation der zeitlichen Ungenauigkeit herangezogen werden.

Standardabweichung. Sie berechnet sich aus der Quadratwurzel der Varianz (vergleiche Abschnitt 2.2.4.3):

$$\sigma = \sqrt{(E[(\mathcal{X} - \mu)^2])}. \tag{5.4}$$

Sie gibt eine Einschätzung darüber, ob die einzelnen Abweichungen vom Referenz-Onset in unmittelbarer Nähe des Mittelwertes liegen. Bei besonders niedriger Standardabweichung kann somit der zeitliche Fehler durch den Mittelwert approximiert und kompensiert werden.

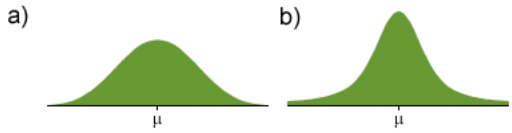

Abbildung 5.1: Verteilungsdichtefunktion: a) flacher Verlauf, b) steiler Verlauf

Kurtosis. Die Kurtosis stellt das Moment vierter Ordnung dar. Sie gibt Auskunft über die Steilheit der Verteilungsdichtefunktion. Die Kurtosis berechnet sich mit

$$kurt(\mathcal{X}) = \frac{E[(\mathcal{X} - \mu)^4]}{\sigma^4}. \tag{5.5}$$

Die Kurtosis hat für die Abschätzung der zeitlichen Genauigkeit eine ähnliche Aussagekraft wie die Standardabweichung. Bei einem sehr hohem Wert liegen die meisten Abweichungen im Bereich des Mittelwertes μ, so dass μ eine gute Approximation der konstanten Abweichung vom Referenz-Onset darstellt. Eine niedrige Standardabweichung bedingt allerdings nicht einen großen Wert der Kurtosis und umgekehrt, so dass es sinnvoll ist, beide Kriterien zu betrachten.

5.2.4 Auswahl und Erstellung des Testmaterials

Wie schon in Abschnitt 5.2.2 erwähnt, ist die Auswahl und Erstellung des Testmaterials ein grundlegendes Problem für die Evaluierung von MIR-Systemen. Allein die Festlegung der zu verwendenden Musikgenres erfordert eine gründliche Diskussion. Im Hinblick auf den Rahmen dieser Arbeit wird deshalb von einer vereinfachten Taxonomie zur Klassifikation der Musiksignale nach Genres ausgegangen. Tabelle 5.1 zeigt die berücksichtigten Genres, zum Teil in Unterkategorien in Anlehnung an [BL04] aufgeteilt.

Genre	**Unterklassen**
Klassik-Orchestermusik	Orchester, Orchester mit Solist, Choralmusik
Klassik-Kammermusik	Kammermusik mit Klavier, Streichquartett, Solo
Klassik-Oper	Oper
Rock	Hard Rock, Rock 'n' Roll
Electronic/Pop	Pop, Techno, Hip-Hop
Jazz/Blues	Jazz, Blues

Tabelle 5.1: Vereinfachte Taxonomie der zur Evaluierung verwendenden Musikgenres

Das größte Problem stellt die Erstellung des Referenzmaterials dar. Um die Robustheit nachzuweisen, muss das System anhand einer größeren Menge von Musiksignalen verschiedenster Komplexität evaluiert werden. Da bis dato keine vorgefertigten standardisierten Testdatenbänke existieren, müssen sämtliche Referenz-Onsets von Hand bzw. visuell und vom Gehör aus gesetzt werden, was nicht nur einen erheblichen Arbeitsaufwand, sondern auch eine hohe Ungenauigkeit mit sich bringt.

Zur besseren Einschätzung der zeitlichen Genauigkeit wird das System deshalb weiterhin mit einem MIDI-Signal getestet. Hierzu werden aus dem MIDI-Signal zwei unterschiedliche Audiosignale generiert, deren Instrumente stark unterschiedliche Einschwingvorgänge aufweisen. Das erste Signal $MIDI_{Klavier}$ verwendet ein Klavier mit deutlichen Anschlägen, demnach weisen die Klänge kurze und klare Einschwingvorgänge auf. Das zweite Signal $MIDI_{Geige}$ verwendet eine Geige mit lang einschwingenden Klängen, die weiterhin mit Vibrato gespielt sind.

Vor den Tests muss der konstante zeitliche Versatz zwischen den Onsets im generierten Audiosignal und den dazugehörigen *MIDI Note On*-Befehlen hinsichtlich des wahrgenommenen Onsets abgeschätzt werden. Hierbei wird von einem konstanten Versatz ausgegangen, da die MIDI-Noten gleiche Lautstärken aufweisen und in einem begrenzten Tonbereich liegen. Unter Berücksichtigung dieser konstanten Abweichung kann die zeitliche Genauigkeit wesentlich exakter bewertet werden als mit Hilfe der manuell gesetzten Referenz-Onsets. Für die Einschätzung des Versatzes wird das in [Jan95] vorgeschlagene Verfahren verwendet,

indem der wahrgenommene Onset-Zeitpunkt mit Hilfe der zweiten Ableitung der Hüllkurve des Signals festgelegt wird (vergleiche Abbildung 5.2).

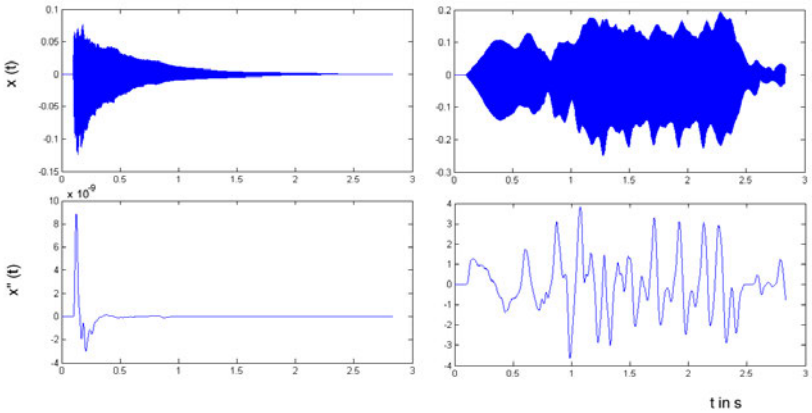

Abbildung 5.2: Signale des Klavierklanges (links) und des Geigenklanges (rechts) zur Abschätzung des wahrgenommenen Onsets nach [Jan95]: Zeitsignal (oben) und zweite Ableitung der Hüllkurve (unten)

Für den Klavierklang ergibt sich zwischen wahrgenommenen und MIDI-Onset ein Versatz von 5.1 ms, für die Geige ein Versatz von 52.8 ms. Bei den Tests wird dieser jeweils von der Differenz zwischen erkanntem Onset und Referenz abgezogen. Bei dem Geigenklang fällt weiterhin auf, dass die zweite Ableitung durch das Vibrato viele aufeinander folgende Maxima aufweist. Bei der Berechnung des Versatzes wird in diesem Fall nicht das absolute, sondern das im Zeitraum der Einschwingphase befindliche lokale Maximum berücksichtigt.

5.2.5 Durchführung

Zur Durchführung werden für die in Tabelle 5.1 bestimmten Unterklassen der Musikgenres jeweils ein bis zwei konkrete Testsignale der Länge einer halben Minute ausgewählt und die Referenz-Onsets manuell bestimmt. Im Vergleich stehen das in dieser Arbeit entwickelte System S_{KLICH} sowie die frequenzbasierten Verfahren mit den Distanzen D_{MKL}, D_{KMPX} und der einfachen D_{GOTO} (vergleiche Abschnitt 3.4.1). D_{EUC} wird nicht betrachtet, da diese Distanz bei den vorhergehenden Tests in Abschnitt 4.1.2 zu schlecht abgeschnitten hat. Für die bekannten Verfahren wird ein normales moving-average-Peakpicking verwendet, wobei der Gewichtungsfaktor basierend auf den Untersuchungen aus Abschnitt 4.1.2 an die Verfahren angepasst wurde, um möglichst optimale Ergebnisse zu erhalten.

Die Kriterien 5.2.3.1 bis 5.2.3.3 werden jeweils für die Gesamtmenge der Testsignale Q_{TEST} und zusätzlich separat für die Musikgenres Klassik-Kammermusik (Q_{K-K}) und Electronic/Pop ($Q_{E/P}$) berechnet. Die statistischen Kriterien werden zum einen für Q_{TEST} und zum anderen für jedes aus dem MIDI-Signal erzeugte Audiosignal bestimmt. Die Ergebnisse werden in Abschnitt 5.3 aufgelistet.

5.3 Ergebnisse der Evaluierung

In diesem Abschnitt werden die Ergebnisse meist in Form von Tabellen dargeboten. Zunächst folgt die Übersicht der absoluten Anzahl von Referenz-Onsets in jeder Testsignalmenge (Tabelle 5.2) und anschließend die Statistik über die Anzahl richtig und falsch erkannter Onsets. Wie bereits erwähnt, gilt ein Onset als richtig erkannt, wenn er sich innerhalb eines Rechteckfensters von 50 ms um den jeweilig betrachteten Referenz-Onset befindet.

Testmaterial	Anzahl der Signale	Anzahl der Referenz-Onsets
Q_{TEST}	18	787
Q_{K-K}	4	182
$Q_{E/P}$	4	164
$MIDI_{Klavier}$	1	53
$MIDI_{Geige}$	1	53

Tabelle 5.2: Anzahl der Signale und Referenz-Onsets in den Testsignalen

5.3.1 Statistik erkannter Onsets

$Anz_{match}(Ons)$	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	669 (85.0%)	640 (81.3%)	422 (53.6%)	563 (71.5%)
Q_{K-K}	147 (80.7%)	145 (79.6%)	131 (71.9%)	142 (78.0%)
$Q_{E/P}$	153 (93.2%)	149 (90.8%)	97 (59.1%)	131 (79.8%)
$MIDI_{Klavier}$	53 (100%)	53 (100%)	43 (81.1%)	53 (100%)
$MIDI_{Geige}$	32 (60.3%)	34 (64.1%)	22 (41.5%)	37 (69.8%)

Tabelle 5.3: Anzahl der richtig erkannten Onsets: Total und in Prozent

$Anz_{spurious}(Ons)$	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	128 (16.2%)	125 (15.8%)	122 (15.5%)	105 (13.3%)
Q_{K-K}	35 (19.2%)	34 (18.7%)	27 (14.8%)	25 (13.7%)
$Q_{E/P}$	21 (12.8%)	23 (14.0%)	26 (15.8%)	18 (10.9%)
$MIDI_{Klavier}$	0 (0.0%)	0 (0.0%)	3 (5.6%)	1 (1.9%)
$MIDI_{Geige}$	44 (83.0%)	43 (81.1%)	37 (69.8%)	42 (79.2%)

Tabelle 5.4: Anzahl der falsch erkannten Onsets: Total und in Prozent

Zu bemerken ist, dass alle Verfahren bei dem per MIDI-Signal generierten Geigenstück eine extrem hohe Anzahl an Fehlerkennungen aufweisen. Grund hierfür ist das Vibrato in den Geigenklängen. Verfahren zur automatischen Onset-Erkennung sind grundsätzlich anfällig gegenüber derartigen Frequenz- und Amplitudenmodulationen.

5.3.2 Ergebnisse der Erkennungsmaße

Die Ergebnisse der folgenden Kriterien (vergleiche Abschnitt 5.2.3) untersuchen die Korrektheit der automatischen Onset-Erkennung. Zur Vereinheitlichung bleiben die Werte dimensionslos, da sich für die CDR mitunter negative Ergebnisse ergeben können (vergleiche Abschnitt 5.2.3.3). Allgemein gilt, das höhere Werte eine bessere Fähigkeit zur korrekten Erkennung von Onsets anzeigen.

QRE	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	85.0	81.3	53.6	71.5
Q_{K-K}	80.7	79.6	71.9	78.0
$Q_{E/P}$	89.5	87.3	53.2	71.9

Tabelle 5.5: Ergebnisse der QRE

CDR	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	67.8	65.4	38.1	58.2
Q_{K-K}	61.5	60.9	57.1	61.9
$Q_{E/P}$	75.8	73.6	43.3	68.9

Tabelle 5.6: Ergebnisse der CDR

OEM	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	73.1	70.1	46.4	63.1
Q_{K-K}	67.7	67.1	61.0	68.5
$Q_{E/P}$	82.7	79.6	52.7	67.3

Tabelle 5.7: Ergebnisse des OEM

Die allgemeine Tendenz der obigen Ergebnisse zeigt klar auf, dass die Optimierungen hinsichtlich der Robustheit des Systems gegenüber dem Musiksignal erfolgreich sind. Es gibt lediglich zwei Beispiele, bei denen ein anderes Verfahren besser abschneidet. In beiden Fällen ist es D_{KMPX}, die durch ihre geringere Anzahl an Fehlerkennungen bei den Kammermusik-Testsignalen (Q_{K-K}) günstigere Ergebnisse liefert. Extrem schlechte Werte weist D_{MKL} auf, was auch den Erwartungen entspricht, da dieses Distanzmaß bereits bei den Tests in Abschnitt 4.1.2 negativ abschnitt. Der Erfolg zeigt sich weiterhin darin, dass das entwickelte System stets höhere Werte aufweist als D_{GOTO}. Da dieses Verfahren als Basisalgorithmus des Systems eingesetzt wird, kann die Steigerung der Robustheit eindeutig den durchgeführten Optimierungen zugeschrieben werden.

5.3.3 Ergebnisse der statistischen Abweichungsmaße

Die statistischen Abweichungsmaße betrachten die zeitliche Genauigkeit der erkannten Onsets. Berücksichtigt werden jedoch nur korrekte Erkennungen. Aus diesem Grund sind die Ergebnisse von $MIDI_{Geige}$ nicht sehr aussagekräftig, da aufgrund der hohen Fehlerrate (vergleiche Tabelle 5.4) nur sehr wenige Onsets statistisch ausgewertet werden. Die Onsets aus $MIDI_{Klavier}$ hingegen werden nahezu vollständig richtig erkannt. Für jedes Verfahren wird deshalb die Häufigkeitsverteilung der zeitlichen Abweichungen von den Referenz-Onsets anhand dieses Signals betrachtet. Dies dient der Veranschaulichung der anschließenden tabellarischen Darstellung der Ergebnisse der statistischen Abweichungsmaße. Abbildung 5.3 zeigt für jedes Verfahren die Verteilungsfunktion. Hierbei wird der in Abschnitt 5.2.4 berechnete konstante Versatz zwischen MIDI- und wahrgenommenen Onset berücksichtigt.

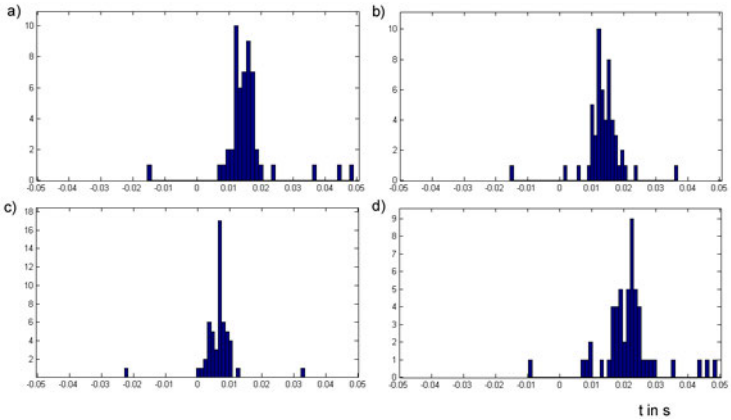

Abbildung 5.3: Verteilungsfunktionen der Abweichungen gegenüber dem Referenz-Onset bei $MIDI_{Klavier}$: a) S_{KLICH}, b) D_{GOTO}, c) D_{MKL}, d) D_{KMPX}

Mittelwert	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	0.0201	0.0180	0.0179	0.0311
$MIDI_{Klavier}$	0.0155	0.0140	0.0065	0.0215
$MIDI_{Geige}$	0.0264	0.0253	0.0171	0.0371

Tabelle 5.8: Ergebnisse der Mittelwertberechnung

Standardabweichung	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	0.0142	0.0160	0.0114	0.0122
$MIDI_{Klavier}$	0.0086	0.0064	0.0060	0.0089
$MIDI_{Geige}$	0.0117	0.0094	0.0108	0.0124

Tabelle 5.9: Ergebnisse der Standardabweichung

Kurtosis	S_{KLICH}	D_{GOTO}	D_{MKL}	D_{KMPX}
Q_{TEST}	6.18	6.52	7.03	3.19
$MIDI_{Klavier}$	10.98	12.58	18.90	6.88
$MIDI_{Geige}$	7.12	9.83	8.04	3.19

Tabelle 5.10: Ergebnisse der Kurtosis

An den Mittelwerten lässt sich erkennen, dass das entwickelte System S_{KLICH} hinsichtlich der zeitlichen Genauigkeit der Erkennung in etwa mit D_{GOTO} vergleichbar ist und bessere Ergebnisse liefert als D_{KMPX}. D_{MKL} weist im Mittel jedoch die deutlich kleinste Abweichung auf. Für alle Verfahren wird festgestellt, dass die Onsets gegenüber der Referenz verzögert erkannt werden. Dies ist sicherlich dadurch bedingt, da es sich um FFT-basierte Verfahren handelt und die Onsetkurven im Postprocessing (vergleiche Abschnitt 3.4) durch Tiefpassfilterung geglättet werden. Die Werte von Standardabweichung und Kurtosis treffen hier eine interessante Aussage. So weisen die Verfahren mit Ausnahme von D_{KMPX} im Mittel sehr konstante Abweichungen vom Referenzwert auf, was sich in geringen Standardabweichungen, hoher Kurtosis und steilen Verteilungsfunktionen widerspiegelt. An führender Position steht wiederum D_{MKL}, so dass dieses Verfahren hinsichtlich der zeitlichen Genauigkeit insgesamt am besten abschneidet.

Die Ergebnisse der Standardabweichung korrespondieren in etwa mit denen in [Sch95]. Hier werden in Abhängigkeit von der Komplexität des Musiksignals Standardabweichungen im Bereich von 10-30 ms gemessen. Die Verfahren in dieser Arbeit weisen sogar mitunter geringere Abweichungen auf.

Diese Ergebnisse sind dennoch mit Vorsicht zu genießen, da hier besonderes Augenmerk auf die gute Onset-Erkennung des Klaviersignals gesetzt wurde. Weiterhin muss beachtet werden, dass den Ergebnissen der Testsignalmenge Q_{TEST} stets die Ungenauigkeit des manuellen Setzens der Referenz-Onsets zugrunde liegt.

Die Ergebnisse zeigen, dass das entwickelte System im Mittel zwar eine höhere Abweichung als D_{MKL} aufweist, der Wert jedoch im Großen und Ganzen konstant bleibt. Beattracking-Systeme reagieren hinsichtlich der Tempoerkennung in der Regel robust gegenüber konstanten Abweichungen. Das System ist somit durchaus für Beattracking-Systeme einsatzfähig. Durch die nachgewiesene Robustheit gegenüber den Musiksignalen am Eingang bietet das in dieser Arbeit entwickelte System zur automatischen Onset-Erkennung eine solide Grundlage für robustes Beattracking.

5.4 Fazit der Evaluierungsmethode

Zum Abschluss der Evaluierung soll ein kurzes Fazit über die angewandte Evaluierungsmethode gezogen werden, da einige Probleme oder Nachteile erst mit- oder nach der Evaluierung selbst klar werden.

Ein wichtiges Hauptproblem dieser Evaluierung ist die fehlende Referenzstruktur. Ohne standardisiertes Testmaterial ist die Aussage der Testergebnisse immer leicht subjektiv belastet. In dieser Arbeit wurde zwar versucht, eine möglichst genreübergreifende Auswahl an Testsignalen zu erstellen, doch auch diese musste in Hinblick auf den zeitlichen Rahmen dieser Arbeit reduziert werden.

Die Auswertung der Testkriterien ergibt eine grundlegende Aussage über die Fähigkeiten des Systems. Das vorgestellte Erkennungsmaß OEM stellt durch die Berücksichtigung sowohl fehlender als auch falsch erkannter Onsets eine Verbesserung bestehender Kriterien dar, da unklare Aussagen aufgrund negativer Werte nicht vorkommen können. Doch gerade hinsichtlich der Einsetzbarkeit in Beattracking-Systemen wäre eine genauere Abschätzung des wahrgenommenen Onset-Zeitpunktes (vergleiche Abschnitt 2.1.1) wünschenswert, um auf dieser Basis Kriterien zu erstellen, die eine derartige Einsetzbarkeit überprüfen.

Kapitel 6

Resümee

6.1 Zusammenfassung

Diese Arbeit hat sich ausführlich mit der Untersuchung und Weiterentwicklung von Systemen zur automatischen Onset-Erkennung in Musiksignalen für Beattracking-Systeme auseinandergesetzt. Zunächst ist der Onset-Begriff erörtert worden. Interessant hierbei ist die Diskrepanz zwischen physikalischem und wahrgenommenem Onset und die Ansätze, die Wahrnehmung von Onsets psychoakustisch zu modellieren.

Die Menge der in der Literatur existierenden Verfahren zur automatischen Onset-Erkennung hat gezeigt, dass auf diesem Gebiet ein enormer Forschungsdrang herrscht. Repräsentativ wurden die naheliegendsten in dieser Arbeit ausgewertet. Werden die verschiedenen betrachteten Ansätze zeitlich eingeordnet, so stellt man einen Übergang von der Betrachtung der Eingangssignale im Zeitbereich in den Frequenzbereich fest, trotz des Problems der zeitlichen Auflösung. Frequenzbasierte Verfahren haben nämlich den Vorteil, nicht von transienten Vorgängen abzuhängen, die sich deutlich im Hüllkurvensignal des Musikstücks abzeichnen. Dies ist eine wichtige Voraussetzung, um die Zielsetzung dieser Arbeit zu erfüllen, einen auf die Robustheit gegenüber den Musiksignalen am Eingang hin optimierten Algorithmus zur automatischen Onset-Erkennung für Beattracking-Systeme zu entwickeln. Ausgangslage für diese Entwicklung war ein bereits existierendes Verfahren, welches aufgrund konkreter Tests an Musiksignalen unterschiedlicher Komplexität als Basisalgorithmus ausgewählt wurde.

Eine entscheidende Neuerung auf dem Gebiet der automatischen Onset-Erkennung stellt die automatische Bandaufteilung aufgrund spektraler Features des Eingangssignals dar. Auf diesem Wege ist es dem System möglich, sich ohne weitere Eingabeinformationen an die spektralen Eigenschaften des Eingangssignals anzupassen. Durch die psychoakustisch orientierte Bandzusammenfassung fließt weiterhin der Aspekt der menschlichen Wahrnehmung mit ein, sehr kurz aufeinander folgende Klangereignisse als Einheit anzusehen. Für die Rhythmuswahrneh-

mung kann dies eine entscheidende Rolle spielen. Die Einschätzung der Relevanz eines Onsets hinsichtlich des Taktes kann weiterhin für einen Beattracker von Nutzen sein, die Taktart zu bestimmen. Der vorgestellte Ansatz liefert jedoch lediglich eine grobe Einschätzung der Lage schwerer Zählzeiten.

Besonders die automatische Anpassung an das Eingangssignal steigert die Robustheit des Systems, was sich in den Ergebnissen der Evaluierung widerspiegelt. Für die Evaluierung selbst war jedoch zunächst wichtig, die angewandte Evaluierungsmethode zu erörtern, da dies in der Literatur meist vernachlässigt wird und somit die dargebotenen Ergebnisse an Aussagekraft verlieren.

Die Ergebnisse fallen im Mittel positiv aus, so dass im Vergleich zu den herkömmlichen Verfahren gerade im Hinblick auf die Robustheit eine Steigerung erzielt wurde. Was die zeitliche Genauigkeit angeht, so ist das System noch verbesserungsfähig, bleibt aber innerhalb einer akzeptablen Toleranz, so dass das entwickelte System in Beattracking-Systemen eingesetzt werden kann.

6.2 Ausblick

Gerade zum Schluss der Arbeit entstanden neue Ideen zur weiteren Optimierung des Systems. Zur Verbesserung der zeitlichen Auflösung könnte beispielsweise im Anschluss an den Algorithmus des bisherigen Systems eine zweite Stufe folgen, in der die zeitliche Genauigkeit aufgrund von zeitbasierten Verfahren verbessert wird. Hierbei können die in Abschnitt 2.1.1 erwähnten Verfahren verwendet oder basierend darauf eigene Ansätze entwickelt werden. Es liegt nahe, den in Abschnitt 3.5 beschriebenen Ansatz mit Linearer Prädiktion nochmals in diesem Kontext zu untersuchen.

Bei allen frequenzbasierten Verfahren traten bei bestimmten Testsignalen, in denen Amplituden und Frequenzmodulationen vorkommen - beispielsweise verursacht durch ein Vibrato - häufig Fehlerkennungen auf. Das System könnte in der ersten Stufe durch die Integration eines Moduls zur Kompensation derartiger Modulationen erweitert werden, um nochmals die Robustheit zu steigern. Erste Ansätze hierzu finden sich in [RDS+99].

Insgesamt ist abzusehen, dass im Bereich des MIR noch viel zu forschen ist - besonders hinsichtlich der Modellierung menschlicher kognitiver Fähigkeiten. Interessant wäre für diese Arbeit eine psychoakustische Untersuchung der Vorgänge im Gehör während der Onset-Wahrnehmung, um für die zeitliche Justierung der Onset-Zeit einen eigenes Modell aufzustellen. Dies könnte den Ansatzpunkt zur Fortsetzung dieser Arbeit darstellen.

Literaturverzeichnis

[AF93] A. Friberg, J. S.: Perception of Just-Noticeable Time Displacement of a Tone Presented in a Metrical Sequence at Different Tempos. In: *ASA 126th Meeting*. Denver, 1993

[AP03] Abdallah, Samer A. ; Plumbley, Mark D.: Unsupervised Onset Detection: A Probabilistic Approach Using ICA and a Hidden Markov Classifier. In: *Proceedings of Cambridge Music Processing Colloquium*. Cambridge, 2003

[AT99] Alghoniemy, Masoud ; Tewfik, Ahmed H.: Rhythm and Periodicity Detection in Polyphonic Music. In: *Proceedings IEEE third Workshop on Multimedia Signal Processing*. Denmark, 1999

[BL04] Burred, Juan J. ; Lerch, Alexander: Hierarchical Automatic Audio Signal Classification. In: *Journal of the Audio Engineering Society* 52 (2004), Nr. 7/8, S. 724–739

[BPM87] B. Pompino-Marschall, B. K.: P-Centers and the Perception of Momentary Tempo. In: *Proceedings of the 11th Congress of Phonetic Science* Bd. 5, 1987

[Bre90] Bregman, A. S.: *Auditory Scene Analysis: The Perceptual Organization of Sound*. A Bradford book, 1990

[BS03] Bello, J. P. ; Sandler, M.: Phase-based Note Onset Detection for Music Signals. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003

[CDK00] Cemgil, A. T. ; Desain, P. ; Kappen, H. J.: Rhythm Quantization for Transcription. In: *Computer Music Journal* 24:2 (2000), S. 60–76

[DBDS03] Duxbury, C. ; Bello, J.P. ; Davies, M. ; Sandler, M.: 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)*. London, 2003

[DDS02] Duxbury, C. ; Davies, M. ; Sandler, M.: A Hybrid Approach to Musical Note Onset Detection. In: *Proceedings of the 5th Int. Conference on Digital Audio Effects (DAFX-02)*. Hamburg, September 2002

[DDS03] Duxbury, C. ; Davies, M. ; Sandler, M.: Complex Domain Onset Detection for Musical Signals. In: *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*. London, October 2003

[Dix99] Dixon, Simon: A Beat Tracking System for Audio Signals. In: *Proceedings of the Conference on Mathematical and Computational Methods in Music*. Wien, 1999

[Dix03] Dixon, Simon: On the Analysis of Musical Expression in Audio Signals. 5021 (2003)

[Dow03] Downie, J.: Toward the Scientific Evaluation of Music Information Retrieval Systems. Baltimore, 2003

[GM95] Goto, M. ; Muraoka, Y.: Music Understanding at the Beat Level - Real-Time Beat Tracking for Audio Signals. In: *Proceedings of IJCAI Workshop on Computational Auditory Scene Analysis*, 1995

[GM99] Goto, Masataka ; Muraoka, Yoichi: Real-Time Beat Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions. In: *Speech Communication* 27 (1999)

[Got01] Goto, M.: An Audio-based Real-Time Beat Tracking System for Music With or Without Drum-sounds. In: *Journal of New Music Research* 30 (2001), Nr. 2

[Han89] Handel, S.: *Hearing*. Cambridge, MA : A Bradford book, 1989

[Hay96] Haykin, Simon: *Adaptive Filter Theory*. 3. Prentice-Hall Inc., 1996

[HM03] Hainsworth, S. ; Macleod, M.: Onset Detection in Musical Audio Signals. In: *Proceedings of International Conference on Computer Music (ICMC)*. Singapore, September/October 2003

[Jan95] Jancker, P. M.: Sprechrhythmus, Silbe, Ereignis / Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation (FIPKM) 33. 1995. – Forschungsbericht

[JV82] J. Vos, R. R.: The Perceptual Onset of Musical Tones. In: *Music, Mind and Brain. A Neuropsychology of Music*. New York, 1982

[Kla99] KLAPURI, A.: Sound Onset Detection by Applying Psychoacoustic Knowledge. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Proceesing (ICASSP,'99)*, 1999

[Kla01] KLAPURI, A.: Multipitch Estimation and Sound Separation by the Spectral Smoothness Principle. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP,'01)*. Salt Lake City, 2001

[Kla03] KLAPURI, Anssi P.: Musical Meter Estimation and Music Transcription. In: *Cambridge Music Processing Colloquium*. Cambridge University, 2003

[Mar81] MARCUS, S. M.: Acoustic Determinants of Perceptual Center (P-Center) Location. In: *Perception and Psychophysics* Bd. 30, 1981

[ME49] MEYER-EPPLER, W.: *Elektrische Klangerzeugung. Elektronische Musik und synthetische Sprache*. Bonn : Dümmler, 1949

[MM02] MATIJA MAROLT, Marko P.: Neural Networks for Note Onset Detection in Piano Music. In: *Proceedings of ICMC 2002*. Cambridge, 2002

[PK02] PAULUS, Jouni ; KLAPURI, Anssi: Measuring the Similarity of Rhythmic Patterns. In: *Proceedings of the 3rd International Conference on Music Information Retrieval*. Paris, 2002

[RDS+99] ROSSIGNOL, S. ; DEPALLE, P. ; SOUMAGNE, J. ; RODET, X. ; COLLETTE, J.-L.: Vibrato: Detection, Estimation, Extraction, Modification. In: *Proceedings of International Conference on Digital Audio Effects (DAFx-99)*, 1999

[Reu95] REUTER, C.: *Der Einschwingvorgang nichtperkussiver Musikinstrumente*. Frankfurt am Main : Lang, 1995

[Rie03] RIEDEL, Thomas: *Klassifikation von Klangdaten*, Technische Universität Chemnitz, Diplomarbeit, 2003

[RLM03] R. LIU, J. W. ; MURPHY, P.: Time Domain Note Average Energy Based Music Onset Detection. In: *Proceedings of Stockholm Music Acoustics Conference 2003*. Stockholm, 2003

[Sch78] SCHÜTTE, H.: Ein Funktionsschema für die Wahrnehmung eines gleichmäßigen Rhythmus in Schallimpulsfolgen. In: *Biological Cybernetics* Bd. 29, 1978

[Sch95] SCHEIRER, Eric D.: *Extracting Expressive Performance Information from Recorded Music*, Massachusetts Institute of Technology, Masters Thesis, September 1995

[Sch97] SCHEIRER, E.: Tempo and Beat Analysis of Acoustic Signals. In: *J. Acoust. Soc. Am.* 103 (1997)

[Sea38] SEASHORE, C. E.: *Psychology of music.* New York : McGraw-Hill, 1938

[Smi96a] SMITH, Leigh M.: Modelling Rhythm Perception by Continuous Time-Frequency Analysis. In: *Proceedings of the International Computer Music Conference*, 1996

[Smi96b] SMITH, L.S.: Onset-based Sound Segmentation. In: *Advances in Neural Information Processing Systems*. Cambridge, 1996

[VM97] VERMA, Tony S. ; MENG, Teresa H.: An Analysis/Synthesis Tool for Transient Signals. (1997)

[ZF99] ZWICKER, E. ; FASTL, H.: *Psychoacoustics. Facts and Models.* 2. Springer, 1999