Ein physiologisch gehörgerechtes Verfahren zur automatisierten Melodietranskription

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt

der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau

von Dipl.-Phys. Thorsten Heinz
geboren in Heessen

1. Gutachter: Herr Univ.-Prof. Dr.-Ing. K. Brandenburg
2. Gutachter: Herr Priv.-Doz. Dr.-Ing. habil. P. Husar
3. Gutachter: Herr Dr. H. Distler

eingereicht am: 06.04.2004
wissenschaftliche Aussprache: 07.06.2006

urn:nbn:de:gbv:ilm1-2006000055

Abstract

This thesis describes the implementation of a method for automatic transcription of music. To date, the human ability of musical perception, and especially that of musical experts, cannot be reproduced by technical systems. It can therefore be considered a plausible approach to make use of perceptually motivated strategies, as far as possible, to bridge this gap for systems aimed at analyzing and understanding music. In the presented work, the principal processing mechanisms used by the human auditory periphery, as well as high-level cognitive brain functions are applied to the analysis of musical input.

A detailed summary describes state-of-the-art algorithms for detection of fundamental frequencies and segmentation of musical phrases. Current systems for monophonic and polyphonic melody transcription are introduced.

The fundamental physiological components of the auditory periphery and principles based on Gestalt psychology are illustrated. Furthermore the models used in this thesis, including active sound preprocessing of the inner ear, are described in detail. In order to take account of auditive post-processing, the principles of pitch perception and a hierarchical model based on assumptions from Gestalt psychology are utilized.

Besides the development of the hierarchical model, the core of this thesis consists of the implementation of the chosen methods for monophonic and polyphonic transcription strategies. Aurally correct pitch extraction, psychoacoustically motivated segmentation, and post-processing based on music theory constitute the basis for monophonic transcription. The investigation of polyphonic parts, as partial interference, pitch hypothesis or octave detection, prepares the fundamentals for subsequent implementations. The thesis concludes with the evaluation of the proposed system. A variety of different test series are described in the context of a metadata search engine. The results show the potential of the method, especially with regard to commercial applications.

Keywords:
music transcription, auditory periphery, Gestalt psychology, metadata

Zusammenfassung

Das Thema dieser Dissertation ist die Implementierung eines Verfahrens zur automatisierten Transkription von Musik. Die Fähigkeit des Menschen, insbesondere die von musikalischen Experten, bezüglich der Wahrnehmung musikalischer Inhalte kann von aktuellen technischen Systemen bei weitem nicht reproduziert werden. Einen plausiblen Ansatz, um diese Lücke für Anwendungen der automatisierten Musikanalyse zu schließen, stellt die Verwendung perzeptuell motivierter Strategien dar. Die vorliegende Arbeit wendet daher konsequent grundlegende Verarbeitungsmechanismen der menschlichen auditorischen Peripherie sowie kognitiv höher angesiedelter Gehirnzentren an.

In einer ausfürlichen Darstellung des Standes der Technik werden die aktuellen Algorithmen zur Bestimmung der Grundfrequenzen und zur Segmentierung musikalischer Phrasen sowie deren Anwendung in monophonen und polyphonen Melodietranskriptionssystemen erläutert.

Nach der Beschreibung der fundamentalen physiologischen Komponenten der auditorischen Peripherie und Prinzipien der Gestaltpsychologie werden die in dieser Arbeit verwendeten Modelle der teilweise aktiven Schallvorverarbeitung des Innenohres erläutert. Im Bereich der auditiven Weiterverarbeitung werden Prozesse der Frequenzwahrnehmung sowie ein auf gestaltbasierenden Annahmen aufgebautes eigenes Hierarchiemodell eingeführt.

Neben der Aufstellung dieses Hierarchiemodells besteht der Kernpunkt der Arbeit in der Implementierung der ausgewählten Modelle bezüglich monophoner und polyphoner Transkriptionsstrategien. Gehörgerechte Pitchextraktion, psychoakustisch motivierte Segmentierung und musiktheoretisch untermauerte Nachbearbeitung bilden die Basis einstimmiger Analyse. Die Untersuchung von Partialtoninterferenzen, polyphonen Pitchhypothesen und Ansätzen zur Oktaverkennung sollen als Grundlage weiterführender Arbeiten im mehrstimmigen Anwendungsfall aufgefasst werden.

Die Arbeit schließt mit der Evaluierung des Verfahrens anhand der Diskussion einer Anzahl verschiedener Testreihen im Umfeld eines Metadaten-Suchsystems. Die erhaltenen Ergebnisse verdeutlichen das (auch kommerzielle) Anwendungspotential der vorgestellten Methode.

Schlagwörter:
Musiktranskription, Auditorische Peripherie, Gestaltpsychologie, Metadaten

Danksagung

Herrn Prof. Dr. Karlheinz Brandenburg und Herrn Dr. Thomas Sporer gilt mein Dank für die Möglichkeit, meine Tätigkeit am Fraunhofer IDMT mit der Umsetzung dieser Arbeit zu kombinieren sowie für deren unterstützende Begleitung.
Herrn Dr. Distler danke ich für die umfassende und konstruktive Durchsicht der Inhalte.
Zu großem Dank verpflichtet bin ich meinen Kollegen am Fraunhofer IDMT für deren wissenschaftliche Unterstützung.
Insbesondere Herrn Andreas Brückmann danke ich für seine wertvollen programmiertechnischen Hinweise.
Herrn Rominger danke ich für die fortlaufende Richtigstellung der wesentlichen Dinge.

Nicht zuletzt und ganz besonders danke ich meiner lieben Frau Katrin, meiner kleinen Tochter Liv-Berit und meiner Mutter für die liebevolle unterstützende Begleitung meines bisherigen Lebensweges.

Und immer sind da Spuren,
und immer ist einer dagewesen,
und immer ist einer noch höher geklettert
als du es je gekonnt hast, noch viel höher.
Das darf dich nicht entmutigen.
Klettere, steige, steige.
Aber es gibt keine Spitze.
Und es gibt keinen Neuschnee.

(Kaspar Hauser, 1931)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

In Zeiten stetig wachsender Datenmengen in der multimedialen Informationsgesellschaft erwächst zunehmend der Bedarf an intuitiven und ergonomisch orientierten Schnittstellen zum effizienten Zugriff auf die gewünschten Inhalte. Herkömmliche Anwendungen beschränken sich auf die Benutzung textbasierter Beschreibungen oder signaltheoretischer Darstellungen (Spektralanalyse, etc.). Eine Strategie zur Verbesserung der Suche in großen Datenbanken besteht in der Einführung der semantischen Analyse von audio-visuellem Datenmaterial, d.h. aus den ursprünglichen Eingangsdaten werden für den Menschen bedeutungsvolle Repräsentationen extrahiert. Durch die Standardisierungen im Umfeld von MPEG-7 [mpe03] ist bereits ein Rahmenwerk geschaffen, das die Beschreibung, Verwaltung und auch die Indexierung der Inhalte in einheitlichem und portablem Format gewährleistet.

Motivation für die vorliegende Arbeit war die Implementierung eines sogenannten „Query-By-Humming"-Systems (QbH, Begriff vermutlich eingeführt von Ghias et al. [GLCS95]) am „Fraunhofer-Institut für digitale Medientechnologie IDMT" [idm04]. Die grundlegende Funktionsweise lässt sich anhand des Schemas in Abbildung 1.1 erläutern. Exemplarisch sei als repräsentativer Anwendungsfall die Suche nach einem im Alltagsgeschehen in den Medien wahrgenommenen „Ohrwurm" betrachtet. Man hört in Radio oder Fernsehen eine Melodie, versäumt aber die Information bezüglich Interpret und Titel. Ein praktikables QbH-System sollte nun in der Lage sein, aufgrund einer natürlichen Eingabe, die gesuchte Melodie in einer Datenbank wiederzufinden. Im üblichen Fall wird die QbH-Applikation über eine benutzerfreundliche Eingabe, der Anwender singt oder spielt mit einem Instrument in ein Mikrofon, mit einer Suchanfrage versehen. Ausgehend von der digi-

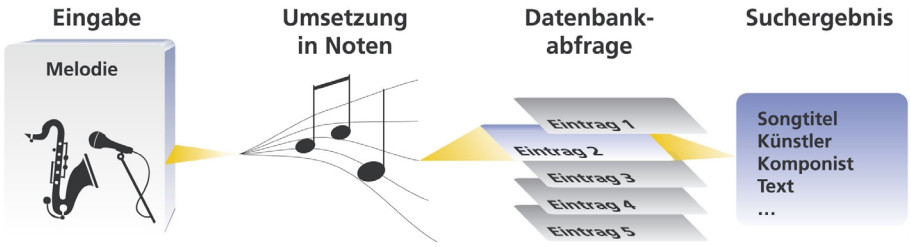

Abbildung 1.1: QbH - Funktionsschema (Quelle: [idm04])

talisierten Schallwellenform erfolgt die Konvertierung in musikalische Noten. Nach der Abfrage der gefundenen Notenfolge in einer Melodiedatenbank werden dem Musikinteressierten die Metadaten der ähnlichsten Einträge einer vorhandenen Melodiedatenbank in Form von Liedtitel, Sänger, Komponist, etc. dargeboten.

Ziel dieser Arbeit ist die Bereitstellung einer zuverlässigen und allgemeingültigen automatisierten Melodietranskription. Ansatzpunkt für die Wahl der hierfür zu benutzenden Verfahren war die Beobachtung, dass die Leistung des Menschen bezüglich der Informationsverarbeitung von komplexen Umweltreizen durch technische Systeme in der Regel bisher nur ansatzweise erreicht werden kann. So versuchen viele aktuelle Verfahren im Bereich der künstlichen Intelligenz, die durch die Evolution über Jahrtausende entwickelten menschlichen Strategien nachzuvollziehen. In dem für diese Arbeit speziell interessierenden Sektor der auditorischen Szenenanalyse, wie aber auch in anderen signalverarbeitenden Audiobereichen [Bre90][Bra99][Ste99] (Audiokodierung, Geräuschemission, etc.), fanden in der jüngsten Vergangenheit psychoakustische Modelle verstärkt Verwendung. In diesem Teilbereich der Psychophysik werden mathematische Beziehungen zwischen physikalischen (akustischen) Signalen und den entsprechenden psychophysischen Reaktionen durch Hörtests gewonnen [ZF01]. Diese werden dann beipielsweise zur Reduktion großer Audiodatenmengen oder auch zur gehörgerechten Beurteilung von Umweltgeräuschen benutzt.

Der interne Aufbau der beteiligten physiologischen Systeme wird dabei aber weitestgehend außer acht gelassen. Hier setzt das vorgestellte Verfahren an und erweitert diese Tendenzen konsequent durch die Benutzung von messtechnisch validierten Erkenntnissen der peripheren auditorischen Gegebenheiten des Menschen, d.h. Modelle der menschlichen Schallverarbeitung werden benutzt, um die zu analysierenden Audiosignale *physiologisch*

gehörgerecht zu interpretieren. Besondere Bedeutung findet dabei die im Innenohr stattfindende Transduktion der mechanisch-akustischen Schwingungsvorgänge in neuronale, elektrische Aktivierungsmuster auf dem Hörnerven.

Auf die explizite Nachbildung der Vorgänge im zentralen Gehör wird verzichtet, da in diesem Bereich die vorliegenden medizinischen Fakten entweder nur ansatzweise erforscht oder noch nicht ausreichend bestätigt sind. Zur Weiterverarbeitung der Ergebnisse der physiologischen Vorverarbeitung sollen daher die aus psychologischen Modellen bekannten kognitiven Strategien des Menschen, soweit wie das für die praktische Anwendbarkeit des Systems sinnvoll erscheint, nachvollzogen werden.

Es werden Ansätze der hierarchischen Informationsverarbeitung benutzt, die zur sukzessiven Generierung mental höherwertigerer und abstrakterer Strukturen dienen. Das Zusammenfassen von Einzelmerkmalen zu übergeordneten Objekten, die Einheiten als Summe ihrer Komponenten darstellen, lässt sich auf praktisch allen Ebenen der Verarbeitung sowohl in der visuellen als auch in der auditiven Wahrnehmung nachweisen. So bilden einzelne harmonische Partialtonkomponenten einen Ton, einzelne Noten ergeben in ihrer Gesamtheit eine Melodie. Besonderes Gewicht wird hierbei gelegt auf die Verwendung von sogenannten Gestaltgesetzen [And01], wonach diese Gruppierungsstrategien der elementaren Komponenten bestimmten Parametern wie Nähe, Ähnlichkeit, guter Fortsetzung oder auch gemeinsamem Schicksal folgen.

Als besondere Herausforderung erweist sich die Allgemeingültigkeit des zu entwickelnden Verfahrens bezüglich unterschiedlichster in der musikalischen Realität vorkommender Eingabemedien bzw. Instrumente. Im Einklang mit dem gewählten physiologisch basierten Ansatz werden die Analysestrategien entsprechend *wahrnehmungsbezogen* ausgelegt, d.h. Vorannahmen über spezielle instrumentenspezifische Schallerzeugungsmechanismen werden bewusst vermieden. Ausschließlich der perzeptive Eindruck der untersuchten Schallwellen soll als Grundlage der Transkription verwendet werden.

Die Untersuchungen beziehen sich auf monaurale Daten, obwohl durch die Einbeziehung von binauralen bzw. Stereoaufnahmen (Richtungsinformationen der beteiligten Schallquellen) eine Verbesserung der Ergebnisse erreicht werden sollte. Aufgrund des ohnehin schon enormen Rechenaufwandes, bedingt durch die Komplexität des verwendeten Ohrmodells, wird aber hierauf verzichtet.

Der Aufbau der Arbeit gliedert sich in die folgenden Abschnitte:

In Kapitel 2 findet sich eine ausführliche Beschreibung des Standes der Technik. Zunächst werden Verfahren der Tonhöhenbestimmung (Pitchextraktion) sowie der Segmentierung (Unterteilung der Frequenzverläufe in einzelne Noten) vorgestellt. Diese dienen so oder in ähnlicher Weise praktisch als Grundpfeiler in allen Melodietranskriptionssystemen. Anschließend erfolgt eine Beschreibung von Gesamtsystemen, die sowohl für separate, monophone Transkription als auch als Bestandteile komplexer „Query-By-Humming"-Systeme benutzt werden. Weiterhin werden aktuelle Verfahren vorgestellt, die sich der Transkription polyphoner Inhalte nähern.

Kapitel 3 fasst, soweit dies zum weiteren Verständnis notwendig ist, die physiologischen Grundlagen der menschlichen Hörwahrnehmung zusammen.

Die wichtigsten in dieser Arbeit verwendeten Modelle werden in Kapitel 4 vorgestellt. Im Bereich der Physiologie beinhalten diese die Funktionsschemata für Außen- und Mittelohr, Innenohr („Erweitertes Analogmodell") sowie innere Haarzellen. Des Weiteren werden die Wirkungsweisen der auditiven Frequenzanalyse („Phase-Locking") und ein adaptiertes Hierarchiemodell nachgebildet.

Die detaillierte Beschreibung der in der Implementierung benutzten Verfahren erfolgt in Kapitel 5. Anhand der Analyse einer von Gesang und Klarinette dargebotenen Beispielmelodie werden die Grundfrequenzerkennung (Pitchextraktion), die psychoakustisch motivierte Unterteilung der erhaltenen Tonhöhenverläufe in einzelne Noten (Segmentierung) sowie eine interpretativ basierte Nachverarbeitungsstufe vorgestellt. Weiterhin finden sich hier die Strategien zur Szenenanalyse polyphoner Musik am Beispiel eines einfachen zweistimmigen Holzbläserfragments.

In Kapitel 6 wird schließlich der Nachweis der ausgereiften Praxistauglichkeit des Verfahrens bezüglich des Einsatzes in kommerziellen Anwendungen gebracht. Die Evaluierung erfolgt durch eine Auswahl statistisch aussagekräftiger Testreihen im Vergleich mit konkurrierenden Transkriptionsansätzen im Rahmen des Fraunhofer „Query-By-Humming"-Systems [idm04]. Weiterhin wird ein unabhängiger externer Test der Universität Ghent, Belgien, vorgestellt, der die Konkurrenzfähigkeit der physiologisch basierten Anwendung unterstreicht.

Kapitel 2

Stand der Technik: Etablierte Verfahren und aktuelle Anwendungen

2.1 Pitchextraktion

Die automatisierte Bestimmung der Frequenzinhalte von Schallsignalen ist ein Teilbereich der Signalverarbeitung mit langer Tradition. So finden sich auch unzählige Verfahren zur Extraktion der (wahrgenommenen) Tonhöhe pitchbehafteter Anteile von Sprache und Musik. Die meisten dieser Algorithmen sind im Umfeld der Spracherkennung entstanden, lassen sich aber auch in der Musikanalyse anwenden. Gemeinsam ist fast allen Verfahren die Suche nach Periodizitäten. Diese Vorgehensweise stützt sich auf die Annahme, dass pitchbehaftete Signale sich im Regelfall aus harmonischen Partialtönen zusammensetzen.

Die Ansätze lassen sich in drei Bereiche aufteilen:
1) Verfahren im Zeitbereich suchen nach Regelmäßigkeiten direkt in der Schallwellenform;
2) Verfahren im Frequenzbereich transformieren das Eingangssignal in die Frequenzdomäne und werten die erhaltenen Koeffizienten aus;
3) kombinierte Zeit-Frequenzverfahren zerlegen zunächst das Signal in einzelne Frequenzbänder und wenden dann Zeitbereichsalgorithmen auf die einzelnen Filterkanäle an.

Eine Reihe von Unwägbarkeiten machen allerdings das Vorhaben, ein

allgemeingültiges Verfahren zu implementieren, zu einer nichttrivialen Aufgabe: schwach ausgeprägter oder fehlender Grundton („virtueller Pitch“), Abweichungen der Partialtonfrequenzen von der idealen harmonischen Obertonreihe („Inharmonizitäten“) und andere Effekte sind beispielsweise für das Hauptproblem der Oktavvertauschung mitverantwortlich.

Aufgrund der großen Zahl unterschiedlicher Pitcherkennungsverfahren kann hier nur eine Auswahl der etabliertesten Methoden beschrieben werden.

Das klassische Verfahren der Grundfrequenzbestimmung im Zeitbereich findet sich in der Benutzung der Autokorrelationsfunktion (AKF) [Rab77b]. Die Autokorrelationsfunktion beschreibt ein Maß für die Ähnlichkeit eines Signals mit sich selbst in Abhängigkeit von einem Verschiebungsparameter τ. Sie ist im kontinuierlichen Bereich definiert als:

$$AKF(\tau) = \int_{-\infty}^{\infty} f(t)f(t+\tau)dt. \tag{2.1}$$

Das zu korrelierende Signal f(t) und seine identische Kopie werden sequentiell zeitlich gegeneinander verschoben und multipliziert, anschließend wird über die entstandene Kurve integriert. Beinhaltet das Signal eine Periodizität, äußert sich diese als eine Folge ausgeprägter Maxima im Abstand der gesuchten Grundperiode in der Autokorrelationsfunktion.

Umfangreiche Untersuchungen zur Verwendung stellen Brown et al. in ihren Arbeiten [BP89][BZ91] vor. Neben der konventionellen Autokorrelation werden Modifikationen zur Verbesserung der Methode beschrieben. Die Autoren weisen nach, dass mit der Einführung von Mischtermen der Form $f(t)f(t+2\tau)$, $f(t)f(t+3\tau)$, ... in das AKF-Integral die Auflösung verbessert werden kann („narrowed autocorrelation“). Desweiteren wird die Verwendung der sogenannten *invertierten* Autokorrelation diskutiert. Hierbei sucht man nach der mit τ periodischen Funktion

$$p(t,\tau) = \frac{f(t) + f(t+\tau) + f(t+2\tau) + ... + f(t+(N-1)\tau)}{N}, \tag{2.2}$$

die f(t) am besten approximiert. In der über die Zeit gemittelten Differenzfunktion

$$E(\tau) = \left\langle (f(t) - p(t,\tau)^2 \right\rangle \tag{2.3}$$

wird über signifikante Minima die Grundperiode bestimmt.

In Experimenten mit repräsentativen Mitgliedern der Instrumentengruppen Tasten-, Streich- und Holzblasinstrumente werden subjektiv gute, wenn auch nicht einheitliche, Ergebnisse bezüglich der gefundenen Fundamentalfrequenzen angegeben. Somit kristallisieren sich je nach Instrument jeweils unterschiedliche Modifikationen der Autokorrelation als optimal heraus.

Das Verfahren von Choi [Cho97] strebt eine zuverlässige Bestimmung der Fundamentalfrequenz auch für sehr kurze Signalausschnitte an. Motivation ist die Implementierung von interaktiven, echtzeitfähigen Musikanwendungen. Seine „Least-Square Fitting"-Methode kommt dabei ohne Fensterfunktion aus, wie sie bei den konventionellen Spektralanalysen benötigt wird.
Der Algorithmus setzt voraus, dass sich das zu untersuchende Signal aus der Summe harmonischer Komponenten ω_k zusammensetzt ($k = 1..N$: Zeitindex). Aufgabe des Verfahrens ist nun, den Fehler einer Sinus-Testfunktion

$$\hat{\omega}_k = a \cdot sin(fk) + b \cdot cos(fk) \tag{2.4}$$

bezüglich des Signals zu minimieren. Die Koeffizienten a und b bestimmen Amplitude und Phase der Testfunktion. Der quadratische Fehler e ergibt sich zu:

$$e = \sum_{k=1}^{N} \left(\hat{\omega}_k - \omega_k\right)^2 . \tag{2.5}$$

Über die Ableitung der Fehlerfunktion nach den Koeffizienten a und b lässt sich diese in reiner Abhängigkeit von der Testfrequenz darstellen:

$$\left(\frac{\delta e}{\delta a} = 0\right) \wedge \left(\frac{\delta e}{\delta b} = 0\right) \Rightarrow a^*(f), b^*(f). \tag{2.6}$$

Die modifizierte Fehlerfunktion

$$e^*(f) = \sum_{k=1}^{N} \left(\hat{\omega}_k(a^*(f), b^*(f)) - \omega_k\right)^2 . \tag{2.7}$$

bestimmt den minimalen quadratischen Fehler für die Frequenz über alle Amplituden und Phasen. Ihr Verlauf zeigt charakteristische Einbrüche bei den Partialtonfrequenzen, woraus sich Spektrum und Grundtonfrequenz bestimmen lassen.

Testergebnisse werden vorgestellt für Gitarrentöne im Oktavabstand im Bereich von 98 - 784Hz (G2-G5) mit Segmentlängen von 5 - 30 ms. Für die

hohen Töne ergeben sich bis zur kleinsten Segmentlänge zufriedenstellende Ergebnisse, während die beiden tiefen Töene ab ca. 10 ms in Richtung kleinerer Segmentabschnitte zunehmend schlechtere Ergebnisse zeigen.

Die Adaptierung eines Verfahrens aus der Bildverarbeitung wird von Richter [Ric01] beschrieben. Die Hough-Transformation [BFRR95] ist eine Methode zur Mustererkennung. Sie kann eingesetzt werden, wenn es darum geht, Abbildungen eines vorher bekannten Mustergegenstandes in einer grapfischen Darstellung zu finden. In der Anwendung auf die Grundfrequenzerkennung sollen im Audiosignal die aufsteigenden Flanken der fundamentalen harmonischen Schwingung detektiert werden. Dazu wird das Eingangssignal in einen Parameterraum abgebildet, d.h. für jeden diskreten Signalwert (t_i, y_i) bestimmt sich die Hough-Transformation wie folgt:

$$\frac{1}{A} = \frac{1}{y_i} \cdot sin(\omega_c t_i - \varphi). \tag{2.8}$$

Bei ω_c handelt es sich um die empirisch bestimmte „Center-Frequenz“, die mit 261 Hz angegeben wird. Somit ergibt sich für jeden Sample eine Sinus-Welle mit Amplitude A. Diese wird im Phasenbereich von 0 bis $\frac{\pi}{2}$ im Abstand der Abtastperiode fortschreitend in ein zweidimensionales Histogramm (Hough-Matrix (HM)) eingetragen. Bei deutlich ausgeprägten Flanken akkumulieren sich zusammengehörige Einzeleinträge zu signifikanten Maxima im Parameterraum (A, φ). Aus dem Abstand der gültigen Extremwerte lässt sich dann die Grundfrequenz bestimmen. Im Vergleich zu Verfahren, die nur einzelne Punkte im Signal zur Charakterisierung der Grundperiode heranziehen (z.B. Nulldurchgangsrate ZCR [Lar77]), lässt sich somit eine gewisse Robustheit gegen Störungen durch die Berücksichtigung kompletter Signalflanken erreichen.

Experimente zum Verfahren zeigen unterschiedliche günstige Modellparameter (Maxima-Schwellwert, etc.) für bestimmte Instrumente. Die folgenden Resultate (Prozent richtig erkannter Tonhöhen) werden mit Klangbeispielen der „McGill Master Samples“ [OW87] bei Wahl optimaler Einstellungen erreicht: Piano 78,4 %, Flöte (vibrato) 97,3 %, Oboe 87,5 %, Klarinette (Bb) 86,5 %.

Als letzter exemplarischer Vertreter der Pitcherkennungsverfahren im Zeitbereich soll die „Average Magnitude Difference Function“ (AMDF) [RSC+74]

vorgestellt werden, die definiert ist als:

$$AMDF(\tau) = \sum_{i=1}^{N} |s(i) - s(i+\tau)| \,. \tag{2.9}$$

Nahe verwandt zur oben erläuterten Autokorrelationsfunktion wird hierbei jedoch nicht das Produkt, sondern die Betragsdifferenz voneinander verschobener Kopien gebildet. Die übliche Vorgehensweise sucht somit nach lokalen Minima unterhalb eines festzulegenden Schwellwertes, die Hinweise auf die Existenz von Periodizitäten im Signal liefern. Die kleinen notwendigen Fensterbreiten und das Fehlen von Multiplikationen prädestiniert das Verfahren für effiziente Rechnerimplementierungen.

Optimierung erfährt die AMDF-Methode in der Arbeit von de Cheveigné [dC02]. Zunächst überführt dieser die AMDF in eine „Differenzfunktion“:

$$d_t(\tau) = \sum_{i=1}^{N} \left(s(i) - s(i+\tau)\right)^2 \,. \tag{2.10}$$

Entwicklung des quadratischen Terms unter Mitnahme der ersten drei Glieder erlaubt die Formulierung als Autokorrelationsformel:

$$d_t(\tau) \simeq AKF_t(0) + AKF_{t+\tau}(0) - 2 \cdot AKF_t(\tau). \tag{2.11}$$

Zur Vermeidung von Oktavvertauschungen, die das Hauptproblem aller Pitchextraktionsverfahren darstellen, wird schließlich durch das Mittel der vorher berechneten kleineren Verzögerungswerte dividiert. De Cheveigné bezeichnet dies als „Cumulative mean difference function“ (CMDF):

$$d'_t(\tau) = \begin{cases} 1 & : \quad \tau = 0 \\ d_t(\tau) / \left(\frac{1}{\tau} \sum_{i=1}^{N} d_t(i)\right) & : \quad sonst. \end{cases} \tag{2.12}$$

Drei weitere Modifikationen der üblichen AMDF-Strategie runden das Verfahren ab: 1) Nicht das absolute sondern das erste Minimum bezogen auf die Verzögerungszeit soll die gesuchte Periodizität repräsentieren; 2) Parabolische Interpolation der Umgebung der Extremwerte verbessert die diskrete Auflösung der Periodendauer; 3) die Pitchhypothesen werden durch zeitlich benachbarte Einschätzungen verworfen oder bestätigt.

Evaluierung der Implementierung erfolgt mit insgesamt 1,9 h Sprache 48 mehrsprachiger Testpersonen. Als Referenz dient das mit Elektroden am

Kehlkopf (Larynx) abgenommene Signal. Akzeptiert man einen Fehler von 20 % bezogen auf die Referenzfrequenz, werden 99 % richtig erkannt. Bei einem Fehler von 5 % beträgt die Trefferquote 94 %, bei 1 % Toleranz sinkt die Erkennungsrate auf 60 %.

Die klassische Methode zur Pitcherkennung im Frequenzbereich ist die Verwendung des Cepstrum [Nol70][Fel84], dessen Wurzeln und hauptsächliche Anwendungen im Bereich der Sprachverarbeitung liegen. Das Verfahren basiert auf der Annahme, dass sich das Spektrum von Sprache und den meisten Instrumentenklängen als Summe aus Anregungsvorgang (z.B. schwingende Saite bei Streichinstrumenten) und Resonanzverhalten (Vokaltrakt, bzw. Instrumentenkorpus) zusammensetzt [FR98][Ben90]. Das Cepstrum ist definiert als die inverse Fouriertransformierte des zur Basis 10 logarithmischen Fourier-Betragsspektrum.

Der Anregungsvorgang kann in der Regel erklärt werden als Sequenz quasi-periodischer Impulse. Fouriertransformation der Anregung ergibt ein Linienspektrum von Harmonischen der Anregungsfrequenz. Logarithmische Betragsbildung verliert zwar die Phaseninformationen, erhält aber die Größenverhältnisse der Partialtonamplituden. Eine anschließende inverse Fouriertransformation ergibt eine wieder quasi-periodische Wellenform von Impulsen.

Das Resonanzverhalten des schallverbreitenden Instrumentenkorpus bzw. Vokaltraktes kann interpretiert werden als ein über die Frequenz kontinuierlicher Filter, mit dem das Anregungsspektrum gefaltet wird. Es lässt sich zeigen, dass bei einem Abfall der zugehörigen Impulsantwort mit $1/t$ die Cepstrum-Wellenform mit $1/t^2$ abnimmt. Somit bündelt das Cepstrum die Impulsantwort des Resonanzfilters in einem kleinen Bereich am Anfang des Analyseergebnisses; Anregung und Resonanz werden also gleichsam „entfaltet“. Aus den im Abstand der Grundperiode angeordneten Spitzenwerten der Cepstrum-Wellenform kann abschließend die Grundfrequenz ermittelt werden.

Einen typischen Vertreter für eine Vielzahl von Anwendungen stellt das Modell von Brown dar [Bro92], welches versucht, das Ergebnis einer Spektralanalyse in Einklang mit einer idealen harmonischen Obertonreihe zu bringen. Im speziellen Fall wird mittels sogenannter „Konstant-Q-Transformation“ das Spektrum ermittelt [BP92], woraus sich ein Muster der harmonischen Komponenten in der logarithmisch dargestellten Frequenzdomäne ergibt. Die Fre-

quenztransformation entspricht einer Filterbank mit relativen konstanten Filterbreiten von 1/24 Oktaven. Für harmonische Komponenten ist der Abstand somit unabhängig von der Grundfrequenz. Anschließend wird die Kreuzkorrelationsfunktion zwischen dem Transformationsergebnis und einer idealen Musterobertonreihe als Funktion der Fundamentalfrequenz berechnet. Das Maximum bestimmt die Einschätzung der momentanen Pitchfrequenz. Ein Vorteil des Verfahrens ist unter anderem die mögliche Bestimmung eines „virtuellen Pitch", d.h. auch bei fehlendem Grundton wird dem Tonkomplex dieser als Grundfrequenz zugeordnet.

Die von Brown angegebenen Testresultate deuten aber auf das den meisten Pitchtracker-Systemen inhärente Problem der Oktavvertauschung hin; die meisten Fehler bei der Grundfrequenzanalyse liegen zweifellos in der Fehleinschätzung bezüglich Ober- und Suboktave. Testergebnisse werden angegeben für Flöte, Geige und Piano, als typische Vertreter ihrer jeweiligen Instrumentengruppe (Holzbläser, Streicher und Tasteninstrumente). Chromatische Skalen werden subjektiv beurteilt als die Abweichung von horizontalen Stufenlinien in der Frequenz-Zeit-Darstellung. Verschiedenartige Ergebnisse werden interpretiert als Folge unterschiedlicher Partialtonstärken. Als kritischer Paramter wird unter anderem die Anzahl der in der idealen Obertonreihe verwendeten Einträge angegeben.

Die Arbeit von Doval und Rodet [DR93] basiert auf einem Wahrscheinlichkeitsmodell pseudo-periodischer Signale. Der Ansatz sucht nach dem perfekt periodischen Signal, dessen Modellparameter das zu analysierende Signal mit größtmöglicher Wahrscheinlichkeit annähern. Zunächst wird eine „Observation" definiert als der Satz der Partialtöne einer Kurzzeitfouriertransformation (STFT) inklusive nichtharmonischer Maxima. Diese „Observation" kann dann im Sinne einer Wahrscheinlichkeitsbetrachtung beschrieben werden durch eine Anzahl von Zufallsvariablen: Grundfrequenz f_0, spektrale Amplitudenhüllkurve, Vorhandensein und Nichtvorhandensein des k-ten Partialtons, Anzahl harmonischer und nichtharmonischer Partialtöne. Weiterhin wird jedem Partialton eine Qualität bezüglich seiner Nähe zur idealen Harmonischen zugeordnet und anschließend für jede mögliche Grundfrequenz die Wahrscheinlichkeit bezüglich der aktuellen „Observation" bestimmt. Die resultierende Wahrscheinlichkeitsdichte ergibt sich aus der Summe der oben beschriebenen Einzelmodell-Wahrscheinlichkeitsdichten.

Zur Detektion einer zeitlichen Pitchtrajektorie nutzen die Autoren ein „Hidden-Markov-Modell" (HMM). Einzige Erweiterung ist dabei die Einfüh-

rung von Übergangswahrscheinlichkeiten zwischen aufeinanderfolgenden Zuständen. Diese sei klein, wenn die Differenz der Grundfrequenz große Abstände aufweist. Die resultierende Trajektorie findet man dann durch die optimale Zustandssequenz mit größtmöglicher Gesamtwahrscheinlichkeit.

Evaluierung erfolgt über eine Datenbank von 20 kontinuierlich gesprochenen Sätzen, welche einer Gesamtheit von ungefähr 20000 Analyseframes entsprechen. Ein allgemeingültiger Ansatz klassifiziert 4,3 % der Einträge falsch, während Vorannahmen über das Signal (Frequenzbereich, etc.) den Fehler auf 3,1 % verringern. Einbeziehung der HMM Tracking-Methode führt zu einem Fehler von nur noch 1,6 %.

Cano [Can98] adaptiert und modifiziert das ursprünglich von Maher und Beauchamp [MB94] entwickelte „Two-way mismatch"-Verfahren (TWM). Man versucht hierbei, die harmonische Obertonreihe zu finden, welche die berechneten Partialtonfrequenzen am besten annähert.

Aus der Kurzzeit-Fouriertransformation (STFT) erhält man über das Betragsspektrum eine Schar potentieller Partialtonkandidaten. Für jede mögliche Fundamentalfrequenz wird daraufhin der Abstand zwischen idealer harmonischer Obertonreihe und den berechneten Einträgen bestimmt. Die Diskrepanz zwischen gemessenen und „vorhergesagten" Teiltönen wird als „Anpassungsfehler" bezeichnet. Die Fehlerberechnung teilt sich in zwei Abschnitte. Im ersten Teil werden die Abweichungen der gemessenen Partialtonfrequenzen von ihren nächsten Nachbarn der idealen harmonischen Struktur bestimmt. Ergebnis ist der sogenannte „Predicted-to-measured"-Anpassungsfehler:

$$\begin{aligned} Err_{p\to m} &= \sum_{n=1}^{N} E_\omega(\Delta f_n, f_n, a_n, A_{max}) \\ &= \sum_{n=1}^{N} \Delta f_n \cdot (f_n)^{-p} + \left(\tfrac{a_n}{A_{max}}\right) \cdot [q\Delta f_n \cdot (f_n)^{-p} - r]. \end{aligned} \tag{2.13}$$

Der zweite Schritt berechnet umgekehrt den Abstand zwischen den vorhergesagten idealen Harmonischen und den jeweils nächsten gemessenen Partialtönen und mündet im „Measured-to-predicted"-Anpassungsfehler:

$$\begin{aligned} Err_{m\to p} &= \sum_{k=1}^{K} E_\omega(\Delta f_k, f_k, a_k, A_{max}) \\ &= \sum_{k=1}^{K} \Delta f_k \cdot (f_k)^{-p} + \left(\tfrac{a_k}{A_{max}}\right) \cdot [q\Delta f_k \cdot (f_k)^{-p} - r], \end{aligned} \tag{2.14}$$

wobei a_i und f_i den Amplituden und Frequenzen der involvierten Partialtöne i entsprechen. Δf_i misst jeweils die Differenz zwischen gemessenem Partialton

und dem nächsten vorhergesagten Teilton und umgekehrt. A_{max} ist die größte gemessene Amplitude. Die Konstanten p, q, r und ρ (s.u.) sind experimentell ermittelte Werte, die das Verfahren optimieren. Der Gesamtfehler für die möglichen Grundfrequenzen ergibt sich somit als Kombination der beiden Fehler:

$$Err_{total} = \frac{Err_{p \to m}}{N} + \rho \cdot \frac{Err_{m \to p}}{K} \tag{2.15}$$

Aufgrund der negativen Bewertung von nichtharmonischen sowie fehlenden Partialtönen hilft die zweigleisige Fehlerbestimmung, die üblichen Oktavfehler signifikant einzuschränken.

Die speziellen Anpassungen von Cano beinhalten pitchabhängige Fensterbreite, geschickte Vorauswahl benutzter Frequenzpeaks und möglicher Fundamentalfrequenzen sowie Miteinbeziehung benachbarter Fenster zur Hypothesenstützung oder -abschwächung.

Eine Methode, die auch bei starkem Hintergrundgeräusch („Multi-talker noise“, ...) und spektraler Verzerrung (Telefonie, Nachhall, ...) eine zuverlässige Bestimmung der Pitchfrequenz gewährleisten soll, stellen Nakatani und Irino vor [NI02]. Sie bedienen sich der sogenannten „Instantaneous frequency“ (IF), d.h. der zeitlichen Ableitung der Phasendifferenz $\dot{\phi}(\omega_c)$ zweier aufeinanderfolgender Frames eines Frequenz-Bins ω_c der Kurzzeit-Fouriertransformation (s.a. [BP93]). Koinzidiert eine dominante Frequenzkomponente mit einem Frequenz-Bin („Fixed point“: $\dot{\phi} = \omega_c$), so haben auch die IF der benachbarten Bins nahezu den gleichen Wert. In der Darstellung der Frequenz-Bins gegen die IF erhält man somit näherungsweise eine Treppenfunktion, auf deren Stufen das Verfahren die Partialtonfrequenzen annimmt. Aus der Differenz benachbarter „Fixed point“-Einträge lässt sich dann die Grundfrequenz bestimmen. Kernpunkt der Implementierung ist die Berechnung der sogenannten „Degree of Dominance“- Funktion $D_0(\omega_c)$:

$$D_0(\omega_c) = log\left(\frac{1}{B(\omega_c)^2}\right) \tag{2.16}$$

mit

$$B(\omega_c)^2 = \frac{\int_{\omega_c - \frac{\Delta\omega}{2}}^{\omega_c + \frac{\Delta\omega}{2}} (\dot{\phi}(\omega) - \omega_c)^2 S(\omega) d\omega}{\int_{\omega_c - \frac{\Delta\omega}{2}}^{\omega_c + \frac{\Delta\omega}{2}} S(\omega)^2 d\omega}. \tag{2.17}$$

Unter Benutzung der IF für Frequenz-Bin ω und Betragsspektrum $S(\omega)$ bestimmt die Funktion $B(\omega_c)^2$ das gewichtete Mittel der quadrierten Differenz

zwischen Frequenz-Bin ω_c und den IF innerhalb einer Umgebung $\Delta\omega$. Bei Koinzidenz der betrachteten Frequenzen und ebenem Verlauf der Treppenfunktion wird $B(\omega_c)^2$ minimal und die „Degree of Dominance"- Funktion liefert ein Maximum. Über die dominanten Einträge harmonisch verwandter Partialtonmaxima wird schließlich die Grundfrequenz abgeschätzt.

Evaluierung erfolgt mittels einer Datenbank von 840 gesprochenen Eingaben. Als Referenz wird analog zu [dC02] das Kehlkopfsignal der Probanden herangezogen. Abweichungen von weniger als 5 % der berechneten Frequenz zur Referenz werden als korrekt klassifiziert. In Abhängigkeit des Signal-Rausch-Abstandes (SNR = Signal-to-Noise-Ratio) bezüglich zugefügten weißen Rauschens und „Multi-talker noise" ergeben sich im Vergleich zum konventionellen Cepstrum-Verfahren je nach Störabstand Verbesserungen der Erkennungsraten bis zu 10 %. Ähnliche Ergebnisse werden für spektral verzerrte Signale erzielt.

Ein „kombiniertes" Verfahren, welches das frequenztransformierte Signal mit Zeitbereichsmethoden weiter auswertet, wird von van Immerseel und Martens [IM92] beschrieben. Details hierzu finden sich in der Darstellung des Transkriptionssystems von Clarisse et al. [Lem02] in Kap. 2.3. Ebenso „kombiniert" zeigt sich der in der dieser Arbeit entwickelte und streng physiologisch basierte Algorithmus, der die vom Innenohr durchgeführte Resonanzfilterung mit Korrelationsmethoden weiterverarbeitet (s. Kap. 5.1.1).

Wie schon oben angedeutet beschäftigt sich eine Vielzahl weiterer Verfahren mit der Erkennung der Grundfrequenz, bzw. der wahrgenommenen Tonhöhe (Pitch) realer nicht-perkussiver Signale. Abschließend sollen noch einige repräsentative Vertreter bestimmter Herangehensweisen nicht unerwähnt bleiben. Eine gebräuchliche Methode, die die Anzahl der Nulldurchgänge im Zeitsignal auswertet, wird vorgestellt von Larrson [Lar77] (ZCR = „Zero Crossing Rate"). Fitch und Shabana [FS99] bedienen sich der Wavelet-Transformation [Dau96], die eine variable Frequenzauflösung über den betrachteten Spektralbereich berücksichtigt. Barnard, et al. [BCVA91] benutzen ein neuronales Netz, Slaney [Sla90] beschreibt ein perzeptuell basiertes Modell. Interessant sind ebenso die Implementierungen von Piszczalski und Galler [PG79], Ney [Ney82], Lane [Lan90] und Kuhn [Kuh90].

2.2 Segmentierung

Dem Teilaspekt der Segmentierung ist als separate Disziplin bei weitem nicht so viel Aufmerksamkeit geschenkt worden wie der vorher besprochenen Pitcherkennung, obwohl sich eine korrekte Unterteilung der im ersten Schritt einer Melodietranskription herausgebildeten Pitchtrajektorien in einzelne musikalische Noten als mindestens genauso wichtig herausstellt. Vielmehr trennt sich nach Einschätzung des Autors die Qualität von Gesamtsystemen zur automatisierten musikalischen Notationsausgabe in der Regel genau an diesem Merkmal, was auch durch die in Kapitel 6.2 vorgestellten Testergebnisse unterstrichen wird.

Es gibt eine Reihe hierarchischer Abstraktionsebenen, in die ein musikalisches Audiosignal segmentiert werden kann. Eine erste Möglichkeit stellt die Einteilung in unterschiedliche Schallquellen dar. So könnte man beispielsweise in gesungene und instrumentale Abschnitte segmentieren. Neben einer solchen klangfarbenrelevanten Unterteilung liegt natürlich eine musikalische bzw. musiktheoretische Abgrenzung der einzelnen Anteile nahe. Je nach Genre sind hier verschiedene Teilmengen möglich: für liedhafte Musikstücke kann man nach Strophen und Refrain aufteilen, bei klassischer Musik sind bestimmte logische Abschnitte innerhalb einer Satzstruktur zu bestimmen. In einer weiteren tieferen Ebene ist es von Interesse, den rhythmischen Aufbau untergeordneter Abschnitte zu analysieren. Dazu muss das Signal auf perzeptive Schwerpunkte mit zeitlich stationärer Korrelation hin untersucht werden. Für die vorliegende Arbeit liegt das Hauptaugenmerk aber in einer noch grundlegenderen Stufe der Segmentierung. Die musikalisch kleinsten sinnvollen Einheiten stellen die einzelnen Noten dar, aus denen durch mannigfaltigste Kombination die schier unendlich erscheinende Motiv- und Melodievielfalt der Musik erzeugt werden kann. Abgesehen von verschiedenen musikalisch ausdrucksbehafteten Parametern wie Dynamik oder Klangfarbe kann eine Note in der höchsten Abstraktionsstufe durch drei Werte beschrieben werden: neben dem im vorherigen Abschnitt beschriebenen Pitch (wahrgenommene Tonhöhe) sind Notenanfang (Onset-Zeitpunkt) und Notenende (Offset-Zeitpunkt) von Bedeutung.

Für die charakterisierende rhythmische Beschreibung von Melodien ist dem Notenanfang die entscheidende Bedeutung zuzuordnen, zumal das Notenende durch physikalische Einflüsse wie Nachschwingen der schallerzeugenden Körper oder raumakustischen Nachhall häufig zeitlich nicht eindeutig aufgelöst werden kann. Der Onsetzeitpunkt wird unterschieden nach physi-

kalischem Beginn der Schallerzeugung und wahrgenommenem Notenbeginn, welche sich unter anderem durch spektrale und zeitliche Verdeckungseffekte [ZF01] voneinander abgrenzen können [VR81]. Der perzeptive Onset ist für die wahrnehmungsgerechte Segmentierung von Audiosignalen von primärem Interesse, da er beispielsweise das wesentliche Element für die musikalische Synchronität im Ensemblespiel darstellt.

Die grundlegende Methodik zur Detektion von Notenanfängen besteht in der Suche nach Intensitätsflanken oder auch plötzlichen Wechseln von Klangfarbe und/oder Pitch. Hauptproblem dabei ist es, graduelle Variationen innerhalb einer Einheit (Vibrato, Klangfarbenänderung, etc.) von wirklichen perzeptiv akzeptierten Onsets zu unterscheiden.

Ein frühes Verfahren zur automatisierten Segmentierung findet sich in der Arbeit von Schloss [Sch85]. Obwohl nicht mehr ganz neu, soll es doch erwähnt werden, da einige aktuelle komplexere Systeme (Kapitel 2.3) den hier vorgestellten Algorithmus im Zeitbereich als eines von mehreren Features zur Segmentierung nutzen. Aus dem gefilterten und geglätteten Zeitsignal wird die Amplitudenhüllkurve als Mittelwert über 20 ms breite Signalausschnitte berechnet. Über eine 4-Punkte lineare Regression wird die Hüllkurvensteigung bestimmmt; Peaks im Steigungsverlauf signalisieren Kandidaten für Onsets. Diese Methode eignet sich insbesondere gut für perkussive Darbietungen aufgrund der dort anzutreffenden starken Anstiegsflanken.

Eine Methode zur Bestimmung von Onsets, die ebenfalls in Gesamtsystemen (Kapitel 2.3) zur Anwendung kommt, wird von Masri und Bates [MB96] vorgestellt. Ursprünglich zur verbesserten Resynthese von transienten Einschwingvorgängen entwickelt, erweist sich die Nutzung der „High Frequency Content Function“ (HFCF) als probates Mittel bezüglich der Segmentierungsproblematik. Ansatzpunkt ist die Annahme, dass bei abrupten Änderungen im Signalverlauf aufgrund von Phasendiskontinuitäten das Frequenzspektrum zum Zeitpunkt der Änderungen von einem starken hochfrequenten Anteil bestimmt wird. Hierzu wird das Signal mit Hilfe der Kurzzeit-Fourier-Transformation (STFT) in die Zeit-Frequenzdarstellung überführt. Die Energiefunktion $E(t)$ in Abhängigkeit von der Zeit ergibt sich somit zu:

$$E(t) = \sum_{k=2}^{N/2+1} \left(|X(k,t)|^2\right) \tag{2.18}$$

($X(k)$: k-ter Frequenz-Bin der Fourier-Transformation).

Zur Messung des Anteils der hochfrequenten Energieanteile werden die einzelnen Komponenten der Energiefunktion mit ihrer Ordnungszahl gewichtet:

$$HFC(t) = \sum_{k=2}^{N/2+1} (|X(k,t)|^2 \cdot k). \tag{2.19}$$

Zur Detektierung von Onsets wird schließlich das normierte relative Verhältnis zweier aufeinanderfolgender HFC-Einträge mit einem Schwellwert verglichen:

$$\frac{HFC(t)}{HFC(t-\Delta t)} \cdot \frac{HFC(t)}{E(t)} > T_O, \tag{2.20}$$

Δt entspricht dem zeitlichen Abstand zweier aufeinanderfolgender Spektren. Überschreitet die Detektierungsfunktion den Schwellwert T_0, so wird der Beginn eines transienten Einschwingvorganges angenommen. Einträge innerhalb eines bestimmten Zeitabstandes werden zu einem Onset fusioniert.

Einen stark auditorisch motivierten Ansatz verfolgt Smith [Smi96], der versucht, die Verarbeitungsschritte der peripheren menschlichen Gehörelemente nachzuvollziehen. Eine detaillierte Beschreibung der physiologischen Grundlagen findet sich in Kapitel 3. Als auditorisches „Frontend“ verwendet er eine gehörgerechte Patterson-Holdsworth-Filterbank [PH90], die das akustische Signal in 28 Frequenzbänder zerlegt. Diese Bandsignale werden anschließend als extreme Simplifikation der inneren Haarzellen des Innenohrs gleichgerichtet und als Näherung der Aktivität auf den Hörnerven verstanden. Faltung mit einem „Gauss-Filter“ detektiert bei stark positivem Output eine deutliche Erhöhung der vorhandenen Intensität. Als weiterer physiologisch motivierter Verarbeitungsschritt höherer kognitiver Strukturen wird ein neuronales „Integrate-and-Fire“-Netzwerk [MS90] eingeführt. Dieses besteht aus einzelnen Einheiten, die ihren gewichteten Input akkumulieren und bei Überschreiten einer Schwellwertaktivität einen Nervenimpuls auslösen:

$$\frac{dA(t)}{dt} = I(t) - \gamma A(t), \tag{2.21}$$

mit Aktivität A(t), Input I(t) und Verlustfaktor γ. Pro Bandsignal wird ein Neuron zugeordnet. Gewichtete Signalzuführung und Rückkopplung der Impulse auf benachbarte Neuronen sorgen für eine Synchronisierung der Feuerraten. Fallen 6 Band-Onsets in ein 10 ms-Fenster wird ein neuer Notenanfang angenommen.

Neben einer kurzen gesprochenen Sequenz werden die Resultate anhand zweier Melodien von Flöte und Gitarre veranschaulicht. Die zugehörigen Darstellungen zeigen für die Flöte tendenziell die richtigen Onsetbereiche, während die Segmentierung der Gitarrenmelodie offensichtlich durch nachklingende Vorgängertöne und andere Effekte unsauber erscheint.

Ebenfalls perzeptiv basiert zeigt sich das Verfahren von Moelants und Rampazzo [MR97]. Sie verwenden für die Vorverarbeitung des Audiosignals ein ursprünglich zur Sprachanalyse entwickeltes gehörgerechtes Verfahren [IM92], mit dessen Hilfe 20 Subbänder in Frequenzgruppenbreite [ZF01] erzeugt werden. Autokorrelation der einzelnen Kanäle liefert eine 56 Elemente umfassende Repräsentation pitchrelevanter Signalinhalte. Der eigentliche Segmentierungsschritt schließt sich dann als Suche nach Intensitätsflanken in den Autokorrelationssignalen an. In Fenstern werden gleitend die gefundenen Flanken aufaddiert und bei Überschreiten eines Schwellwertes als Onset gespeichert. Onsets mit einem Abstand kleiner ungefähr 50 ms werden zu einem Einzelevent fusioniert. Die Autoren merken an, dass hier ein komplexerer Ansatz in Abhängigkeit von zusätzlichen Parametern wie Pitch, Klangfarbe und Lautheit angemessen erscheint. Zur Vermeidung von Vibrato-Onsets wird die Korrelation der 56 Einzelbänder mit ihrem zeitlichen Vorgänger berechnet. Ein hoher Wert deutet auf Vibrato und würde eine potentielle Onseteinstufung revidieren.

Validierung des Systems erfolgt anhand der Untersuchung eines mehrere Stücke unfassenden Testdatensatzes von Einzelinstrumenten und kleineren Ensemblebesetzungen in diversen musikalischen Genres. Je nach Komplexität der Darbietung erreichen die Autoren eine Treffergenauigkeit von 60 - 100 %. Probleme ergeben sich insbesondere bei höheren Frequenzen, resultierend aus der ursprünglichen Optimierung des Vorverarbeitungschrittes auf Sprachsignale.

Pragmatischer zeigt sich das Verfahren von Scheirer [Sch98]. Polyphone Inputs werden bezüglich Tempo und Beat (äquidistante Impulse, die das Tempo bestimmen) analysiert. Nachdem das Eingangssignal in potentielle Onsetkandidaten segmentiert wurde, werden diese im Nachverarbeitungsschritt in rhythmische Strukturen zusammengefasst. Für die hier interessante empirisch entwickelte Erkennung von Notenanfängen wird das Analysematerial mittels einer Filterbank in 6 nichtüberlappende, oktavbreite Subbänder im Bereich zwischen 0 - 3200 Hz zerlegt. Je Band wird die Am-

plitudenhüllkurve mit Hilfe der „smooth-and-rectify“-Methode erzeugt, d.h. nach Gleichrichtung erfolgt die Faltung mit einem 200 ms breiten Hamming-Fenster [But98], was einer Tiefpassfilterung entspricht. Ableitung der Hüllkurven (Differenzfunktion 1. Ordnung) und erneuter Gleichrichtung, bezogen auf die positiven Halbwellen, folgt die kanalweise Zuführung einer Resonanzfilterbank. Hierbei handelt es sich um einen Resonator mit sogenanntem „phase-lock“, bei dem die Resonanzfrequenz mit der Periodiziät der vorher berechneten Ableitung übereinstimmt (ähnlich dem Autokorrelationsverfahren). Ausgeführt ist dies als Kammfilterbank: die verwendeten Periodizitätszeiten liegen im Bereich möglicher Tempi. Anschließend werden über alle Kanäle die Filter mit maximalem Energieinhalt zur Summen-Energiefunktion aufaddiert. Über ein übliches Extrempunkt-Suchverfahren bestimmen die gefundenen lokalen Maxima die Positionen der metrischen Schwerpunkte, die dann auch zur Bestimmung des Tempos herangezogen werden können.

Zur Validierung wird ein Datensatz von 60 realen Aufnahmen diverser Genres á 15 s benutzt. Zwei Stufen werden angegeben: im ersten Schritt wird subjektiv im qualitativen Vergleich der mit Klicks versehene Input mit dem Analyseergebnis verglichen. 68 % werden als sehr akkurat und 11 % als mäßig akkurat gewertet. Für den zweiten Test werden die spontanen Beat-Einschätzungen von 5 Hörprobanden sowie die Transkription eines musikalischen Experten mit dem Algorithmus verglichen. Berechnet werden die RMS-Abweichung und die Varianz der Inter-Onset-Intervalle. Hier sind keine Einzelergebnisse angegeben, die Erfolgsquote wird im Bereich der ersten Testreihe angesiedelt.

Den vielleicht komplexesten Ansatz zur Segmentierung stellen Rossignol et al. [RRS+99] vor. Sie unterteilen ihre Implementierung in 3 Abschnitte:

1) „Source segmentation scheme“: In der höchsten Abstraktionsstufe wird das zu untersuchende Signal in Sprache und Musik getrennt. Die Verwendung von Varianz und Mittelwert bezüglich „Spectral Flux“ [SS97], „Spectral Centroid“(Schwerpunkt der Spektralverteilung) und Nulldurchgangsrate (ZCR) ergeben einen 6 Elemente umfassenden Vektor von Klassifizierungsmerkmalen. Sprache beinhaltet neben stationären Abschnitten im Vergleich zu musikalischen Signalen einen großen Anteil geräuschartiger Komponenten (stimmlose Phoneme, etc.). Als Folge dessen ergeben sich somit signifikante Schwankungen der betrachteten Parameter. Charakteristische Unterscheidungsmerkmale sind somit große Varianzen für Sprache und ausgeprägte Mittelwerte für Musik. Als Klassifizierungsmethode erweist sich die Verwendung

von „k-nächste Nachbarn“ bzw. „Multilayer Perceptron“ [Sch96] [DH00] als vorteilhaft.

2) „Vibrato segmentation scheme“: Als Vorverarbeitungsschritt für die dritte Segmentierungsstufe wird in diesem Abschnitt das Eingangssignal von Vibrato, d.h. zeitlichen Schwankungen der Frequenzwerte, befreit, um „Pseudo-Onsets“zu vermeiden. Vereinfacht ausgedrückt werden dabei lokale Extrema der Frequenztrajektorie gesucht und auf ihre Korrelation bezüglich typischer Vibrato-Periodizitäten (150 - 250 ms) hin untersucht. Bei positivem Ergebnis ergibt sich der geglättete Frequenzverlauf als Mittelwert der lokalen Minima und Maxima.

3) „Note & phones segmentation scheme“: Abschließend findet dann die Unterteilung in stationäre und transiente Signalabschnitte statt, wie sie für eine Detektion von Notenanfängen notwendig ist. Eine breite Palette von unterschiedlichen Merkmalen wird mit einem Takt von 100 Hz (10 ms) berechnet:

1. Ableitung der Pitchtrajektorie ($d(i) = |f(i+1) - f(i-1)|$),
2. relative Ableitung der Pitchtrajektorie ($\delta(i) = d(i)/f_0(i)$),
3. Ableitung der Energie (berechnet in 20 ms-Fenster),
4. relative Ableitung der Energie,
5. Inharmonizität der Partialtöne ($H(i) = \sum_{n=2}^{N}(|f_n - n \cdot f_0|/n \cdot f_0)$),
6. „Voicing Coefficient“ (Korrelation von Partialton- und F_0-Energien),
7. „Spectral Flux“ [SS97],
8. Wahrscheinlichkeitsmaß für F_0-Werte in Fenstern (50ms) vor und nach aktueller Zeit,
9. Entropiemodell der Pitchtrajektorie [BN93a].

Zur Unterscheidung von bedeutungsvollen und rauschartigen Maxima wird zur Festlegung der Schwellwerte unter Annahme einer Normalverteilung die 3σ-Regel angewandt, d.h. σ wird unter Einbeziehung von 90 % der niedrigen Werte ermittelt. Als Entscheidungsfunktion für oder gegen einen Onset wird schließlich aus der Summe der Einzelfeatureentscheidungen (0/1) in 50 ms-Fenstern die Dichte der Pulse herangezogen.

Einen aussergewöhnlichen Ansatz stellt Raphael [Rap99] mit der Umsetzung von Hidden-Markov-Modellen [Rab77a] vor. Hierbei handelt es sich um einen typischen „Top-Down"-Ansatz, d.h. Vorwissen über das zu untersuchende System erleichtert die Analyse, limitiert aber auch die Allgemeingültigkeit bezüglich alternativer Anwendungen. Ziel ist die Implementierung eines Systems zur automatisierten musikalischen Begleitung einer Solostimme. Mit Hilfe der originalen Notenschrift soll das zugehörige Audiosignal in Noten und Pausen segmentiert werden. Der Algorithmus weist zunächst *a priori* unterschiedlichen Segmentierungen bestimmte Wahrscheinlichkeiten zu. Daraus wird mittels des verwendeten Hidden-Markov-Ansatzes ein Modell entwickelt, das die Ähnlichkeit der akustischen Daten mit einer hypothetischen Segmentierung beschreibt; Ermittlung der Modellparameter erfolgt über nicht-überwachtes Lernen. Mittels dynamischer Programmierung [Gus97][Wat95] wird schließlich die global optimale Segmentierung als Minimierung der *posterior* erwarteten Anzahl von Segmentierungsfehlern identifiziert. Validiert wird das System mit einem einfachen monophonen Input von ein paar Takten Länge. Eine Solo-Kadenz aus einem Oboenkonzert von Mozart dient als Testexempel, welches einen anspruchsvollen Input mit großen Temposchwankungen und anderen Variabilitäten darstellt. Das Segmentierungsbild ist wegen der Kadenzeigenheiten schwer objektiv bewertbar. Nach Aussage des Autors entspricht das Segmentierungsergebnis aber weitgehend dem Original.

Den wohl psychoakustisch konsequentesten Ansatz zur Detektion von Onsetzeitpunkt und -intensität stellt Klapuri [Kla99] vor. Zunächst transformiert er das Eingangssignal mittels einer gehörgerechten Filterbank von 21 nichtüberlappenden Bändern mit Frequenzgruppenbreite in die Zeit-Frequenzdarstellung. Anschließend wird versucht, die Energieintegration des menschlichen Hörapparates nachzubilden, um plötzliche Wechsel zu erhalten, aber graduelle Variationen zu verdecken. Dazu wird durch Faltung jedes Bandsignals mit einem 100 ms breiten Hamming-Fenster [But98] die Amplitudenhüllkurve berechnet (vgl. Verfahren von Scheirer [Sch98]).
Die weitere Verarbeitung stützt sich auf die Angaben von Moore [Moo95], wonach die kleinste wahrnehmbare Änderung in der Intensität (JND = „just noticable difference") näherungsweise proportional ist zur Intensität des Signals, d.h. der wahrgenommene Anstieg der Intensität steht in direkter Beziehung zum zugehörigen Ausgangspegel. Formelmäßig zusammenfassen lässt

sich dies in der Konstanz des sogenannten Weber-Bruch:

$$\frac{\Delta I}{I} = const. \tag{2.22}$$

Die Beziehung gilt für Intensitäten I im Bereich zwischen 20 dB bis ungefähr 100 dB über der absoluten Hörschwelle. Es lässt sich zeigen, dass dies äquivalent ist zur zeitlichen Ableitung des Logarithmus des Signalverlaufs $A(t)$:

$$W(t) = \frac{d}{dt}(logA(t)). \tag{2.23}$$

In der Funktion W(t) wird nun nach signifikanten lokalen Maxima gesucht, die einen globalen Schwellwert überschreiten. Gefundene Einzelband-Onsets mit einem Abstand von weniger als 50 ms werden fusioniert. Zur Bestimmung von gültigen Notenanfängen werden die lokalen Kanal-Onsets gleitend in 50 ms-Fenstern aufaddiert und erneut einer Schwellwertprüfung unterzogen.

Zur Validierung finden sich 10 s-Samples diverser Genres. Mittels eines Vergleichs mit einer manuellen Transkription ergeben sich je nach Komplexität des Signals Erkennungsraten von 7 % - 95 %.

Goto [Got01a] stützt seine Suche nach der rhythmischen Struktur in musikalischen Audiosignalen auf Onset-Zeiten, Harmoniewechsel und Schlagzeugmuster. Hieraus gewinnt er als Ergebnis höherwertiger Interpretationsstufen Aussagen über die Position von Taktwechseln, Viertel- und halben Noten unter der Annahme, dass die Klangbeispiele im 4/4-Takt vorliegen und die verwendeten Tempi zwischen 61 und maximal 120 Schlägen pro Minute für schlagzeuglose Musik bzw. maximal 185 Schlägen für schlagzeugbesetzte Beispiele liegen. Für die hier interessante Bestimmung der Onsetzeitpunkte wird das Signal im Bereich von 0 - 11 kHz in 7 Frequenzbänder mit annähernd Oktavbreite zerlegt. Im ermittelten Spektralverlauf wird dann nach Leistungsflanken in benachbarten Zeit-Frequenz-Regionen mittels der sogenannten „Degree-of-Onset-Function“gesucht:

$$D(t) = \sum_f d(t, f), \tag{2.24}$$

$$d(t,f) = \begin{cases} max(p(t,f), p(t+\Delta t, f)) - PrevPow \\ \quad falls \quad min(p(t,f), p(t+\Delta t, f) > PrevPow \\ 0 \quad sonst \end{cases}, \tag{2.25}$$

$$PrevPow = max(p(t - \Delta t, f), p(t - \Delta t, f \pm 1)), \tag{2.26}$$

(p(t,f) = Leistung im Frequenzband f zur Zeit t).

Anschließend an die Glättung der Funktion $D(t)$ wird dort nach signifikanten Peakwerten, die einen empirisch ermittelten Schwellwert überschreiten, gesucht. Die Maxima in den einzelnen Frequenzbändern müssen also durch ihre Umgebung verifiziert sein, um eine gültige Onsethypothese aufzustellen.

Die Anwendbarkeit der Methode weist Goto anhand eines Testdatensatzes von 85 CD-Aufnahmen mit einer Länge von mindestens 1 Minute und konstantem Tempo nach. Die Signale teilen sich in 45 schlagzeugbesetzte und 40 schlagzeuglose Inputs. Die Erkennungsrate der korrekten rhythmischen Einschätzung liegt für alle drei Ebenen (Takt, Viertel- und halbe Noten) bei über 86 %.

Im Rahmen seiner Untersuchungen zur Expressivität von Darbietungen von Pianomusik beschäftigt sich Dixon [Dix01] mit der Onsetdetektion polyphoner Klaviermusk. Trotz der zeitlich wohldefinierten Anschlagcharakteristik der untersuchten Signale wird die Aufgabe aber durch die vorhandene Mehrstimmigkeit nichttrivial. Er betrachtet die Problematik als Klassifikationsaufgabe, d.h. eine Reihe zeitlicher und spektraler Parameter werden in Feature-Vektoren zusammengefasst, trainiert und klassifiziert. Wie bei den meisten Verfahren wird auch hier nach starken Anstiegen im Zeitsignal und korrespondierenden Frequenzbändern gesucht. Im Wesentlichen werden dabei die schon beschriebenen Ideen von Schloss [Sch85] zu den Hüllkurvensteigungen im Zeitsignal auf eine Anzahl von Frequenzbändern adaptiert. Ein nicht näher erläuterter „genetischer“ Algorithmus dient dem Mustererkennungsprozess. Als Testdatensatz werden 10 Mozart-Klaviersonaten, dargeboten auf einem elektronischen Klavier, benutzt. Vorteilhaft ist die Überprüfbarkeit der Ergebnisse durch die Originalmitschrift des Instruments als MIDI-Notation. Trainiert wird mit jeweils einer Sonate, während der Klassifizierungsvorgang auf die Gesamtheit aller 10 Werke angewendet wird. Dixon gibt an, 90 % aller Notenanfänge richtig erkannt zu haben mit einer Genauigkeit von $\Delta t < 10$ ms.

Jensen und Murphy [JM01] begegnen dem allgemeinen Problem, die optimalen Schwellwerte für eine korrekte Segmentierung zu bestimmen, durch den Einsatz eines neuronalen Netzwerkes. Dazu bereiten Sie die Gesamtheit

der benutzten Daten mittels manueller Analyse und Segmentierung für den Trainingsprozess vor. Im Abstand von 10 ms werden folgende 12 Features berechnet: „Brightness“, „modifizierte Tristimuli“ T1, T2 und T3 [PJ82], Amplitude des Zeitsignals, Grundfrequenzanalyse über Autokorrelation, „High Frequency Content“ und HFC-Detektierungsfunktion [MB96], „Spectral Flux“ [SS97], Zero-Crossing-Rate (ZCR) sowie die Ableitung der Grundfrequenz. Diese Parameter werden als „Mikro-Level-Parameter“ in ein neuronales Elman-Netzwerk [DB00] gegeben und über Standard-Backpropagation trainiert.

Im Rahmen des eigentlichen Segmentierungsprozesses wird zusätzlich zu den gewonnenen Gewichten als weiterer „Makro-Level-Parameter“ ein Rhythmusmodell eingeführt. Hierbei werden im Verlauf der Analyse gefundene Inter-Onset-Zeitintervalle zum Aufbau einer Statistik wahrscheinlicher Onset-Abstände verwendet und diese Werte als unterstützende Segmentierungselemente genutzt.

Die Evaluierung des Ansatzes erfolgt mit einer monophonen Melodiedatenbank bestehend aus 7 Melodien (6 Instrumente und Gesang) mit einer durchschnittlichen Länge von ca. 60 s. Die Tests unterteilen sich in zwei Methoden. Im ersten Durchlauf werden Fragmente der Melodien segmentiert, während ein weiterer Test die „Leave-One-Out“-Methode (LOO) verwendet, d.h. bis auf eine Melodie werden alle übrigen zum Trainieren verwendet und auf den ausgelassenen Input getestet. Als noch nicht zufriedenstellendes Resultat geben die Autoren einen verbleibenden Fehler von ca. 20 % an.

Angelehnt an die „Phase Vocoder“-Theorie [AKZ02] segmentiert das Verfahren von Duxburry et al. [DDS01]. Hierbei wird versucht, aus Phaseninformationen der Frequenzinhalte das Signal in stationäre und transiente Abschnitte zu zerlegen. Die Phasendifferenz $\Delta\phi(k,t) = \phi(k,t) - \phi(k,t-\Delta t)$ eines Frequenz-Bins zwischen zwei aufeinanderfolgenden Analyseframes bestimmt die Augenblicksfrequenz f_i:

$$f_i = \frac{\Delta\phi(k,t)}{2\pi\Delta t}. \tag{2.27}$$

Die Schwankung der Augenblicksfrequenz, also die Abweichung von der idealen Bin-Frequenz f_k, dient als Maß zur Detektion von nichtstationären Signalverläufen. Um die Fensterbreiten der Kurzzeit-Fourier-Transformationen auf die untersuchten Frequenzbereiche anzupassen, wird, ähnlich der Vorgehensweise bei der Wavelet-Transformation, die Frequenztransformation als „Konstant Q“-Filterbank durchgeführt („Multiresolution analysis“). Der

Schwellwert T_t zur Detektion von Transienten wird eingeführt als:

$$\phi(k,t) - 2 \cdot \phi(k,t-\Delta t) + \phi(k,t-2\cdot\Delta t) > T_t. \tag{2.28}$$

Zur verbesserten Anpassung an die zeitliche Signalentwicklung wird die Detektionsschranke variabel gehalten, d.h. in Abhängigkeit von den vorhergehenden Analyseschritten passt sich der adaptive Schwellwert A_t gemäß

$$A_t = T_t + \alpha T_t + \beta T_t \tag{2.29}$$

an die Vorgaben der vergangenen Frames an. Die Parameter α und β sind empirisch ermittelte reelle Zahlenwerte.

In der Anwendung wird das Verfahren verglichen mit der vorher beschriebenen HFC-Methode. Im subjektiven Vergleich der Segmentierungsergebnisse eines Melodieverlaufes mit 11 Noten zeigt das Verfahren das eindeutigere und akkuratere Resultat.

2.3 Monophone Melodietranskription

Die in den Abschnitten 2.1 und 2.2 vorgestellten Einzeldisziplinen „Pitchextraktion“ und „Segmentierung“ stellen in ihrer Zusammenführung die Grundpfeiler monophoner Transkriptionssysteme dar. Einstimmige Melodieverläufe werden aus ihrer hierarchisch niederstufigen Repräsentation als Audiosignale überführt in symbolische, abstraktere Darstellungen. Diese können als einfache Contour-Linien (aufwärts, abwärts, gleichbleibend), aber auch als grob quantisierte Intervallstufen sowie gebräuchliche musikalische Zwölfton-Notation vorliegen.

In der Regel wird zunächst der zeitliche Verlauf der Grundfrequenzen der harmonischen Schallsignale bestimmt und anschließend in einzelnen Segmenten (Noten) zusammengefasst. Den resultierenden Abschnitten kann dann gemäß deren Frequenzinhalten eine bestimmte Notenhöhe zugeordnet werden.

2.3.1 Transkriptionssysteme

Das erste von drei repräsentativen Melodietranskriptionssystemen findet sich in der Arbeit von Haus und Pollastri [HP00][HP01]. Es ist konzipiert für die Transkription von Gesang. Der Algorithmus beginnt im Gegensatz zu

den meisten anderen Systemen mit der Segmentierung. Potentielle Segmentgrenzen werden mit Hilfe des Signal-Rausch-Abstandes (SNR) detektiert. Dazu wird der Rauschpegel mit 15% über dem Effektivwert (RMS) der ersten 60 ms des Audiosignals festgelegt. Eine verbesserte Auflösung von Notenanfängen und -enden wird erreicht durch die Unterscheidung von stimmhaften und nichtstimmhaften Bereichen, also durch die Suche nach signifikanten pitchbehafteten Vokalen sowie nichtharmonischen Abschnitten. Hierbei sind somit Segmentierung und Pitchextraktion unmittelbar miteinander verknüpft. Der Pitch eines stimmhaften Bereiches ergibt sich durch Kurzzeitfouriertransformation (STFT) und Interpretation des erhaltenen Spektrums. Der Bereich möglicher Grundtonfrequenzen wird auf ca. 80 - 800 Hz limitiert. Nach der blockweisen Bestimmung der Pitchfrequenzen werden diese in drei aufeinanderfolgenden Einheiten mediangefiltert und auf Oktavfehler getestet. Vier zusammenhängende Blöcke mit identischer Grundfrequenz werden in einer Einheit zusammengefasst. Unterscheiden sich solche Einheiten um mehr als 0,8 Halbtöne, wird ein Notenwechsel prognostiziert und eine zusätzliche Segmentgrenze hinzugefügt. In einem Nachbearbeitungsschritt wird versucht, die Abweichung des Sängers von der Standardstimmung (440 Hz) zu berücksichtigen. Hierzu wird angenommen, dass der Sänger die Tonhöhe seiner dargebotenen Melodie implizit auf einen Referenzton bezieht. Sämtliche zugeordnete Notenfrequenzen werden verglichen mit einer Anzahl möglicher Stimmungen. Die Überlappungen zwischen gefundenen Noten und vorgegebenen Skaleneinträgen werden in einem Histogramm aufsummiert, dessen Einträge 0,2 Halbtöne breit sind. Die Inhalte des maximalen Peaks werden gemittelt und das Ergebnis der Intonation des Interpreten zugeordnet.

Validierung erfolgt über jeweils 4 kurze Melodien von 5 Sängern. Demzufolge werden 90 % der Noten richtig erkannt. Weitere erläuterte Testergebnisse dieser Implementierung finden sich neben denen anderer frei zugänglicher Transkriptionssysteme in Kapitel 6.1.3.

Pragmatisch und konsequent ist die Implementierung von Monti und Sandler [MS00]. Die Grundfrequenzbestimmung erfolgt mittels Autokorrelation des Zeitsignals. Die Autoren verweisen explizit auf die Motivation des gewählten Ansatzes durch die Arbeiten von Brown [BP89][BZ91] (s. Kapitel 2.1). Wie üblich wird nach Maxima der Autokorrelationsfunktion in Abhängigkeit des Verschiebungsparameters gesucht, die Hinweise auf die Fundamentalperiode und Subharmonische geben. Das Verfahren erweist sich

somit auch gut geeignet für Signale mit nur schwach ausgebildetem Grundton. Der berücksichtigte Frequenzbereich erstreckt sich von 120 - 2700 Hz. Aus Effizienzgründen ist die Autokorrelation als schnelle Fouriertransformation implementiert (Zusammenhang über „Energiedichtespektrum“). Die Segmentierung wird als rein pitchbasiert angegeben, was robusteres Verhalten bei Notenwechseln im Glissando oder Legato bereitstellen soll. Im sogenannten „Kollektor“ wird nach konstanten Pitchverläufen gesucht, die einzelne Noten darstellen. Bei signifikanten Frequenzwechseln wird eine Segmentgrenze (Notenintervall) angenommen. Implizit wird durch die Berechnung der Hüllkurve zur Detektion von Pausen, in denen die Hüllkurvenamplitude einen Schwellwert nicht überschreitet, auch eine Art der „konventionellen“ energiebasierten Segmentierung benutzt.

Als Resultate werden nur subjektive Betrachtungen kurzer Melodien von Blechblasinstrumenten angegeben. Dies erfolgt durch Vergleich des resynthetisierten „Csound“-Outputs [cso03] mit dem Original und wird als vielversprechend bewertet.

Clarisse, et al. [Lem02] beschreiben ein System, dass sich hinter dem in dieser Arbeit beschriebenen Verfahren am konsequentesten an den physiologischen Gegebenheiten der menschlichen auditorischen Peripherie orientiert. Es wird versucht, eine angemessene Beschreibung der im Innenohr vorliegenden Signalrepräsentation nachzuvollziehen. Das auditorische Modell beginnt mit der Simulierung der Transferfunktion von Außen- und Mittelohr als Tiefpassfilterung des Audiosignals. Anschließend wird die hydromechanische Verarbeitung der Cochlea (s. Kap. 3) als 23-kanalige Filterbank mit Mittenfrequenzen im Frequenzgruppenabstand (140 Hz - 6 kHz) unter Berücksichtigung von Mitverdeckungseffekten [ZF01] umgesetzt. Die im Innenohr vorhandenen Haarzellen sind verantwortlich für den Transduktionsvorgang von mechanischer Schwingung in neuronale Nervenimpulse. Diese werden für jeden Filterkanal nachgebildet als AGC-Verstärker inklusive Halbwellengleichrichtung und Dynamikkompression zur Nachbildung der auftretenden Effekte.

Die Schritte zur Extraktion der Pitchverläufe beginnen mit der Berechnung der Autokorrelation pro Filterkanal. Die erhaltenen Ergebnisse werden als Summenautokorrelation über alle Teilbänder aufaddiert. Mögliche Pitchkandidaten und zugehörige Wertigkeiten erhält man aus den Maxima der Summenautokorrelationsfunktion (SACF). Eine Kontinuitätsbetrachtung der Frequenzhypothesen liefert mögliche Trajektorien als Repräsentanten der

Melodienoten. Die Suche nach nichtstimmhaften Anteilen im Audiosignal wird benutzt als Segmentierungsalgorithmus. Parameter hierfür sind niedrige Wertigkeiten der Maxima aus den Korrelationsfunktionen sowie Diskontinuitäten im Pitchverlauf. Testresultate sind in Kapitel 6.1.3 angegeben.

2.3.2 Query-By-Humming

Eine Reihe sogenannter „Query-By-Humming"-Systeme (QbH) (s.a. Kapitel 1 und 6.1) verwenden als akustisches „Frontend" monophone Melodietranskriptionseinheiten. Semantische Inhalte in Form von Notenverläufen werden benötigt, um mittels dieser symbolischen Darstellung die gesuchten Melodien in Datenbanken zu suchen. Die meisten Systeme legen ihr Hauptaugenmerk auf die effizienten Mustererkennungsverfahren der extrahierten Melodien bezüglich vorgegebener Datenbanken. Eine exemplarische Darstellung eines solchen Suchverfahrens findet sich in Kapitel 6.1. Die nachfolgende Beschreibung repräsentativer Vertreter solcher QbH-Systeme beschränkt sich auf die Zusammenfassung der für diese Arbeit relevanten Transkriptionseinheiten. Soweit nicht anders angegeben, schränken diese Systeme die zu analysierenden Signale weitgehend ein. Bei gesungenen Melodien sollen in der Regel die Noten auf Silben beginnend mit einem Stopkonsonanten, gefolgt von einem deutlich ausgehaltenen Vokal, artikuliert werden. Häufig werden zusätzlich deutlich voneinander abgesetzte Noten gefordert. Nach Ansicht des Autors verdeutlicht dies, dass sich die Qualität von Transkriptionssystemen hauptsächlich in der zuverlässigen Segmentierungsstufe widerspiegelt. Eine robuste Pitchextraktion wird von den meisten Implementierungen bereitgestellt. Eigene Experimente deuten allerdings daraufhin, dass nicht-textuell dargebotene Melodien bei musikalischen Laien die gesungenen Intervallfehler reduzieren.

Der Urvater der QbH-Systeme findet sich wohl in der Arbeit von Ghias et al. [GLCS95]. Nach Vergleich mit „Cepstrum-Analyse" und „Maximum-Likelihood" entscheiden sich die Autoren zur Pitchextraktion für das Autokorrelationsverfahren (s. Kapitel 2.1). Über eine nicht näher erläuterte pitchbedingte Segmentierung werden Sequenzen unterschiedlicher Grundfrequenzen transkribiert (die gesungene Eingabe soll auf voneinander abgesetzten Silben „haaa" erfolgen). Das Ergebnis zeigt eine „Contour"-Linie, d.h. der globale Verlauf wird über die Parameter „D" (abwärts), „U" (aufwärts) und „S" (gleichbleibend) dargestellt.

Evaluiert wird das System über die Robustheit der Suche in einer Datenbank von 183 Liedern (in MIDI-Repräsentation) verschiedener Genres. Bezüglich der erhaltenen Resultate wird nur angemerkt, dass 10 - 12 Noten notwendig sind, um 90 % der Daten auseinanderzuhalten.

McNab et al. [MSW96][MSWH00] beschreiben das ursprünglich unter dem Namen „Meldex“ bekannte Transkriptionssystem, das seit dem Jahr 2000 im neuseeländischen „Greenstone Digital Library Software“-Projekt angesiedelt ist. Die Pitchberechnung erfolgt über eine Art Autokorrelationsverfahren. Mittels des „Gold-Rabiner-Algorithmus“ [GR69] werden im Zeitbereich Extremalwerte des Schallsignals ausgewertet. Die Segmentierung basiert ausschließlich auf dem Effektivwert (RMS) des Zeitsignals. Sobald der RMS-Wert einen voreingestellten Schwellwert überschreitet, wird ein Notenbeginn angenommen (analog bei Unterschreitung eines etwas niedrigeren Schwellwertes das Notenende). Der Schwellwert wird bei 55 % für den Onset bzw. 35 % für den Offset bezogen auf das Mittel des Effektivwertes über das komplette Audiosignal angenommen. Die Zuweisung einer Note für einen segmentierten Bereich erfolgt über die Identifizierung des höchsten Peaks im Histogramm der Fundamentalfrequenzen des aktuellen Segments; anschließend wird die Pitchfrequenz ermittelt als Durchschnitt der Einträge im Maximum-Bin des Histogramms. Eine Zuordnung von Midinoten aus den ermittelten Segmentfrequenzen wird dynamisch aufgebaut, d.h. es werden sukzessive die Intervalle zwischen zwei aufeinderfolgenden Noten berechnet und fortschreitend nach jeder neuen Note die Intonation aktualisiert. Die Autoren geben selbst keine Validierungsergebnisse ihres Systems an, schränken aber für eine zuverlässige Segmentierung eine gesungene Eingabe mittels scharf akzentuierter Silben wie „da“ oder „ta“ ein. Vergleichende Testergebnisse finden sich in Kapitel 6.1.3.

Chai [Cha01] entscheidet sich bezüglich der Pitchextraktion ebenfalls für das Autokorrelationsverfahren, gibt aber zu Bedenken, dass wegen der stark ausgeprägten Formantstruktur der menschlichen Stimme eine gewisse Anfälligkeit gegenüber harmonischen Vertauschungen besteht. Da die Suche in der Datenbank großen Wert auf rhythmische Aspekte legt, zeigt sich der Segmentierungsschritt aufwändiger als bei anderen Systemen. Der Anwender wird aber auch hier zur sauber abgesetzten Gesangseingabe aufgefordert, diesmal über die Silbe „Da“. Aufgrund des Stopkonsonanten, der

den anregenden Luftstrom im Vokaltrakt unterbricht, wird eine Segmentierung über die Signalamplitude möglich. Das in Zeitfenstern berechnete Betragsspektrum wird im Hauptbereich der menschlichen Stimme (im System $\leq$ 1000 Hz) integriert und dient als modifizierte Amplitudenhüllkurve. Die Einführung eines dynamischen Schwellwertes reagiert auf eine veränderliche Gesamtamplitude. Dies ist aber offensichtlich nicht robust genug bezüglich Amplitudenänderungen innerhalb von Noten und resultiert in Mehrfachsegmentierungen.

Experimente zur Evaluierung beinhalten jeweils 5 Melodien von 10 Probanden, die in einer Datenbank mit 8000 Liedern gesucht werden. Die Auswertung erfolgt personengebunden und zeigt stark gefächerte Erkennungsraten zwischen 17 % und 100 % je nach Proband. Des Weiteren untersucht die Autorin die Qualität der Segmentierung durch subjektiven Vergleich von gesungenem Original und Transkription. Sie gibt den Fehler für ausgelassene Noten mit 3,5 %, den für zusätzliche Noten mit 0,9 % an.

Ein sprachgesteuertes QbH-System stellt Pauws [Pau02] vor. Hierdurch wird die Natürlichkeit der Eingabeschnittstelle bezüglich einer intuitiven Bedienung deutlich verbessert. Die Überführung des Audiosignals in die semantische Beschreibung in Form von Noten erfolgt in der üblichen Vorgehensweise. Zunächst wird versucht, den Noten-Pitch als „Summe“ harmonischer Partialtöne über „subharmonische Summation“ [Her88] zu bestimmen. Dazu werden nach Frequenztransformation die Teiltöne der potentiellen Obertonreihen gemäß einfacher auditorischer Überlegungen (bervorzuge niedrige Teiltonnummern bezüglich der Pitchperzeption) gewichtet aufsummiert. Die maximale Wertigkeit bestimmt die Pitchhypothese. Der anschließende Segmentierungsschritt enthält eine Reihe bekannter Einzelverfahren, dürfte aber aufgrund der Kombination unterschiedlicher Methoden recht effektiv sein. Neben der Berechung der Kurzzeit-Energie zur Detektion von Signalpausen zwischen einzelnen abgesetzten Noten findet sich ein weiterer als „Surf“-Algorithmus bezeichneter Schritt. Das hochpassgefilterte Signal wird mittels Polynomfit in eine geglättete Hüllkurve überführt, deren starke Flanken Indizien für Onsets darstellen. Weiterhin wird die von Masri und Bates eingeführte Methode der „High Frequency Content Function“ (HFCF) (s. Kap. 2.2) sowie der Verlauf der Pitchtrajektorien als Pitchsegmentierung verwendet. Die abschließende Quantisierung der Notenhöhe erfolgt als Median der Grundfrequenzen aus dem jeweiligen Notensegment mit anschließender Zuordnung auf das Stimmungsraster der ersten erkannten Note.

Es werden keine Evaluierungsergebnisse angeführt, allerdings lassen sich Hinweise auf Probleme bei hohen Frequenzen und textuellem Input erkennen. Die gesungene Melodie soll auf Silben mit nichtstimmhaftem Frikativlaut zu Beginn und anschließendem langen Vokal erfolgen.

Abschließend seien noch drei weitere Verfahren erwähnt, die aber ähnliche Ansätze verfolgen wie die detaillierter beschriebenen Implementierungen.

Carré [Car02] nutzt das ursprünglich für die Sprachverarbeitung entwickelte „enhanced Super Resolution Pitch Determinator"-Verfahren (eSRPD) [Bag94] zur Pitchbestimmung. Die Segmentierung erfolgt über den zeitlichen Verlauf der Stimmhaftigkeit als Ausprägung harmonischer Anteile im Signal.

Rao und Raju [RR02] stellen ein anwendungsbezogenes Qbh-System für Hindi-Filmsongs vor. Die benutzten Ansätze sind konventionell: Autokorrelation und minimale Pitchsegmentierung münden in einer Contour-Darstellung der Melodie. Von der Idee fast identisch zeigt sich das System von Lu et al. [LYZ]. Einziger wesentlicher Unterschied besteht offensichtlich in der Benutzung der Nulldurchgangsrate zur Grundfrequenzbestimmung.

2.4 Polyphone Ansätze

Während im Bereich der monophonen Transkription verschiedene Anwendungen ein Niveau erreicht haben, dass Praxistauglichkeit erkennen lässt, gibt es bei der Analyse polyphoner Musik bis dato noch kein Verfahren, dass auch nur annähernd einigermaßen allgemeingültige und zuverlässige Ergebnisse liefert. In der Regel beschreiben die bekannten Arbeiten bestimmte Spezialfälle, auf die dann die Systemparameter angepasst werden. Bis heute ist also die Fähigkeit der menschlichen musikalischen Szenenanalyse von solchen künstlichen Ansätzen weitestgehend unerreicht.

Die Arbeit von Baumann [Bau95] stellt das Verfahren dar, welches am konsequentesten versucht, die Strategien höherer kognitiver Strukturen auf die polyphone Analyse nachzuvollziehen. Die aus der Psychologie bekannten „Gestaltgesetze" [And01] werden auf die vorverarbeiteten Audiosignale angewandt. Dazu gehören „gemeinsames Schicksal", Kontinuität, Nähe, „gute Fortsetzung" und Ähnlichkeit zeitlich und spektral benachbarter Elemente. Grundlegender Ansatz ist das Auffinden akustischer Objekte, d.h. die Grup-

pierung und Zuordnung von Spektralkomponenten zu einer der menschlichen Wahrnehmung entsprechenden Einheit.

Im Vorverarbeitungsschritt wird das Signal einer gehörgerechten Spektralanalyse unterzogen und anschließend das Kurzzeitspektrum nach lokalen Peaks abgesucht. Diese dienen über Pegelüberschuss- und Spektraltonhöhengewichtbetrachtungen [Ter79][TSS82] als Maß für perzeptiv relevante Partialtöne. Hieraus gewonnene „Teiltonlinien“ werden in nachfolgenden Verarbeitungsstufen zu hierarchisch aufsteigenden Repräsentationen zusammengefasst. Zunächst werden Spektralkomponenten mit synchronen Onset- und Offsetzeitpunkten in Objekte gruppiert („Einsatzintegration“), denen dann eine oder mehrere potentielle Tonhöhen zugewiesen werden („subharmonische Koinzidenz“ [TSS82]). „Homophone Trennung“ und „Kollisionserkennung“ spalten die Cluster der Teilfrequenzen wieder in mögliche gleichzeitige Einzeltöne auf. Im Gegensatz zu Bregman [Bre90], der das Prinzip der „exklusiven Allokation“ favorisiert, werden hier also Mehrfachinterpretationen zugelassen. Abgeschlossen wird das Verfahren durch die Anwendung der „sequentiellen Integration“, mittels derer die erkannten Einzelobjekte zu auditorischen Strömen (Melodien) fusioniert werden.

Die Bewertung der Implementierungsergebnisse stützt sich auf subjektive Urteile des Autors bezüglich eines kurzen zweistimmigen Klavierstückes sowie eines ausgehaltenen Trompetentons mit kurzen Klarinetten- und Tubaeinwürfen. Resynthetisierte Analyseergebnisse liefern eine ordentliche Erkennung der originalen Inhalte.

Die Verwendung von sogenannten „Blackboard“-Systemen zur Analyse komplexer musikalischer Audiosignale wird eingeführt von Martin. Die Herangehensweise leitet sich ab von der metaphorischen Beschreibung einer Expertengruppe, die sich problemlösend um eine Tafel gruppiert. Ein solches System besteht aus einem zentralen Datenraum, dem „Blackboard“, in dem in unterschiedlichen Abstraktionsebenen verschiedene hierarchische Stufen des Schallsignals und der zugehörigen semantischen Inhalte verwaltet werden. Zusätzlich dazu existieren eine Reihe von sogenannten „Wissensquellen“ (Knowledge Sources KS), die Vorwissen über den zu analysierenden Kontext in die Interpretation der Daten mit einbringen. Es finden sich hier somit Beschreibungen von auditorischer Physiologie, den physikalischen Schallerzeugungsmechanismen sowie musiktheoretischer Gesetzmäßigkeiten. Steuerung von Datenaustausch und Prozessabläufen wird durch den „Scheduler“ organisiert.

Im ersten technischen Report [Mar96b] wird eine Implementierung zur Transkription vierstimmiger Klaviermusik vorgestellt. Die ausschließliche Verwendung Bachscher Choralmusik schränkt das Aufgabengebiet weitestgehend ein, besteht diese doch aus strengen kontrapunktischen Regeln mit moderatem Tempo, synchronen Notenanfängen und klar abgegrenzten harmonischen Strukturen. Als Vorverarbeitungsstufe wird das Audiosignal mittels Kurzzeit-Fouriertransformation (STFT) in die Frequenz-Zeit-Darstellung überführt. Berechnung der Kurzzeit-Energie des Zeitsignals (Quadrierung und Tiefpassfilterung) bestimmmt in Form von starken Flanken die Segmentgrenzen der Akkorde. Innerhalb dieser Grenzen werden die spektralen Inhalte gemittelt und dienen über die Parameter Startzeit, Frequenzen und Amplituden als Eingangsdaten für das „Blackboard"-System.

Die Datenhierarchie teilt sich mit zunehmend abstrakterem Inhalt auf in die Bereiche Rohdaten (Eingangsparameter), Partialtöne, Noten, Intervalle und Akkorde. Die Bestimmung der Noten erfolgt über einfache gewichtete Evidenz-Kumulation der Präsenz der ersten 5 Partialtöne nach Davis et al. [DBS77]. Als gleichzeitg auftretende Intervalle werden nur kleine und große Terzen akzeptiert. Die Grenzen dieser Implementierung, fehlende Oktavunterscheidung sowie kleiner Notenbereich von ca. $120 - 440$ Hz, werden nach Ansicht des Autors hauptsächlich verursacht durch die nicht adäquate Vorverarbeitungsstufe der Frequenzanalyse.

Aufgrund dessen wird im zweiten technischen Report [Mar96a] ein modifiziertes System vorgestellt. Als auditorisches Frontend wird das von Ellis entwickelte „Log-Lag Korrelogramm" [Ell96] benutzt. Mittels einer Filterbank, bestehend aus 40 Gammatonfiltern mit den üblichen Patterson-Holdworth-Parametern, wird eine Auflösung von 6 Filtern pro Oktave erreicht. Anschließend wird für jeden Filterkanal die Autokorrelation zwischen 20 und 1000 Hz mit 48 Verschiebungswerten pro Oktave berechnet. Im „Periodogramm" wird die Summenautokorrelation als normalisierter Durchschnittswert der Bänder zum jeweiligen Verschiebungsparameter ermittelt. Die Hierarchie des zentralen Datenraums wird im Vergleich zur ersten Implementierung auf niederwertige Abstraktionsebenen beschränkt. Es finden sich 7 Hierarchiestufen: Korrelogramm, Periodogramm, Peaks (lokale Maxima im Periodogramm), Periodizitäten (zeitlich kontinuierliche Peaks), Bandhüllkurven, Onsets (lokale Maxima der Hüllkurven) sowie einzelne Noten. Die Wissensdatenbank beschränkt sich nun auf rudimentäre beschreibende Merkmale der verwendeten Hierarchiestufen, die keine höherwertigen „Weltinformationen" beinhalten und wird damit eigentlich ad absurdum geführt.

Die Behauptung des Autors, dass die korrelationsgestützte Analyse für die gewählte Problemstellung bessere Ergebnisse bereitstellt, lässt sich aus den beschriebenen Resultaten nicht nachvollziehen. Weder die Analyse eines Piano-Oktavintervalls noch einer kurzen monophonen Phrase einer Bach-Fuge liefern eindeutige Ergebnisse. Die Eignung bezüglich polyphoner Transkription soll anhand eines kurzen 4-stimmigen Bach-Chorals veranschaulicht werden. Die Systemparameter sind so justiert, dass die ersten 5 Harmonien korrekt transkribiert werden. Danach bricht die Analyse aber vollständig ein, sodass der zweite Teil das Signals keine verwertbaren Ergebnisse mehr liefert.

Zwei ergänzende Systeme, die sich explizit auf den von Martin entwickelten Ansatz des „Blackboard"-System beziehen, werden vorgestellt von Bello [BMS00][BS00] und Monti [MS02]. Im Wesentlichen wird die Struktur beibehalten. Neu ist jeweils die Einführung einer zusätzlichen „Wissensquelle". Während Bello ein neuronales Netz benutzt, um die Erkennung von Akkorden möglich zu machen, findet sich bei Monti als neues Element das sogenannte „Fuzzy Inference System" (FIS), welches über Fuzzy-Logic-Ansätze versucht, die menschliche Strategie von Adaption und Prädiktion bezüglich veränderlicher Interpretation wahrgenommener Signale nachzubilden. Soweit angegeben, lassen die erzielten Ergebnisse aber nicht auf eine deutliche Verbesserung der von Martin dargelegten Resultate schließen.

Kashino et al. [KNKT98] basieren ihr System auf der Interpretation musikalischer Perzeption als Ausprägung auditorischer Szenenanalyse [Bre90]. Ziel ist die Erkennung von harmonischen Strukturen und nach Schallquellen getrennten Noten. Analog zu Martin [Mar96b][Mar96a] wird eine „Blackboard"-Architektur benutzt. Die Erweiterung um ein Wahrscheinlichkeitsmodell („Bayesian Probability Network" [Pea86]) soll die Integration verschiedener Informationsquellen ohne globale Kontrollstruktur ermöglichen.

In einer Vorverarbeitungsstufe wird das monaurale Musiksignal mittels Frequenzanalyse in seine akustische Energierepräsentation (Spektrogramm) überführt. Mittels der sogenanntem „Pinching plane"-Methode werden kontinuierliche und zeitlich ausgedehnte Frequenzkomponenten extrahiert, d.h. über „Least-Square-Fitting" werden sattelartige Ebenen an die Spektrogrammamplituden ausgeprägter Partialtöne angepasst. Anschließend werden die Analysedaten aus der Vorverarbeitungsstufe in einem dreischichtigen Hypothesen-Netzwerk entsprechend ihres Abstraktionsgrades verteilt. Es werden Wahrscheinlichkeitsverteilungen für einzelne Frequenzkomponenten, Noten

und Akkorde hierarchisch aufgebaut. Über das Wahrscheinlichkeitsmodell werden dann unter Einbeziehung der theoretischen Vorgaben der Wissensquellen neue Informationen innerhalb des Netzwerks zwischen den Elementen propagiert und bei zeitlicher Synchronität Elemente unterer Schichten zu höheren Einheiten fusioniert. Die Gesamtheit der Wissensquellen teilt sich in drei Bereiche. Neben musiktheoretischen Grundlagen (statistische Akkordübergangswahrscheinlichkeiten und -noteninhalte) werden physikalische Charakteristika von 5 Instrumenten in einer Notendatenbank als klangfarbenbeschreibende Feature-Vektoren verwendet. Zusätzlich finden sich noch Annahmen über Regeln auditorischer Perzeption als Speicher höherschichtigen Vorwissens.

Evaluiert wird der Ansatz mittels Testmuster von Repräsentanten unterschiedlicher Instrumentengruppen (Klarinette, Klavier, Flöte, Trompete und Geige). Jeweils 2 - 3 Noten werden gleichzeitg dargeboten. Unterteilt werden die Testmuster in 3 Klassen: mindestens 2 Noten befinden sich in Oktavverwandtschaft, mindestens 2 Noten befinden sich in Quintverwandtschaft sowie übrige Notenkombinationen. Diese Ausgangsdaten werden getestet mit drei unterschiedlichen Informationsstufen des vorgestellten Systems: reine perzeptuelle Schallorganisation, Benutzung von Teilinformation sowie Einbeziehung sämtlicher vorhandener Informationen. Der Bereich der richtig erkannten Noten und Instrumente schwankt je nach Testmusterklasse beträchtlich innerhalb eines Informationsniveaus um bis zu 50 Prozent. Global betrachtet verbessert aber die Integration höherer Wissensstufen die Erkennungsraten innerhalb aller untersuchten Testmustergruppen.

Ein Verfahren zur Transkription mehrstimmiger harmonischer, nichtperkussiver Musiksignale in das MIDI-Format beschreiben Martins und Ferreira [MF02]. Die Implementierung basiert ausschließlich auf technischen Ansätzen, d.h. die von vielen anderen aktuellen Algorithmen verwendeten perzeptiven Grundlagen werden weitestgehend nicht berücksichtigt. So findet sich als Vorverarbeitungsstufe eine Frequenztransformation umgesetzt als „Odd DFT“ [Fer98] mit 50%-Overlap-Add-Schema. In der anschließenden harmonischen Analyse werden quasistationäre Sinusschwingungen als Repräsentanten der Partialtöne gesucht. Das genaue Vorgehen folgt allgemein bekannten Strategien. Im Leistungsspektrum werden signifikante Peaks gesucht und diese zu Obertonreihen kombiniert. Einschränkung ist die Präsenz des Grundtons, der somit nicht durch andere spektrale Komponenten verdeckt sein darf. Probleme treten offensichtlich wie üblich bei harmonisch eng

verwandten Intervallen auf (Oktaven, etc.). Positiv erscheint die verbesserte spektrale Auflösung der Frequenztransformation mittels Interpolation umliegender Transformations-Bins.

Anschließend werden dann die harmonischen Strukturen in der Zeit verfolgt und bei Einhaltung gewisser Kontinuitätsbedingungen (Frequenz und Zeit) zu Trajektorien fusioniert. Der Autor hält diese Vorgehensweise für robuster als die Verfolgung von einzelnen Partialtontrajektorien, begründet diese Behauptung aber nur unzureichend.

Der hieraus resultierende Satz von Pitchtrajektorien wird in einem Nachverarbeitungsschritt mittels bekannter Energiesegmentierung in einzelne Subsegmente unterteilt, sowie einem Fine-Tuning bezüglich der Trajektorienanfänge unterzogen. Abschließend werden dann die erhaltenen Einheiten in zeitlich und harmonisch verwandte Cluster zusammengefasst. Die zugehörige Berechnung von Notenwahrscheinlichkeiten soll Oktavvertauschungen und andere harmonische Fehleinschätzungen eliminieren.

Nicht adäquat erscheint die Zuordnung von Noten zu den Trajektorien. Berechung des Mittelwertes und Quantisierung auf „Kammerton A“-Stimmung (440 Hz) setzen perfekte Standardintonation voraus.

In den angegebenen Transkriptionsbeispielen finden sich nur konstruierte Laborsignale, deren Qualität subjektiv beurteilt wird. Neben synthetisierten Dreiklängen (Flöte, Cello, Klarinette) wird ein kurzer bis zu vierstimmiger Digitalpianoauszug untersucht. Viele Noten sind richtig, aber der Fehler liegt schätzungsweise bei ungefähr 50 Prozent. Die Eignung des Verfahrens für reale Signale bleibt ungeklärt.

Die Arbeiten von Klapuri et al. [KVH00][Kla01b][Kla01a][KVES01] beschäftigen sich im Umfeld polyphoner Transkription hauptsächlich mit der Bereitstellung eines Verfahrens zur zuverlässigen Bestimmung multipler Pitchfrequenzen. Der vorgestellte „Multipitch Estimator“ versucht, wie üblich, über geeignete Interpretation harmonischer Obertonstrukturen die Grundfrequenzen zu ermitteln. Iterativ werden nacheinander die jeweils dominanten Pitchfrequenzen bestimmt und deren zugehörige Partialtonelemente vom Spektrum subtrahiert. Eine Reihe teilweise auditorisch motivierter Modifikationen soll die Robustheit des Ansatzes erhöhen.

Im Vorverarbeitungsschritt wird ein relativ breiter „Hamming“-gefensterter Ausschnitt (Länge bis zu 200 ms) fouriertransformiert und das Spektrum gehörgerecht in 18 Bänder zwischen 50 Hz und 6000 Hz mit 50 % Überlappung aufgeteilt. Unter Benutzung von Amplituden und Frequen-

zen wird mit Hilfe eines Inharmonizitätsfaktors, bezogen auf die Abweichung von der idealen Obertonreihe, für jeden Frequenz-Bin ein lokaler F_0-Ähnlichkeitsvektor aufgebaut. Die Summierung überlappender Frequenz-Bins der Frequenzbänder ergibt einen globalen Ähnlichkeitsvektor, dessen Maximum die wahrscheinlichste Partialtonreihe repräsentiert.

Grundlegende Probleme ergeben sich bei der nachfolgenden Entfernung der signifikantesten Teiltonstruktur. Aufgrund harmonischer Verwandtschaft „teilen“ sich mitunter musikalische Töne die gleichen Teilfrequenzen, welche nach simpler Subtraktion der entsprechenden spektralen Analyseoutputs im nächsten Iterationsschritt „fehlen“. Zur Lösung des Problems bedienen sich die Autoren einer Erkenntnis der Arbeit von Bregman [Bre90], wonach das menschliche auditorische System ein glattes Spektrum mit nach steigender Frequenz abnehmenden Amplituden zur Zusammenfassung einzelner Partialtöne in eine gemeinsame Struktur bevorzugt. Glättung des gefundenen Partialtonspektrums soll diesen Effekt nachbilden. Als Residualspektrum bleibt für den folgenden Analyseschritt für jeden Frequenz-Bin das Minimum aus Originalspektrum und dem geglätteten Partialtonverlauf übrig.

Die vorgestellten Resultate zeigen vielversprechende Ergebnisse für Mixturen von 1 - 6 Einzeltönen diverser Instrumentengattungen und bewegen sich im Qualitätsbereich ausgebildeter Musiker. Untersuchungen zur Transkription realer Signale können aber die Praxistauglichkeit des Verfahrens nicht nachweisen. Hier verhindern offensichtlich die für den Ansatz notwendigen langen Fensterbreiten eine adäquate Zeitauflösung. Ebenso unterteilt der in Kap. 2.2 beschriebene Segmentierungsalgorithmus [Kla99] lange Noten häufig in mehrere Abschnitte.

Virtanen [VK01][VK02] versucht in weiterführenden Arbeiten, den groben zeitlichen Verlauf der vom „Multipitch Estimator“ bereitgestellten Frequenzen detaillierter aufzulösen. Über „Least-Square“-Methodik [DH97] werden in kürzeren Zeitfenstern die Amplituden, Frequenzen und Phasen von harmonischen Sinustonreihen so angepasst, dass dadurch das Audiosignal am besten angenähert wird. Aufgrund der notwendigen Anzahl von 5 - 10 Iterationsschritten pro Zeitpunkt zeigt sich das Verfahren aber extrem rechenaufwendig ohne die Ergebnisse von Klapuri signifikant zu verbessern.

Ein hierzu analoges Verfahren wird vorgestellt von de Cheveigné und Kawahara [dCK02], das über kaskadierte Kammfilter im Zeitbereich eine vorher bekannte Anzahl von Periodizitäten annähert. Mangelnde Flexibilität bezüglich variierender Stimmenanzahl und Voraussetzung näherungsweise perfekt harmonischer Partialtöne deuten aber nicht auf eine mögliche all-

gemeingültige Anwendung auf reale Signale hin.

Den konkretesten Ansatz bezüglich einer realen Anwendung stellen Goto und Hayamizu vor [GH99][Got00]. Die wesentliche Annahme besteht darin, dass für das Verständnis polyphoner Musik keine komplette Transkription notwendig ist. Vielmehr genügt die Extraktion der dominanten Inhalte in symbolische Darstellungen. Leider versäumen es die Autoren, eine adäquate Repräsentation der extrahierten Frequenzverläufe im Sinne abstrakter Metadaten bereitzustellen. Möglicherweise ist dies auf unzureichende Segmentierungsergebnisse zurückzuführen. Ziel der Implementierung ist die Bereitstellung der zeitlichen Verläufe dominanter Frequenzlinien für Hauptmelodie und Bass-Stimme.

Der Algorithmus versucht über die „Expectation-Maximization"-Methode (EM) [DLR77] iterativ für jeden Zeitpunkt die plausibelsten Obertonreihen zu finden. Dazu wird das Ergebnis einer Kurzzeit-Fouriertransformation bandpassgefiltert. Bei einer Trennfrequenz von ungefähr 261 Hz werden die Basslinie im unteren und die Melodie im oberen Frequenzbereich angenommen. Für etwaige Grundfrequenzen werden jeweils Wahrscheinlichkeitsdichtefunktionen (PDF „Probability Density Function") bezüglich der Partialtöne berechnet und anschließend signifikante Peaks innerhalb der PDF über die Zeit in Grundfrequenzlinien zusammengefasst. Die evidentesten Trajektorien werden als Ergebnis ausgewählt.

Erweiterungen der Pitcherkennung erläutert Goto in [Got01b]. Verallgemeinerung der Obertonreihenstruktur sowie adaptive Einschätzung des aktuellen Pitch durch Ergebnisse vorangegangener Analysetakte verbessern das Verfahren.

Praktische Ergebnisse werden in der Anwendung auf 10 ausgewählte kurze Ausschnitte von Aufnahmen verschiedener Genres mit dominanten Melodielinien vorgestellt. In den relevanten Abschnitten entsprechen die extrahierten Frequenzen meist den dominanten Inhalten. Allerdings werden sowohl in solistischen Passagen als auch in Nebenstellen häufig sekundäre Stimmen detektiert. Zudem treten mitunter Oktavfehler auf. Die Autoren geben die Erkennungsraten mit 88,4 % für die Melodie und 79,9 % für die Basslinie an. Allerdings fehlt die Erläuterung der Vorgehensweise zur Bestimmung der Zahlenwerte.

Eine Anzahl von Arbeiten findet sich, denen die Einschränkung auf Pianomusik gemeinsam ist. Diese Vorgehensweise versucht, von der wohldefinier-

ten Form der zu analysierenden Signale zu profitieren. Die spezielle Bauweise von Klavieren sorgt für eindeutig scharfe Notenanfänge sowie vibrato- und schwankungsfreie Notenverläufe. Allerdings geht dies in der Regel auch einher mit dem Verlust allgemeingültiger Prinzipien bezüglich anderer Instrumentengruppen.

Den hierbei am wenigsten eingeengten Ansatz stellt Marolt [Mar01] vor, dessen Arbeit stark auf der Verwendung neuronaler Netze beruht. Zur Vorverarbeitung verwendet er ein auditorisches Modell, bestehend aus einer Gammaton-Filterbank mit anschließender Haarzellentransduktion. Ein Netzwerk adaptiver Oszillatoren sucht im Ergebnis nach Partialtönen. Diese werden mittels eines weiteren Netzwerkes zeitverzögerter Neuronen zur Notenerkennung genutzt.

Weitaus spezieller zeigen sich drei untereinander ähnliche Ansätze, die allesamt auf dem charakteristischen Spektrum von Pianonoten basieren. Unterschiede finden sich bei Bereitstellung der Datenbank mit der Information über die zu erwartenden Frequenzverteilungen. Während Raphael [Rap02] und Bello [BDS02] aus Originalaufnahmen der verwendeten Instrumente ihre spektralen Modelle aufbauen, nutzen Ortiz-Berenguer und Casajus-Quiros [OBCQ02] physikalische Prinzipien der Klaviermechanik zur Bereitstellung der Frequenzinformation.

Der Vollständigkeit halber zum Abschluss noch einige Systeme mit unterschiedlichen Schwerpunkten. Solbach [Sol98] befasst sich mit der Extraktion von Partialtonverläufen und der Lokalisierung von Notenanfängen. Kernpunkt der Arbeit ist die Implementierung einer Gammaton-Filterbank über „Wavelets“ unter Einbeziehung zeitlicher und spektraler Verdeckungseffekte.

In Kombination mit üblichen Annahmen über harmonische Frequenzverteilungen werden verschiedene Wahrscheinlichkeitsmodelle benutzt von den Verfahren von Verfaille et al. [VDC01], Walmsley et al. [WGR99] (Bayes) sowie Rosier und Grenier [RG02], die jeweils aus den plausibelsten Hypothesen Einschätzungen der musikalischen Inhalte kreieren.

Kapitel 3

Grundlagen der Hörwahrnehmung

3.1 Physiologie: Auditorische Peripherie und zentrales Gehör

Die physiologischen Gegebenheiten der menschlichen auditorischen Peripherie sind mittlerweile gut erforscht und können in einer Vielzahl wissenschaftlicher Abhandlungen [SS01][Zen94][DS94][DPF96] [SSZ97][HMPF95][Moo95] nachgeschlagen werden. Daher sollen an dieser Stelle nur die wesentlichen und zum weiteren Verständnis späterer Kapitel notwendigen Grundlagen dargestellt werden.

Der periphere Schallverarbeitungsapparat des Menschen (s. Abb. 3.1) besteht aus der Gesamtheit von Außen-, Mittel- und Innenohr. Durch den äußeren Gehörgang gelangt Schall zum Trommelfell und wird im Mittelohr über die Gehörknöchelchen weitergeleitet. Anschließende Verarbeitung im Innenohr bewirkt eine frequenzabhängige Transduktion der mechanischen Schwingungen in neuronale Nervenaktionspotentiale und Weitergabe dieser an die angeschlossenen Hörnervenfasern.

3.1.1 Außenohr

Das äußere Ohr bildet einen Trichter, der die einfallenden Schallwellen zum Trommelfell leitet. Ohrmuschel, Gehörgang, Schädelform und Schulter modifizieren das Schallsignal.

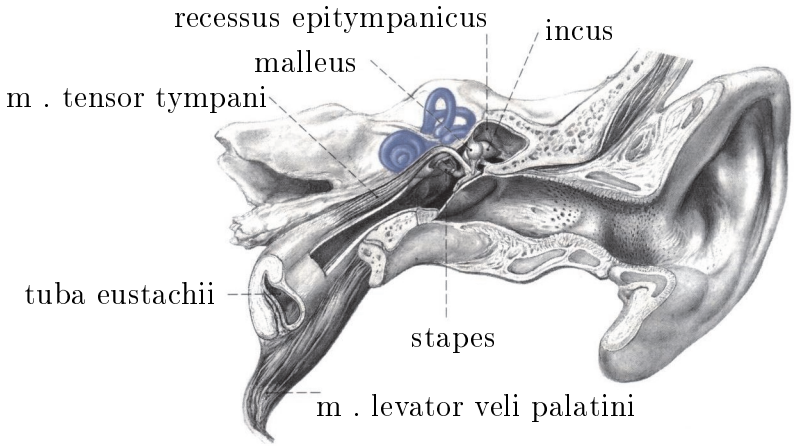

Abbildung 3.1: Auditorische Peripherie (aus [DS94])

Da der Gehörgang (inkl. Ohrmuschel) an einem Ende geöffnet und am anderen geschlossen ist, wird er physikalisch näherungsweise als halboffenes Rohr aufgefasst. Somit kann im Resonanzfall, d.h. wenn ein Viertel der Schallwellenlänge der effektiven Gehörkanallänge entspricht, ein Schalldruckpegelgewinn beobachtet werden. Im Resonanzmaximum bei ungefähr 2500 Hz beträgt die Verstärkung bis zu 20 dB. Eine zweite Resonanz („Cavum-Conchae-Resonanz“) wird zwischen 2000 Hz und 2500 Hz allein durch die Ohrmuschel hervorgerufen.

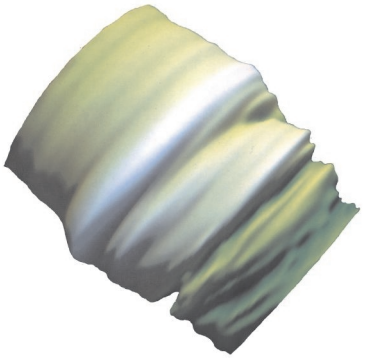

Abbildung 3.2: Außenohrübertragungsfunktion - Zylindrische Oberflächendarstellung der Betragsantwort als Funktion des Azimuthwinkel auf radialer Achse (aus [Beg94])

In Abhängigkeit von der Schalleinfallsrichtung werden als Resultat der Außenohrform durch sogenannte „richtungsbestimmende Bänder“ [Bla96] einzelne schmale Frequenzbereiche angehoben bzw. abgesenkt. Dadurch wird

bis zu einem gewissen Maß die Lokalisation eintreffenden Schalls auch ohne binaurale Zeit- und Intensitätsunterschiede insbesondere in der vertikalen Ebene (Median-Sagittal-Ebene) möglich.

Zusammenfassen kann man die beschriebenen Phänomene in der Außenohrübertragungsfunktion („Head related transfer function“ HRTF), die in Abbildung 3.2 dargestellt ist.

3.1.2 Mittelohr

Die wesentliche Aufgabe des Mittelohres (MO) besteht in der Anpassung der Schallkennimpedanzen von Luft und den Flüssigkeiten im Innenohr (s. [SS01] und [Zen94]). Bei Fehlen einer solchen Funktionalität würden wie im Fall der Schallleitungsschwerhörigkeit bis zu 98 % der einfallenden Schallenergie reflektiert. Bei gesundem Mittelohr können aber ungefähr 60 % der Signalintensität an das Innenohr weitergegeben werden. Die hierfür notwendige Schalldruckverstärkung wird möglich durch die aneinandergereihte Kopplung von Trommelfell, den drei Gehörknöchelchen (Hammer, Amboss und Steigbügel) sowie dem ovalen Fenster als Kontaktstelle zum Innenohr (s. Abb. 3.3).

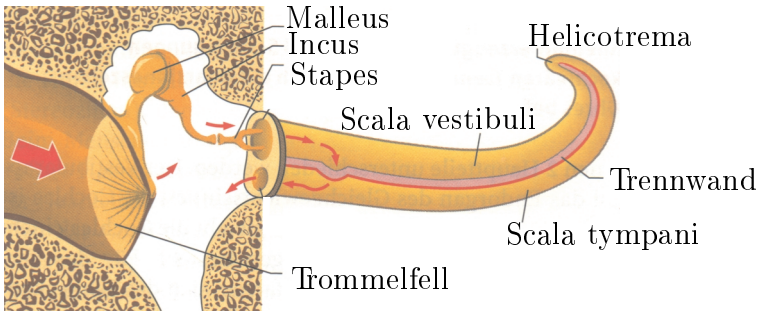

Abbildung 3.3: Schema von Mittelohr und aufgerollter Cochlea (aus [SS01])

Drei unterschiedliche Mechanismen sind verantwortlich für diese Impedanztransformation:

1. Flächenverhältnis zwischen Trommelfell A_T und Steigbügelfußplatte A_S:

$$\frac{A_T}{A_S} \simeq 17 \,, \tag{3.1}$$

2. Verhältnis der Hebelarme von Hammer l_H und Amboss l_A:

$$\frac{l_H}{l_A} \simeq 1,3 \ , \tag{3.2}$$

3. Hebelarm durch die Biegung des Trommelfells und die unsymmetrische Aufhängung des Hammers:

$$F_T \simeq 1,4 \ . \tag{3.3}$$

Die Gesamtverstärkung errechnet sich zu:

$$\frac{p_{ges}}{p_T} = F_T \cdot \frac{A_T}{A_S} \cdot \frac{l_H}{l_A} \simeq 30 \ dB \tag{3.4}$$

(p_T: Schalldruck am Trommelfell).

Bemerkenswert ist die Bedeutung der Transferfunktion des MO [DRS01], die sich wie ein Bandpassfilter mit breitem Durchlassbereich verhält. Im Niederfrequenzbereich wird sie begrenzt durch die mechanischen Eigenschaften von Trommelfell und ovalem Fenster. Bei hohen Frequenzen limitieren die Trägheitsmomente und Reibungs- bzw. Biegungsverluste der Gehörknöchelchen die Übertragung.

Vergleicht man den Verlauf der MO-Übertragungsfunktion mit dem der Hörschwelle (s. Abb. 3.4), so sieht man, dass die Hörempfindlichkeitskurve weitestgehend durch die mechanischen Eigenschaften von mittlerem und äußerem Ohr bestimmt wird.

Eine zusätzliche Aufgabe erfüllen die Muskeln des MO (M. tensor tympanus und M. stapedius, s. Abb. 3.1). Durch reflektorische Kontraktion kann die MO-Steifigkeit erhöht und so eine Dämpfung tiefer Frequenzen erreicht werden. Begrenzter Schutz gegenüber hohen Pegeln und Verringerung der Wahrnehmung selbstproduzierter Laute sind die Folge.

3.1.3 Innenohr

Die Struktur des Innenohres setzt sich aus zwei Einheiten zusammen. Während das Vestibularorgan einen Bestandteil des Gleichgewichtssystems darstellt, bildet der Aufbau der Cochlea den abschließenden Teil der auditorischen Peripherie (s. eingefärbte Sektion in Abb. 3.1). Anatomisch gleicht die Cochlea einem Schneckenhaus mit zweieinhalb Windungen. Sie ist durch

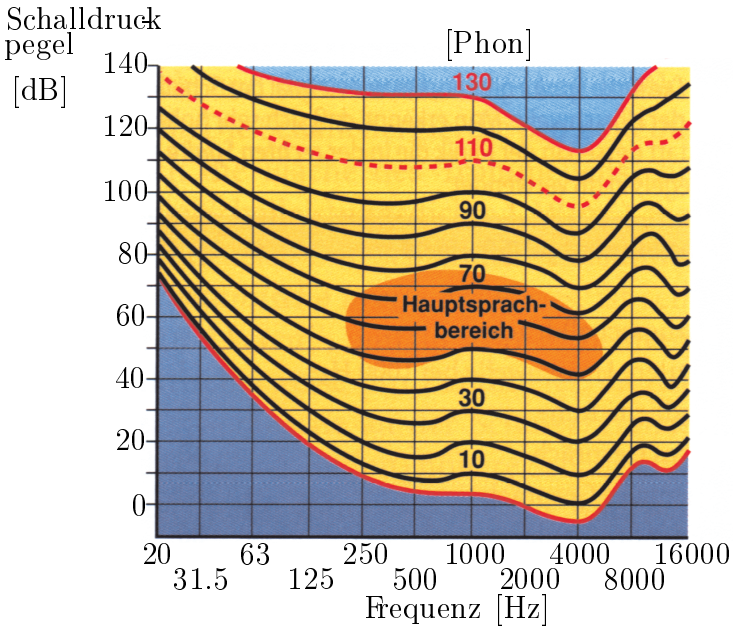

Abbildung 3.4:
Hörfläche
(aus [SS01])

die cochleäre Trennwand (CT) in die zwei Perilymphflüssigkeit enthaltenden Kammern „Scala vestibuli“ (SV) und „Scala tympani“ (ST) geteilt (s. Abb 3.3).

Die Arbeitsweise der Cochlea lässt sich wiederum in zwei Abschnitten beschreiben. Der hydromechanische Teil wird bestimmt durch die makro- und mikromechanischen Eigenschaften des Schneckeninneren. Die eigentliche Funktionseinheit zur Umwandlung der Eingangssignale in neuronale Repräsentationen befindet sich innerhalb der cochleären Trennwand.

Die Scala Vestibuli ist mit dem Mittelohr verbunden über das ovale Fenster (OF). Dieses schwingt mit der Steigbügelbewegung und zwingt so die inkompressible Lymphflüssigkeit zum Ausweichen. Die Ausweichbewegung wird an die cochleäre Trennwand weitergegeben und bildet eine Wanderwelle in Richtung des Helicotrema (HC), Cochlea-Spitze, aus. Aufgrund der längs ihrer Ausdehnung kontinuierlich veränderlichen mechanischen Eigenschaften (Massenbelag, Steifigkeit, Breite, etc.) bildet die Trennwand an bestimmten Stellen frequenzabhängige Resonanzen aus. Diese tonotopische Frequenzselektivität wird auch als Ortstheorie bezeichnet. Den charakteristischen Frequenzen können auf der Trennwand Orte der maximalen Wellenamplituden zugeordnet werden, die kontinuierlich von hohen Frequenzen im Bereich des

ovalen Fensters bis zu tiefen Frequenzen am Helicotrema reichen. Über diese Dispersionseigenschaft können Frequenzinhalte im einkommenden Audiosignal bis zu einem gewissen Grad aufgespalten werden.

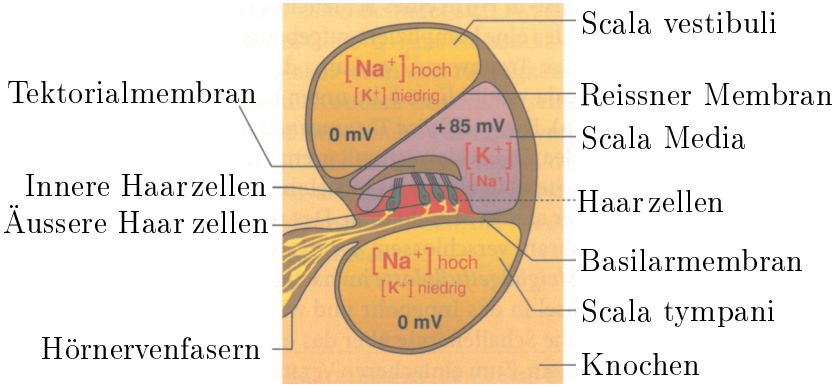

Abbildung 3.5: Querschnitt der Cochlea (aus [SS01])

Unterstützt wird diese Funktionalität durch die Eigenschaften der cochleären Trennwand (s. Abb 3.5). Diese wird zur Scala Vestibuli hin abgeschlossen durch die Reissnersche Membran (RM). Die Grenzfläche zur Scala Tympani besteht aus der Basilarmembran (BM) mit aufsitzendem Cortischen Organ (CO), auf dessen Oberseite sich in Längsrichtung drei Reihen äußerer Haarzellen und eine Reihe innerer Haarzellen befinden. Diese Haarzellen werden wiederum überspannt von der Tektorialmembran (TM). Im Bereich dazwischen befindet sich die Endolymphflüssigkeit der Scala Media. Bei Bewegung der cochleären Trennwand geraten die Tektorialmembran und das Cortische Organ in Relativbewegung, was zu einer Auslenkung der auf den Haarzellen befindlichen Sinneshärchen führt. Dies geschieht teils durch direkten Kontakt, teils aber auch durch hydrodynamische Kopplung. Die äußeren Haarzellen besitzen nun die Fähigkeit, sich in Abhängigkeit der Trennwandschwingung sehr schnell zu verkürzen bzw. zu verlängern. Dies führt zu einer bis zu 1000-fachen Verstärkung der Wanderwellenamplituden und liefert spitze und ausgeprägte Schwingungsmaxima.

Ebenso wie die Sinneshärchen der äußeren Haarzellen werden diejenigen der inneren Haarzellen durch die Relativbewegung von Tektorialmembran und Cortischem Organ ausgelenkt. In Folge dieser Bewegung werden biochemische Prozesse in Gang gesetzt, die eine Transduktion der mechanischen Bewegungen in neuronale Aktionspotentiale bewirken (s. Abb. 3.6).

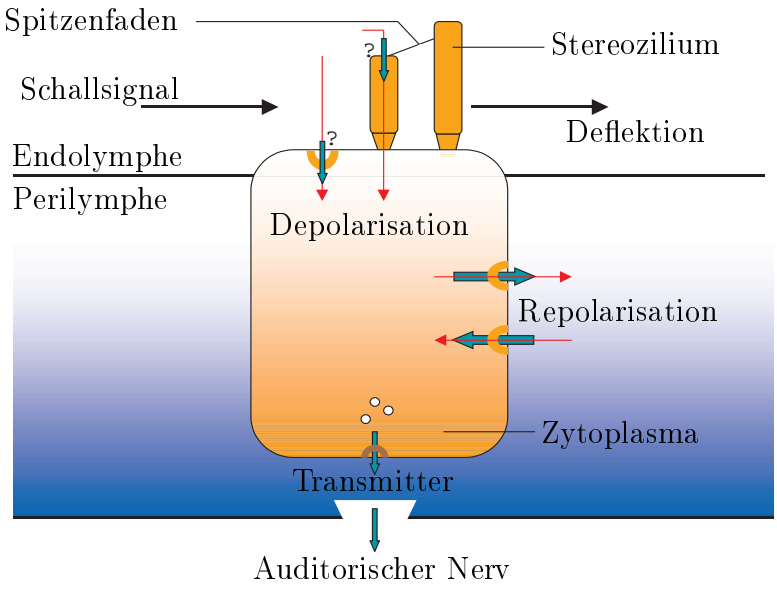

Abbildung 3.6: Schema der Haarzelle (in Anlehnung an [SS01])

Im Ruhezustand besitzen die inneren Haarzellen ein Ruhemembranpotential von ungefähr -40 mV sowie eine niedrige Kalium-Konzentration. Die umgebende Flüssigkeit der Scala Media weist hingegen einen ungewöhnlich hohen Anteil an Kaliumionen auf und ist positiv geladen. Bei Deflektion der Sinneshärchen in eine Richtung öffnen sich sogenannte Transduktionsionenkanäle, durch die zwecks Potentialausgleich ein Einstrom positiv geladener Kaliumionen in die Haarzelle erfolgt. Auslenkung der Sinneshärchen in die entgegengesetzte Richtung verschließt diese Kanäle und durch Ionenverbindungen in die basolaterale Zellmembran kann das ursprüngliche Potential wiederhergestellt werden. Bei geöffneten Kanälen bewirkt das geänderte Sensorpotential eine vermehrte Freisetzung von afferenter Transmittersubstanz. Diese diffundiert durch den synaptischen Spalt in Richtung Hörnerv. In Abhängigkeit von der Transmitterkonzentration im synaptischen Spalt steigt die Wahrscheinlichkeit der Auslösung eines Nervenaktionspotentials (NAP).

Bis zu einer Frequenz von knapp 5000 Hz folgt die Freisetzung von Transmittersubstanz hochgradig synchron der Deflektion der Sinneshärchen. Somit kann eine lineare Frequenzübertragung über zeitliche Kodierung erfolgen, was

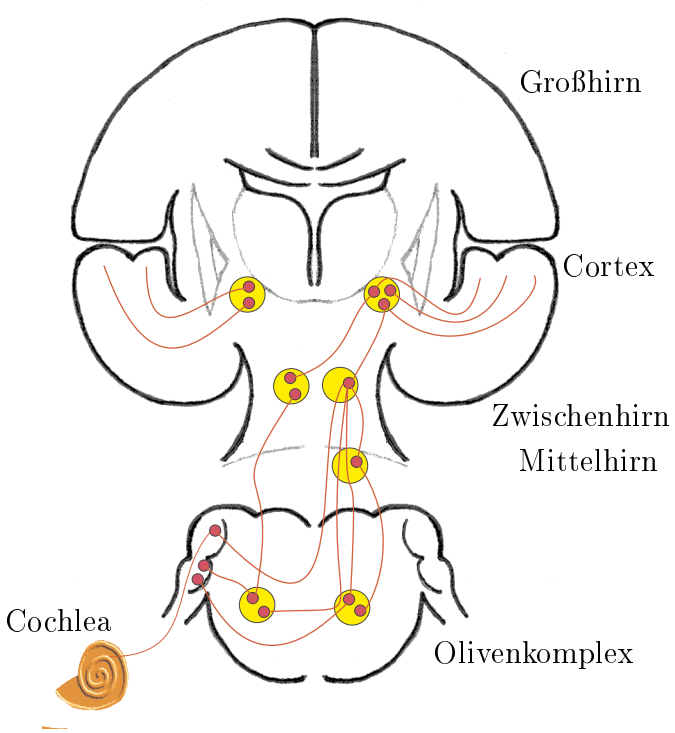

Abbildung 3.7: Hörbahn (in Anlehnung an [SS01])

in der Literatur auch unter dem Begriff „Phase-Locking“ zusammengefasst wird.

3.1.4 Zentrales Gehör

Die zeitlich veränderliche Konzentration von Transmittersubstanz, als Folge der Transduktionsfunktionalität der inneren Haarzellen, führt zu einer Kette neuronaler Erregungen längs der zentralen Hörbahn (s. Abb 3.7). Mit Hilfe von Nervenaktionspotentialen werden die Signale über die Hörnerven, den Hirnstamm und weitere Zwischenstufem der Hörbahn bis zum auditorischen Cortex im Temporallappen weitergeleitet. Zu Beginn der Verarbeitungsfolge ist das Schallsignal durch die Entladungsrate (Schalldruckpegel), Zeitdauer der Aktivierung (Länge des Schallreizes) und die Anregung frequenzspezifischer Haarzellenbereiche kodiert. Nachfolgende Neurone sind zunehmend auf komplexere Schallmuster spezialisiert.

Als Beispiel sei die ausgeprägte Fähigkeit räumlichen Hörens genannt [Bla96]. Hochspezialisierte Neurone reagieren auf Laufzeit- und Intensitätsunterschiede zwischen rechtem und linkem Ohr. Dies wird erst möglich durch die Verknüpfung der kontralateralen Hörbahnstränge, wodurch die binauralen Schallsignale vergleichbar gemacht werden.

3.2 Gestaltpsychologie

Da sich die Quantisierung der im zentralen Gehör stattfindenden Verarbeitungsprozesse durch physiologische Messungen noch in einer Frühphase befindet, sollen die dort vorkommenden Schritte der menschlichen Schallverarbeitung mit Ansätzen aus der kognitiven Psychologie [And01] nachempfunden werden.

In dieser Arbeit werden Prinzipien der sogenannten „Gestaltpsychologie“ verwendet, die sich mit dem Aufbau der Welt der Wahrnehmung beschäftigen. Diese Theorie entstand in der ersten Hälfte des 20. Jahrhunderts als Gegenbewegung zur „Elementenpsychologie“ [Gol97]. Obwohl nicht mehr ganz neu, haben die dort aufgestellten wahrnehmungstheoretischen Postulate bis heute nicht an Bedeutung verloren und können nach wie vor eine breite Palette von Phänomenen erklären.

Die „Grundformel“ des Ansatzes kann man wie folgt zusammenfassen:

> *„Das Ganze, die Gestalt, ist etwas Anderes als die Summe Ihrer Einzelteile“.*

Oder wie es Wertheimer [Wer25], der maßgebliche Begründer der Theorie, beschreibt:

> *„Man könnte das Grundproblem der Gestalttheorie etwa so zu formulieren suchen: Es gibt Zusammenhänge, bei denen nicht, was im Ganzen geschieht, sich daraus herleitet, wie die einzelnen Stücke sind und sich zusammensetzen, sondern umgekehrt, wo - im prägnanten Fall - sich das, was an einem Teil dieses Ganzen geschieht, bestimmt von inneren Strukturgesetzen dieses seines Ganzen“.*

Mittels implizit vorhandener Gruppierungsstrategien und einer Abfolge hypothetischer Interpretationen der durch die sensorischen Organe aufgenommenen Informationen werden elementare Bausteine zu semantisch bedeu-

tungsvolleren Objekten verschmolzen. Dabei besteht mitunter ein signifikanter Unterschied zwischen den real existierenden physikalischen Objekten und den erzeugten mentalen Darstellungen.

Grundsätzlich erwartet man in der Gestaltpsychologie, dass sich Objekte bzw. die Reizkonfigurationen, die diesen zugrunde liegen, als ganzheitliche Gebilde wahrgenommen werden. Diese Einheiten setzen sich aus zueinander in Beziehung stehenden Teilen zusammen, wobei diese Elemente durch ihren inneren Zusammenhalt als auch durch ihre Beziehungen zum Ganzen festgelegt sind. Dies bedeutet im Fall der Objekterkennung, dass der Wahrnehmungsprozess in der gesamten Reizkonfiguration, die von einem Objekt ausgeht, bestimmte Reizstrukturen nach irgendwelchen Kriterien als enger zusammengehörig als andere organisiert, sodass ein aus Teilen bestehendes, bedeutungshaltiges Objekt wahrgenommen wird.

Die Gestaltpsychologie versucht, die Gesetzmäßigkeiten zu finden, die die Organisation von Teilen zu einem Ganzen erklären; sie beschäftigt sich also mit der Frage nach den Einheiten der Wahrnehmung. Die wichtigsten Einflussfaktoren der Reizelemente (Ähnlichkeit, Nähe, Kontinuität und Verbundenheit) sind in den Gestaltgesetzen manifestiert:

1. Gesetz der Nähe:
   Elemente (eines Reizmerkmals) mit geringen Abständen werden als zusammengehörig wahrgenommen.

2. Gesetz der Ähnlichkeit:
   Einander ähnliche Elemente werden eher als zusammengehörig erlebt als einander unähnliche.

3. Gesetz der guten Fortsetzung:
   Elemente, die eine gegebene Tendenz fortsetzen, bilden bevorzugt eine Gruppe.

4. Gesetz des gemeinsamen Schicksals:
   Elemente, die eine gleichartig, gleichgerichtete Veränderung erfahren, werden häufig perzeptiv fusioniert.

Da diese Prinzipien sowohl im visuellen als auch im auditiven Bereich ähnlich angewandt werden können, sollen die Gestaltgesetze an zwei einfachen, aber prägnanten grafischen Darstellungen veranschaulicht werden.

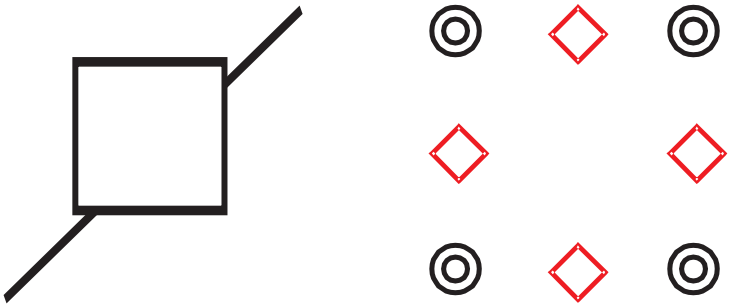

Abbildung 3.8: Gestaltprinzipien

In Abbildung 3.8 findet sich auf der linken Seite ein Beispiel für die Gruppierungsfaktoren Nähe und gute Fortsetzung. Die beiden gleichgerichtet verlaufenden Geradenabschnitte werden bei nicht zu großem räumlichen Abstand, wie in der Abbildung, mental zu einer zusammenhängenden Geraden verschmolzen, unabhängig von dem nicht sichtbaren Bereich hinter dem Rechteck.

Der Tatsache, dass sich die verschiedenen Gruppierungsstrategien mitunter konkurrierend gegenüberstehen, trägt das Beispiel auf der rechten Seite Rechnung. Wie vorher führen auch hier Nähe und gute Fortsetzung zur Wahrnehmung *eines* Quadrates, das sich aus sämtlichen Elementen zusammensetzt. Einflüsse der Ähnlichkeit können aber eine andere Interpretation zulassen. Sowohl die schwarzen Kreise als auch die roten Rauten werden in diesem Beispiel mitunter als eigenständige Einheiten interpretiert. Zwei zueinander um 45 Grad gedrehte Rechtecke sind das Ergebnis. Interessanterweise hat der Betrachter sogar die Möglichkeit, bewusst die eine oder andere Möglichkeit der Gruppierung zu steuern.

Die beiden simplen Beispiele verdeutlichen, dass es schon bei relativ einfachen Konfigurationen zu einer Reihe hypothetischer Interpretationen durch das menschliche Gehirn kommt, die abhängig von der individuellen Deutung unterschiedlichste Repräsentationen der Umwelt zulassen.

Im Bereich der musikalischen Wahrnehmung spielen die Erkenntnisse der Gestaltpsychologie ebenfalls eine gewichtige Rolle. Eine ausführliche Darstellung, die sich hauptsächlich auf die höheren Verarbeitungsschichten (Noten- und Akkordebene) bezieht, findet man bei Bregman [Bre90]. Dieser weist nach, dass vertikale und horizontale Gruppierungsstrategien angewandt wer-

den, um einzelne Noten in auditorische Ströme zusammenzufassen bzw. zu trennen („Sequentielle Integration“). Die enge Verwandtschaft zur visuellen Vorgehensweise lässt sich in analoger Interpretation der zugehörigen Notenschrift zeigen.

Beispielsweise bei der Untersuchung von in der Tonhöhe alternierenden Notenfolgen zerfällt die Wahrnehmung bei höherem Tempo und größeren Intervallen in zwei einzelne Ströme. Distanzen in Zeit *und* Pitch bestimmen die Gruppierung. Schon in den barocken Kompositionen wurde dieses Phänomen von den Komponisten implizit als beliebtes Stilmittel eingesetzt, um bei Verwendung nur eines einzelnen Instrumentes die Illusion von Mehrstimmigkeit hervorzurufen.

Eine Reihe weiterer Anwendungsfälle der Gestaltgesetze lassen sich im musikalischen Bereich finden. Die Klangfarbe von Musikinstrumenten kann dem Gesetz der Ähnlichkeit zugeordnet werden. Ein synchroner Modulationsverlauf von Partialtönen im Vibrato entspricht gemeinsamem Schicksal. Vorgänge der guten Fortsetzung erkennt man in der Obertonstruktur harmonischer Klänge oder auch in kreuzenden Melodielinien.

Neben den von Bregman beschriebenen übergeordneten Wahrnehmungsschichten werden in dieser Arbeit aber hauptsächlich die strukturell niedrigeren Verarbeitungsschritte untersucht. Ein Modell der hierarchischen musikalisch-melodischen Informationsverarbeitung wird in Kapitel 4.5 beschrieben. Dort wird ein Ansatz vorgestellt, wie prinzipiell die sukzessive Abstrahierung der sensorischen Daten in bedeutungsgeladene Muster umgesetzt sein könnte. Die Verwendung gestaltpsychologischer Kriterien ermöglicht dabei die Einführung einer Methode, die mit nur minimalem Vorwissen, wie etwa der Annahme harmonisch verteilter Obertonreihen realer Melodieinstrumente, eine Transkription in Melodien bereitstellt. Diese Vorgehensweise beinhaltet den großen Vorteil, allgemeingültig eine breite Palette unbekannter, auch synthetischer, Instrumentenklänge zu verarbeiten.

Im Sinne einer Optimierung der statistischen Ergebnisse des Transkriptionssystems bezüglich seines Einsatzes in einer anwendungsfähigen „Query-By-Humming“-Software wurden allerdings einige kulturell basierte Vorannahmen eingeführt. Solche angelernten Strategien werden im Kontext der Gestaltpsychologie als „schema-basiert“ bezeichnet und unterscheiden sich somit von angeborenen Vorgängen. Dementsprechend wird in späteren Analyseschritten (s. Kapitel 5.1.3) im Bezug auf musiktheoretische Betrachtungen beispielsweise vom westlichen Zwölftonsystem in wohltemperierter Stimmung oder auch bestimmten harmonischen Gesetzmäßigkeiten ausgegangen.

Kapitel 4

Modelle

Zur Nachbildung des Gehörs auf physiologischer Ebene wird im wesentlichen die von Baumgarte [Bau00] vorgestellte Arbeit zur Modellierung auditiver Wahrnehmungsschwellen verwendet. Die Beschreibung der inneren Haarzellen wird ersetzt durch das etablierte Modell von Meddis et al. [Med86][Med88] [MHS90], das aufgrund guter Übereinstimmung mit physiologischen Daten den Transduktionsprozess im Vergleich zu weiteren Modellen [HM91] am adäquatesten nachbilden kann.

Zur Implementierung der menschlichen Frequenz- bzw. Tonhöhenwahrnehmung findet das sogenannte „Phase-Locking“-Modell [Zen94][DRS01] Anwendung.

Ein auf die speziellen Anforderungen der musikalischen Melodiewahrnehmung vom Autor aufgebautes Hierarchiemodell interpretiert die im physiologischen Analyseprozess erhaltenen Ergebnisse.

4.1 Außen- und Mittelohr

Bei Vernachlässigung der Wirkung der Mittelohrmuskeln bei Pegeln oberhalb 80 dB SPL können Außen- und Mittelohr als linearer Filter aufgefasst werden. Aufgrund in der Regel unbekannter Schalleinfallsrichtung wird eine „mittlere“ Übertragungsfunktion angenommen. Unterhalb einer Frequenz von ungefähr 1 kHz erfolgt somit ein Anstieg von 6 dB pro Oktave. Die signifikante Gehörgangsresonanz findet sich als Überhöhung bei ca. 3 kHz. Oberhalb dieser Frequenz erfolgt ein konstanter Filterverlauf, d.h. stark personenvariierende Resonanzen werden nicht berücksichtigt. Implementiert wird diese

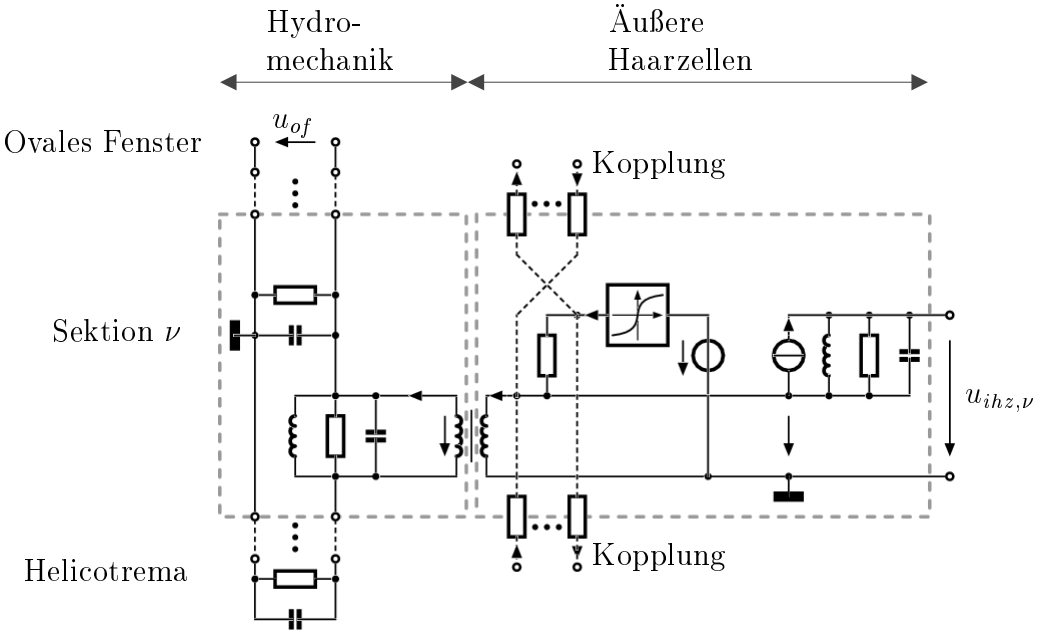

Abbildung 4.1: „Erweitertes Analogmodell“ (in Anlehnung an [Bau00])

Verarbeitungsstufe als passives elektrisches Netzwerk ohne Berücksichtigung von Phasengängen der Impulsantwort.

4.2 Innenohr (Cochlea)

Die Beschreibung der hydromechanischen Elemente des Innenohres sowie der äußeren Haarzellen erfolgt über das „Erweiterte Analogmodell“ von Zwicker und Peisl[ZP90]. Es handelt sich dabei um eine eindimensionale Darstellung der Cochlea, d.h. Abhängigkeiten von radialer und axialer Position werden ohne nennenswerte Einbußen der Genauigkeit vernachlässigt.

Die mechanischen Anteile des Modells erlauben die Simulation der im Innenohr durchgeführten Frequenz-Orts-Transformation sowie der damit verbundenen Frequenzselektivität (s. Kapitel 3.1.3). Durch aktive und nichtlineare Elemente wird die Verstärkungswirkung der äußeren Haarzellen, die unter anderem für Dynamikkompression, Verzerrungsprodukte und Suppression verantwortlich zeichnet, nachgebildet.

Das Modell setzt sich zusammen aus 251 gleichartigen, in Serie geschalteten Sektionen, die kleine longitudinale Abschnitte der Cochlea repräsentieren.

Der Tonheitsabstand benachbarter Teile beträgt somit 0.1 Bark. Die Subeinheiten können formuliert werden als System gekoppelter Differentialgleichungen. Die Benutzung der elektro-akustischen Analogien [ZZ87] erlaubt eine Repräsentation als elektrisches Netzwerk bestehend aus konzentrierten Elementen.

Das resultierende Schaltbild eines Cochlea-Abschnittes findet sich in Abbildung 4.1. Im Bereich der Hydromechanik (HM) modelliert der Parallelschwingkreis örtliche Nachgiebigkeit, Masse und Reibungsverluste der cochleären Trennwand. Die anderen Elemente beschreiben Masse und zugehörige Reibungsverluste der longitudinal bewegten Lymphflüssgkeit.

Die aktive und nichtlineare Wirkungsweise der äußeren Haarzellen, die eine Verstärkung der Basilarmembranschnelle bewirkt, wird als spannungsgesteuerte Spannungsquelle und punktsymmetrische Sättigungskennlinie umgesetzt. Innerhalb einer Rückkopplungsschleife wird das Ausgangssignal auf den Hydromechanikteil zurückgeführt. Die Kopplungswiderstände berücksichtigen außerdem die laterale Kopplung der einzelnen Sektionen über die äußeren Haarzellen.

Bei der zweiten Verstärkerstufe (Stromquelle und Parallelschwingkreis) handelt es sich um eine Erweiterung von Baumgarte, die Instabilitäten bei großen Verstärkungen vermeiden soll.

Die Simulation erfolgt mit Hilfe sogenannter Wellendigitalfilter (WDF) [Fet86] im Zeitbereich. Dies führt zu einer ausgezeichneten Zeitauflösung, wie sie sich für die späteren Schritte der Signalsegmentierung (s. Kapitel 5.1.2) als notwendig herausstellt.

4.3 Innere Haarzellen

An die Ausgänge der einzelnen Sektionen des Baumgarte-Modells wird jeweils eine Beschreibung einer inneren Haarzelle angeschlossen. Die limitierte Anzahl von Sektionen längs der Basilarmembran modelliert das Verhalten von benachbarten Haarzellen- bzw. Nervenfaserpopulationen.

Das hier verwendete Modell von Meddis et al. versucht eine Wahrscheinlichkeitsbeschreibung der Transduktionsvorgänge. Grundlegende Annahme ist, dass die Menge von Transmittersubstanz im synaptischen Spalt eine Funktion der Stimulusintensität darstellt. Weiterhin sei die Wahrscheinlichkeit der Auslösung eines Aktionspotentials auf den Hörnervenfasern proportional zur Spaltkonzentration (s. Kapitel 3.1.3).

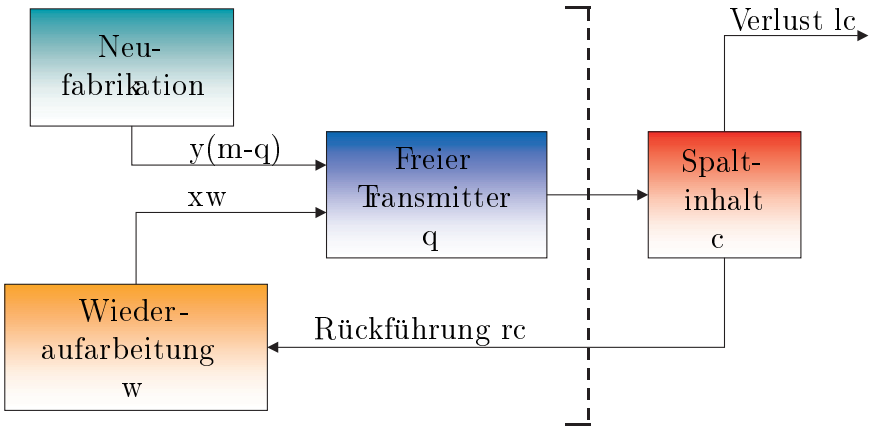

Abbildung 4.2: Haarzellenmodell von Meddis

In Abbildung 4.2 ist das Schema des Modells wiedergegeben. Die Transmittersubstanz wird zwischen verschiedenen Reservoirs in Abhängigkeit von den vorhandenen Konzentrationen und der Stärke des Eingangsstimulus ausgetauscht. Zentrale Größe für die Diffusion von Transmittersubstanz aus der Haarzelle in den synaptischen Spalt ist die Durchlässigkeit der Haarzellenmembran, die durch die Membranpermeabilität k wiedergegeben wird:

$$k = \begin{cases} gdt \cdot \left(\frac{s+A}{s+A+B}\right) & : \quad s + A \geq 0 \\ 0 & : \quad s + A < 0. \end{cases} \tag{4.1}$$

Der Parameter A beschreibt die niedrigste Anregungsamplitude, bei der die Membran permeabel wird. B bestimmt die Steilheit der Permeabilitätskurve und s die Stimulusintensität. Der Simulationstakt wird bestimmt durch das im Idealfall infinitesimale Zeitintervall dt.

Wenn q den freien Transmitter innerhalb der Haarzelle darstellt, so handelt es sich bei $kqdt$ um die Transmittermenge, die pro Simulationstakt in den synaptischen Spalt befördert wird. Während ein Teil der Transmitterkonzentration c im Spalt verloren geht (lc), wird ein anderer wieder zurückgeführt (rc) und kann im weiteren Verlauf erneut genutzt werden (xw). Im Reservoir „Neufabrikation“ wird neuer Transmitter produziert, der in Abhängigkeit von der vorhandenen Konzentration die Substanzverluste ausgleicht $(y(m-q))$. Die beschriebenen Vorgänge können in einem System von Differentialgleichungen folgendermaßen dargestellt werden:

$$\frac{dq}{dt} = y(m - q) + xw - kq, \tag{4.2}$$

$$\frac{dc}{dt} = kq - lc - rc, \tag{4.3}$$

$$\frac{dw}{dt} = rc - xw. \tag{4.4}$$

In Tabelle 4.1 sind die Modellparameter zusammengefasst. Die Werte gelten

A	=	10
B	=	3000
x	=	66.31
g	=	1000
y	=	5.05
m	=	1
l	=	2500
r	=	6580

Tabelle 4.1: Parameter für mittlere spontane Aktionspotentialraten

für Haarzellen mit mittleren spontanen Aktionspotentialraten, die eine Dynamik von 45 dB bereitstellen. Die Einbeziehung des von Meddis ursprünglich verwendeten Wahrscheinlichkeitsmodells zur Auslösung von neuronalen Impulsen wird in dieser Arbeit nicht benutzt. Anstatt dessen repräsentiert die Konzentration von Transmittersubstanz im synaptischen Spalt direkt die Aktivität auf den Hörnervenfasern.

4.4 Phase-Locking

Eine Art der Kodierung von Frequenzinhalten eines Schallsignals geschieht im Innenohr über tonotopische Abbildung der einzelnen Anteile entlang der Basilarmembran (s. Kapitel 3.1.3). In dieser als Frequenz-Orts-Transformation bezeichneten Funktionalität erzeugen die teils nichtlinearen Eigenschaften des Innenohres an charakteristischen Stellen Bereiche resonanter Schwingungen. Allerdings ist diese ortsabhängige Kodierung der spektralen Inhalte für die Vielzahl praktischer Schallsignale nicht ausreichend. Bei Gegenwart von

Hintergrundrauschen wird beispielsweise das charakteristische räumliche Resonanzmuster weitestgehend verdeckt [Kli87]. Ebenso erweist sich für sehr tiefe, aber noch hörbare Frequenzen die Auslenkung der zugehörigen Basilarmembranbereiche als näherungsweise konstant.

Unter dem Begriff „Phase-Locking“ bezeichnet man hingegen die Kopplung der Auslösung von Aktionspotentialen an die Phasenlage der Schallschwingungen [DRS01]. Somit werden im Innenohr die spektralen Informationen zusätzlich zeitlich kodiert. Aus den Pausenlängen zwischen einzelnen oder Gruppen von Aktionspotentialen kann daraus die Frequenz der anregenden Schwingung bestimmt werden. Diese ist umgekehrt proprtional zur Periode des Schallsignals. Roederer [Roe00] erläutert, dass nur durch diesen Mechanismus die Wahrnehmung von Tonhöhen und musikalischen Intervallen möglich ist.

Wie schon in den physiologischen Grundlagen beschrieben (s. Kapitel 4.3), findet in den inneren Haarzellen praktisch eine Halbwellengleichrichtung statt. Die aufsitzenden Stereozilien führen nur bei Auslenkung in *eine* Richtung zur Depolarisation und Freisetzung von Transmittersubstanz. Eine Stimulusauslösung findet bevorzugt im Maximum einer Halbphase, also bei Auslenkung der cochleären Trennwand und der Stereozilien in die stimulierende Richtung, statt.

Die so maximal zeitlich kodierbaren Frequenzen liegen im Bereich zwischen 5000 und 6000 Hz. Das Fehlen des Mechanismus für höhere Frequenzen oberhalb dieses Grenzbereiches wird erklärt mit den elektrischen Kapazitätseigenschaften der Zellmembranen und statistischen Streuungen beim Auslösungsprozess der Aktionspotentiale.

Des Weiteren können einzelne Haarzellen aufgrund von Refraktärzeiten zwischen aufeinderfolgenden neuronalen Impulsen (Spikes) maximal mit einer Rate von 800 Hz der Anregung folgen. Nach dem Salvenprinzip wird aber eine vollständige Reproduktion mittels Gruppen von Haarzellen erreicht. Die Einteilung von benachbarten Populationen in zusammengehörige Einheiten ist im angewandten Modell implizit durch die Aufteilung der Basilarmembran in 251 Sektionen erfüllt.

4.5 Hierarchiemodell

Im zentralen Gehör werden die durch die peripheren Hörorgane erhaltenen Signalinformationen letzendlich einer bewussten Wahrnehmung zugeführt.

Wie bereits in Kapitel 3.2 dargestellt, weisen die Erkenntnisse physiologischer Untersuchungen höherer Verarbeitungsschichten im zentralen Gehör aktuell allerdings noch keine für ein künstliches Modell praktikable Qualität auf. Daher werden für die Nachbildung dieser neuronalen Aktivitäten Ansätze aus der kognitiven Psychologie, hier speziell der Gestaltpsychologie, verwendet.

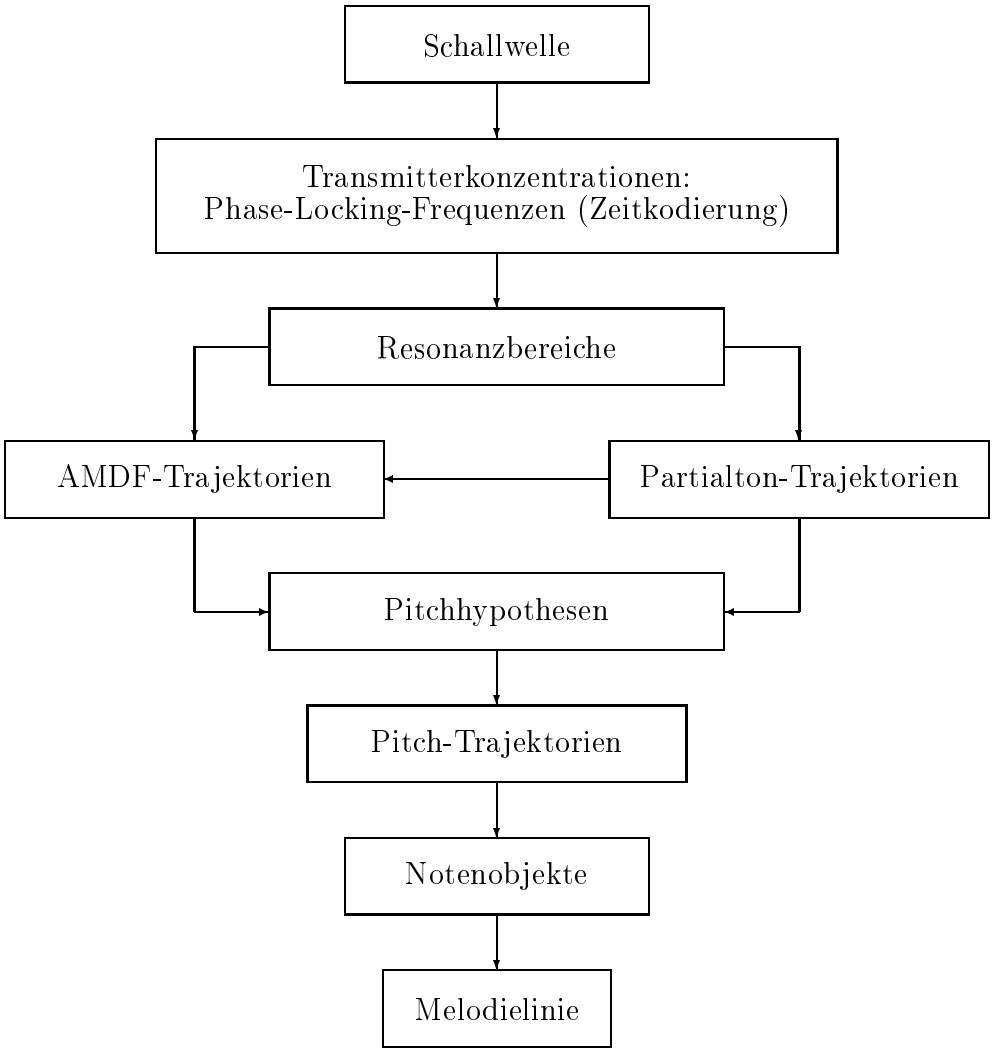

Abbildung 4.3: Hierarchiemodell

Anhand Abbildung 4.3 soll im weiteren das vom Autor aufgestellte Hierarchiemodell erläutert werden. Dieser Ansatz kann natürlich der komplexen

Verabeitungsweise, welche im menschlichen Gehirn vonstatten geht, nur unzureichend gerecht werden. Für eine auf die Perzeption musikalischer Melodien eingeschränkte „Welt“ erweist sich aber diese Darstellung, insbesondere im Hinblick auf die Anwendbarkeit in einem praktikablen System zur automatisierten Melodietranskription, als äußerst zweckmäßig. Wert wird hierbei gelegt auf die durch die gestaltpsychologischen Prinzipien sich ergebende Fusionierung einzelner Elemente in hierarchisch zunehmend abstraktere und bedeutungsgeladenere Objekte.

Ausgehend von der einfallenden Schallwelle wird zunächst im peripheren Gehör, gemäß den physiologischen Grundlagen aus Kapitel 3 und mittels der in den Abschnitten 4.1 - 4.3 vorgestellten Modelle, diese gehörgerecht in neuronale Aktivität auf den Hörnerven transformiert. Aufgrund wahrscheinlichkeitsbelasteter Vorgänge bei der Auslösung von Nervenimpulsen wird die weitere Interpretation auf den Transmittersubstanzen im synaptischen Spalt unmittelbar vor den Nervenfasern basiert. Dies kann gerechtfertigt werden durch die im Gehörmodell verwendeten abschnittsweisen Sektionen, die im Sinne einer statistischen Gesamtheit Populationen von Nervenzellen zusammenfassen.

Über die „Phase-Locking“-Methode (s. Kapitel 5.1.1) können aus den Substanzkonzentrationen innerhalb der inneren Haarzellen die vorhandenen Frequenzen bestimmt werden. Aufgrund der mechanischen Eigenschaften des Cochlea-Systems (Massen, Steifigkeiten, Elastizitäten, etc.) wird bei Resonanzanregung der Basilarmembran nicht nur ein einzelner Punkt in Schwingungen versetzt. Vielmehr entsteht eine glockenförmige Amplitudenhüllkurve, die sich je nach Anregungsstärke über einen ausgedehnten Bereich erstreckt. Im Modell werden daher benachbarte Haarzellen mit „gleichen“ Frequenzen im Sinne der gestaltpsychologischen Fusionierungstendenzen zu Resonanzbereichen zusammengefasst.

Die Wahrnehmung von Tönen ist immer verbunden mit einer je nach Frequenzbereich variierenden minimalen Anzahl von Wiederholungen der zugehörigen Periode, woraus ein zeitlich ausgedehnter Verlauf solcher Frequenzen resultiert. Mit Hilfe bestimmten Kontinuitätskriterien genügenden Eigenschaften, wie zeitlichem und frequenzmäßigem Abstand (s. Abschnitt 5.1.1), werden aus den einzelnen detektierten Frequenzeinträgen sogenannte Partialtontrajektorien gewonnen.

Weitere (für die monophone Transkription aber weniger relevante) Einheiten werden mit den vom Autor als AMDF-Trajektorien bezeichneten Objekten extrahiert. Aufgrund spektraler Verdeckungseffekte können Resonan-

zen von benachbarten Bereichen ausgeprägterer Amplituden teilweise oder ganz überlagert werden. Genauere Betrachtung der Feinstruktur der zeitlichen Amplitudenverläufe der Transmitterkonzentrationen zeigt, dass sich aufgrund der Interferenzen benachbarter Frequenzbeiträge periodische Modulationen finden lassen, die Aussagen über verdeckte spektrale Inhalte ermöglichen (s. Kapitel 5.2.1). Dies entspricht praktisch der Suche nach Autokorrelationsmaxima der Haarzelleninhalte. Die Bezeichnung „AMDF“ erhalten die Trajektorieneinheiten nach dem gewählten Analyseverfahren „Average Magnitude Difference Function“. Die gefundenen Frequenzbahnen werden anschließend einem Längenkriterium unterzogen, d.h. es wird vorausgesetzt, dass Beiträge, die eine gewisse Mindestdauer unterschreiten, keinen wesentlichen Einfluss auf das perzeptierte Hörerlebnis ausüben. Eine Vielzahl sporadischer, irrelevanter Frequenzeinträge kann auf diese Weise bereits eliminiert werden.

Schließlich werden dann aus den übriggebliebenen Frequenzeinträgen zu äquidistanten Zeitpunkten sogenannte „Pitch-Hypothesen“ geformt. Eine Vielzahl psychoakustischer Untersuchungen zeigt, dass harmonische Obertonreihen, wie Sie üblicherweise von gebräuchlichen Musikinstrumenten und auch der menschlichen Gesangsstimme erzeugt werden, zu einer Tonhöhenwahrnehmung führen, die in erster und für den Anwendungsfall hinreichenden Näherung proportional zur Grundfrequenz des zugehörigen Partialtonkomplexes ist. Diese auditive Kenngröße wird in der Regel unter der Bezeichnung „Pitch“ zusammengefasst. Die Hypothesenbildung erweist sich für den polyphonen Fall als wesentlich komplexer und schließt die oben beschriebenen AMDF-Trajektorien mit ein (s. Kapitel 5.2.2).

Analog zur Vorgehensweise bei der Herausbildung von Partialtontrajektorien wird dann bei der Extraktion der „Pitchtrajektorien“ verfahren. Die gleichen gestaltpsychologischen Kriterien werden herangezogen, um die in der vorherigen Stufe erhaltenen Hypothesen zu zeitlich andauernden Pitcheindrücken zusammenzufassen, wobei die Abstandsparameter der veränderten Abstraktionsstufe angepasst sind.

Wie schon in Kapitel 2 („Stand der Technik“) erläutert, existieren eine Reihe von Methoden zur Bestimmung von Frequenz- bzw. Pitchinhalten in Musik- und Sprachsignalen. Die in vielen Verfahren nach Ansicht des Autors stiefmütterlich behandelte Disziplin der Segmentierung stellt einen wesentlichen Baustein eines zuverlässigen Transkriptionssystems dar. Im Hierarchiemodell führt die Anwendung des in Kapitel 5.1.2 beschriebenen Verfahrens zur Einteilung der gefundenen Pitchtrajektorien in einzelne Notenobjekte, die

daran anschließend die Grundelemente möglicher Melodien bilden. Im vorgestellten Algorithmus kommen dabei eine Reihe psychoakustisch motivierter Ansätze zum Einsatz (Onsetdetektion, Onsetfusion, etc.).

Abschließend dienen dann die von Bregman [Bre90] motivierten Erkenntnisse zur Extraktion auditorischer Ströme aus musikalischem Material in der höchsten in diesem Modell erzeugten Abstraktionsschicht zur Herausbildung vorhandener Melodieinhalte. Weitere musiktheoretisch höher angesiedelte Strukturen wie Strophen, Refrain oder ähnliche Einheiten wären untersuchbar, sollen aber nicht Inhalt der Arbeit sein. Eine wesentliche zusätzliche Eigenschaft der polyphonen Anwendung stellt die Möglichkeit zur Analyse und Darstellung gleichzeitig parallel verlaufender Melodien dar.

Kapitel 5

Implementierung

Wie in den vorherigen Kapiteln dargelegt, soll die automatisierte Melodietranskription basiert werden auf den physiologischen und psychologischen Wahrnehmungsmechanismen des Menschen. Gegenstand dieses Kapitels ist nun die praktische Anwendung der beschriebenen Modelle und die nachfolgende Interpretation der resultierenden Ergebnisse.

Nach der Erläuterung der grundlegenden Algorithmen am Beispiel monophoner Musiksignale in Kapitel 5.1, werden in Abschnitt 5.2 die Strategien zur Deutung der komplexen Vorgänge in polyphoner Musik vorgestellt.

Es wird jeweils anhand eines konkreten musikalischen Exempels die Vorgehensweise illustriert. Die gewählten Schallsignale entsprechen repräsentativen Darbietungen gesungener und instrumentaler Eingaben.

5.1 Monophone Melodietranskription

Mit Hilfe der Hauptmelodie aus Sergej Prokofiews „Peter und der Wolf“ sollen die durchgeführten Analyseschritte veranschaulicht werden. Die abstrakte musikalische Notation, ohne Phrasierungs- und Artikulationszeichen, ist in Abbildung 5.1 dargestellt.

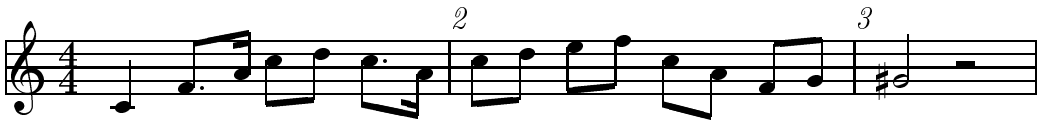

Abbildung 5.1: „Peter und der Wolf“ - Musikalische Notation

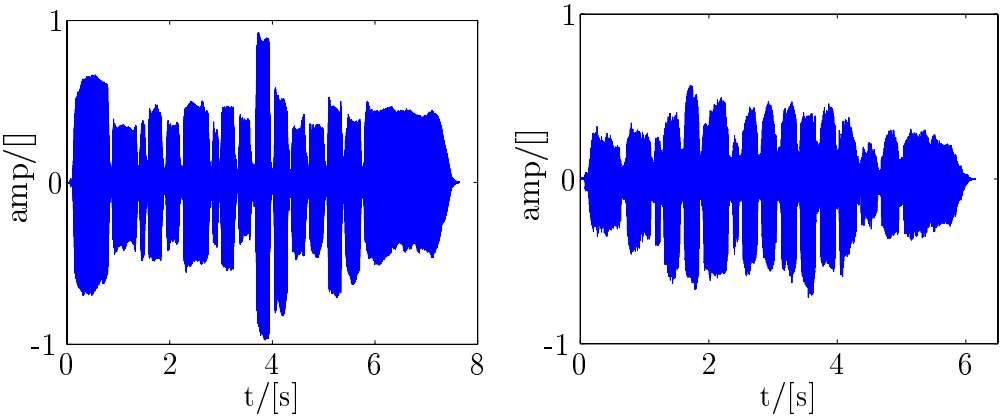

Abbildung 5.2: „Peter und der Wolf“ - Schallwellenform
(links: Klarinette; rechts: Gesang)

Die Schallwellenformen zweier konkreter Eingaben finden sich in Abbildung 5.2. Auf der linken Seite handelt es sich um die Darbietung einer Klarinette, einem Mitglied der Familie der Holzblasinstrumente. Rechterhand ist die von einer Männerstimme eingesungene Melodie zu sehen. Die im weiteren Verlauf illustrierten Ergebnisse beziehen sich, wo nicht anders angegeben, auf diese beiden Eingangssignale.

Die erste Problematik bei der Verarbeitung der Melodiebeispiele ergibt sich bereits beim Einlesen der digitalisierten Wellenformen in den Rechner. Aus der Motivation heraus, das hier vorgestellte Verfahren in multimedialen „Query-By-Humming“-Anwendungen einzusetzen, muss man in der Regel von unkalibrierten Schallsignalen ausgehen, d.h. in technischen Umgebungen, wie sie im Konsumentenbereich anzutreffen sind, liegen keine Informationen über die tatsächlichen Schalldruckpegel vor. Die Empfindlichkeit bzw. die Auflösung des menschlichen Ohres, und somit auch des entwickelten Modellapparates, variieren aber als Funktion der dargebotenen Schallintensitäten. Bezüglich der zu erwartenden durchschnittlichen Pegelverteilungen können somit nur verallgemeinernde Annahmen gemacht werden. In einer Vorverarbeitungsstufe wird das zu untersuchende Schallsignal daher zunächst auf Vollaussteuerung normalisiert. Aufgrund eigener heuristischer Untersuchungen wird der resultierende maximale Pegel für alle Fälle mit 90 dB(SPL) angenommen. Die Benutzung von kalibrierten Signalen würde das nachfolgend beschriebene Verfahren sicherlich verbessern. Auf die Betrachtung solcher Eingaben wurde aber zugunsten der allgemeinen Anwendbarkeit des

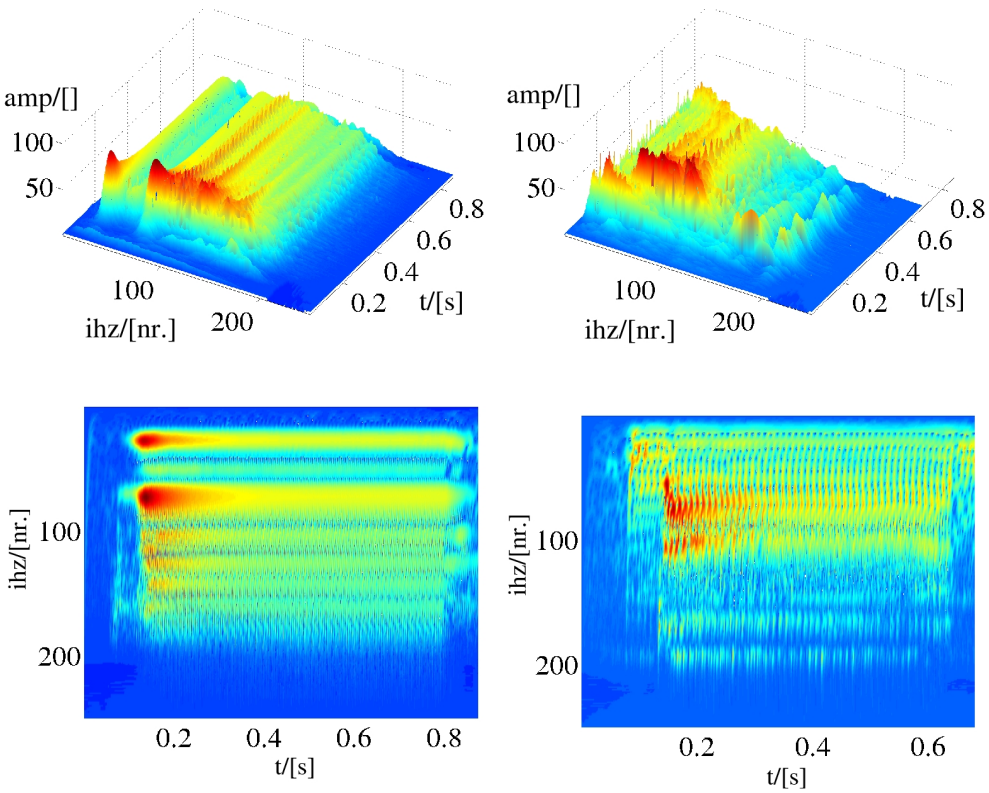

Abbildung 5.3: Transmitterkonzentration der inneren Haarzellen

Transkriptionssystems verzichtet.

Ausgangspunkt für die weiteren Betrachtungen ist das vom Ohrmodell selbst bereitgestellte Signal der physiologisch gehörgerechten Analyse. Abbildung 5.3 zeigt die Hüllkurven der Transmitterkonzentrationen der inneren Haarzellen für den jeweils ersten Ton der Beispielmelodien. Eine gewisse Struktur kann in diesen Bildern erkannt werden, die auf die zu erwartende Obertonanordnung hindeutet. Wie im Folgenden gezeigt wird, ist aber die Untersuchung der Feinstruktur der betrachteten Spaltinhalte notwendig, um eine zuverlässige Bestimmung der Frequenz- bzw. Pitchinhalte zu gewährleisten.

5.1.1 Pitchextraktion

In Kapitel 4.4 wurde bereits erläutert, dass zwei grundlegende Theorien versuchen, die vorkommenden Effekte zu beschreiben [Roe00][HA01]. Im Rahmen eines ortsbasierten Ansatzes wird die Stimulation der Basilarmembran an frequenzcharakteristischen Positionen zur Auswertung vorhandener Teiltonfrequenzen herangezogen. Demnach erhöht sich die neuronale Aktivität an den von den vorliegenden Frequenzkomponenten in Resonanz versetzten Bereichen und führt zu den in Abbildung 5.3 zu erkennenden Teiltonmustern. Eine feste Zuordnung von lokaler Anregungsposition und charakeristischer Frequenz würde die Bestimmung spektraler Anteile erlauben.

Aus Hörexperimenten resultieren allerdings einige über diesen Ansatz nicht erklärbare Effekte. So steht beispielsweise die Feinauflösung der Pitchwahrnehmung in Kontrast zu den Breiten der Frequenzgruppenfilterbank [ZF01]. Demnach entspricht die gerade wahrnehmbare Änderung einer Tonhöhe („Just-Noticable-Difference“ JND) ungefähr einem $\frac{1}{30}$ der Breite einer Frequenzgruppe. Weiterhin zeigt sich das Anregungsmuster auf der Basilarmembran für Frequenzen kleiner 50 Hz als weitgehend konstant.

Diese Effekte können über eine zeitbasierte Theorie unter Betrachtung der angedeuteten Feinstruktur der ortsabhängigen Anregungsfunktion erklärt werden. Wie schon in Kapitel 4.4 erwähnt, bezeichnet man das Feuern von Spikes im Takt der lokalen Signalflanken als „Phase-Locking“. Ein gewisser Nachteil der Zeittheorie besteht in der Einschränkung dieser Funktionalität auf Frequenzen unterhalb von 5 - 6 kHz. Dies steht in Einklang mit der verschlechterten Fähigkeit des Menschen, den Pitch höherfrequenter Eingangssignale eindeutig zu bestimmen.

In der vorliegenden Arbeit wird zur Extraktion von Pitchwerten ausschließlich auf den zeitlichen Ansatz zurückgegriffen, da in der musikalischen Aufführungspraxis in der Regel keine Grundtonfrequenzen oberhalb von 5 kHz verwendet werden. Allerdings führt die Vernachlässigung ortsspezifischer Frequenzerkennungsstrategien zum Verlust einiger relevanter Obertöne bei hohen Grundfrequenzen. Eine Vereinigung beider Modelle im Sinne einer Steigerung der Robustheit des Gesamtsystems soll in weiterführenden Arbeiten implementiert werden.

Zur Veranschaulichung des weiteren Vorgehens zur Pitchbestimmung sind in Abbildung 5.4 die Hüllkurven der charakteristischen, also mit maximaler Amplitude angeregten, Haarzelleninhalte zu sehen. Von oben nach unten handelt es sich hierbei um die den ersten 3 Partialtönen entsprechenden re-

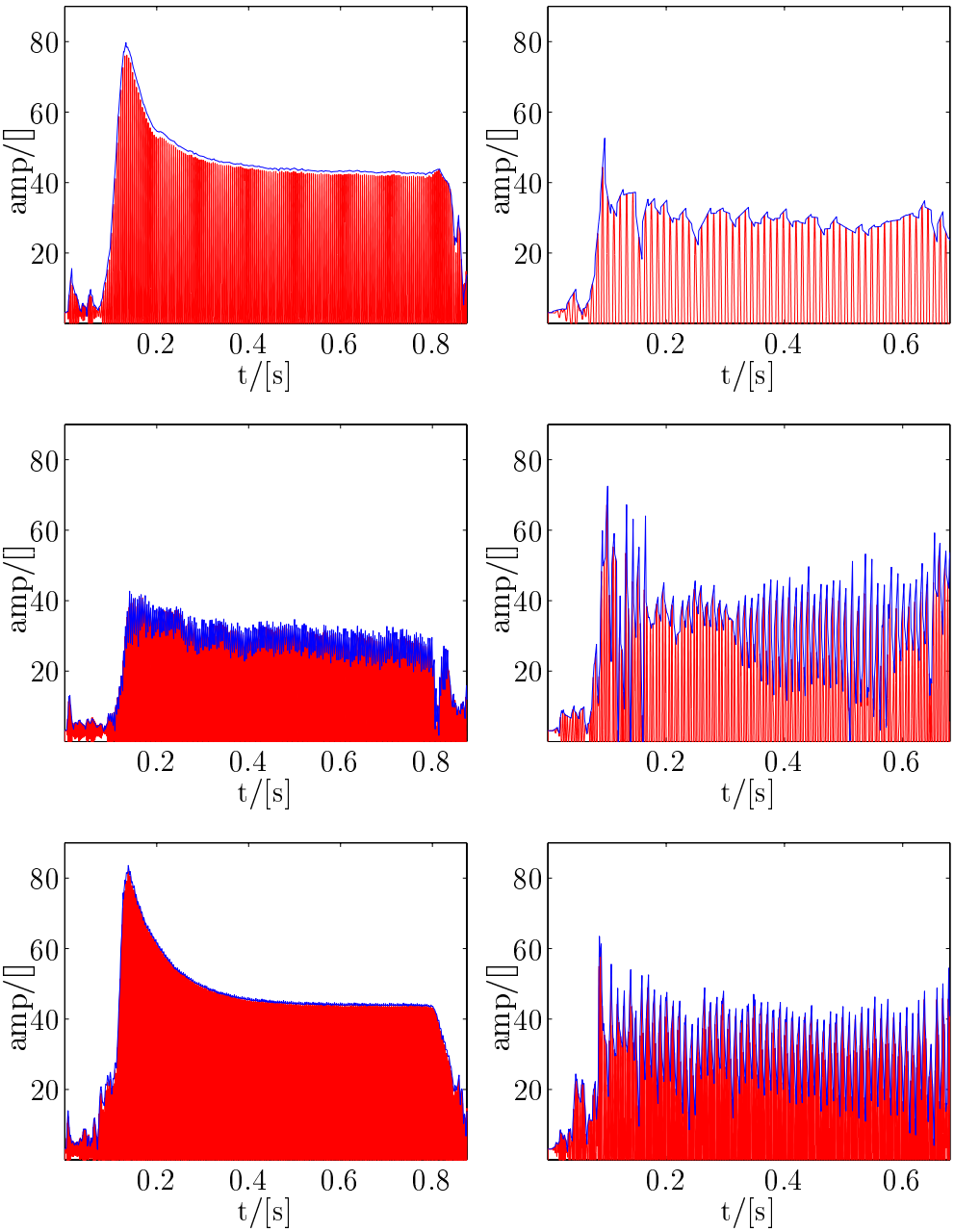

Abbildung 5.4: Transmitterkonzentration charakteristischer Haarzellen

sonanten Verläufe. Die Grafiken der linken Spalte entsprechen wieder dem Klarinettensignal; dargestellt sind die Haarzellen 29, 49 und 76. Die rechte Spalte repräsentiert die Gesangsstimme, vertreten durch die Haarzellen 12, 25 und 40. Wie schon in Abbildung 5.3 zu erkennen, zeigt sich die Instrumentendarstellung wesentlich gleichmäßiger als der menschliche, durch diverse, unter anderem phonemisch bedingte, Instationaritäten während des Verlaufes beeinflusste, Gegenpart. Dies ist charakteristisch für die Mehrzahl solcher Eingaben und erfordert insbesondere bei der späteren Segmentierung besondere Methoden.

In detaillierter zeitlicher Auflösung sind die Transmitterkonzentrationen für den jeweils ersten Partialton in Abbildung 5.5 dargestellt. Hier sieht man deutlich die äquidistanten Abstände der Maxima-Verläufe, die im weiteren zur Bestimmung der Teiltonfrequenzen herangezogen werden. Der Abstand zweier Maxima entspricht der Periode der zugehörigen Frequenz und ist im Bild mit T_i bezeichnet.

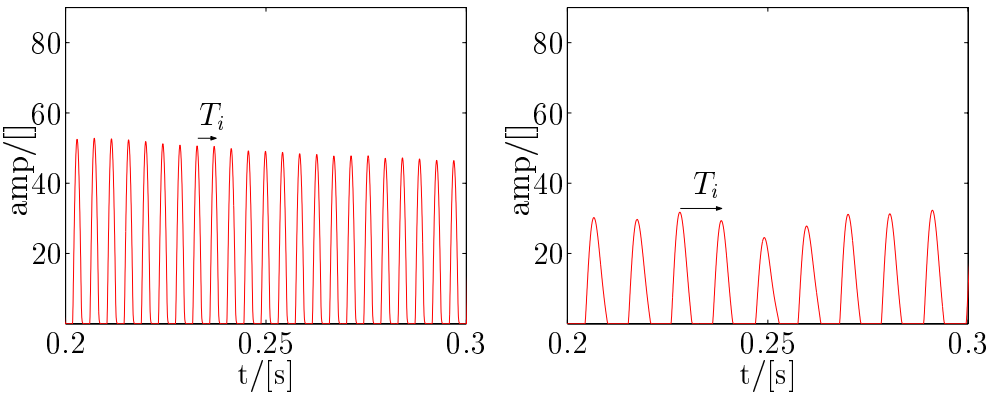

Abbildung 5.5: Transmitterkonzentration in Detail-Darstellung

Nach dem von Meddis [MO98] vorgeschlagenem Verfahren wird die Zuordnung einer Pitchfrequenz über die Methode der „Summen-Autokorrelation“ vorgenommen. Dabei werden eine Anzahl von Inter-Maxima-Abständen in einem Histogramm aufgetragen. Die Benutzung nur einer Periodendauer ist nicht ausreichend. Dies ist notwendig aus zwei verschiedenen Beweggründen. Zunächst ist der Abstand der Spaltmaxima wegen rundungstechnischer Ungenauigkeiten und nichtdeterminiertem Eingangssignal nicht immer „exakt“ gleich der anregenden Periode. Daher empfiehlt es sich, eine mittlere Periodenlänge über einen längeren Zeitraum zu mitteln. Auch bildet der Mensch

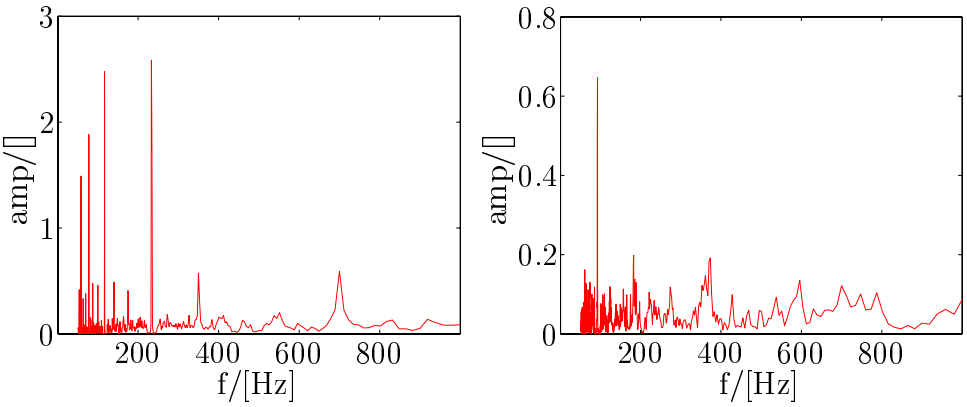

Abbildung 5.6: Histogramm Summen-Autokorrelations-Funktion

seine Hypothese einer Grundtonhöhe aufgrund der statistischen Gesamtheit der Abstände über einen bestimmten Zeitabschnitt. Der zweite Grund für die Benutzung einer Schar auch nicht direkt benachbarter Maxima-Abstände liegt in der Tatsache, dass die Obertöne, entsprechend ihrer Position in der Partialtonreihe, pro Fundamentalperiode mehrere Schwingungen ausführen und somit auch mehrere Transmittermaxima erzeugen. Im Gegensatz zu einer Autokorrelationsbetrachtung, die nur direkte Nachbarn auswertet, erhält man also Einträge für Vielfache und Teiler der Grundperiodendauer. Neben dem zusätzlichen Informationsgewinn durch die gefundenen Obertonintervalle gilt es anschließend, diese von dem tatsächlichen Grundtoneintrag zu trennen. In der Summe addieren sich aber in der Regel die der Grundfrequenz entsprechenden Abstände stärker und sollten in einem prägnanten Maximum im Histogramm münden.

In der beschriebenen Implementierung werden für 6 aufeinanderfolgende Maxima jeweils die Abstände zu den nächsten 10 Nachbarn ausgewertet.

Typischerweise ergeben sich für jeden Analysezeitpunkt Histogrammdarstellungen wie in Abbildung 5.6. Für die eingesungene Variante auf der rechten Seite erhält man *einen* stark ausgeprägten Spitzenwert, der auch die gesuchte Pitchfrequenz von etwas unter 100 Hz repräsentiert.

Komplizierter sieht die Auswertung in der Klarinettengrafik aus. Eine Reihe von Maxima stellen Kandidaten potentieller Fundamentalperioden dar. Mittels zweier Nebenbedingungen lässt sich die Problematik aber in der überwiegenden Mehrzahl der Fälle bewältigen. Zunächst sucht man in der Gruppe möglicher Einträge, die einen bestimmten Schwellenwert nicht un-

terschreiten, nach harmonischen Beziehungen, wie sie aus den oben beschriebenen verschiedenen Konstellationen resultieren. Daran anschließend vergleicht man innerhalb der harmonisch relevantesten Gruppe die relative Ausgeprägtheit der gefundenen Mitglieder. Unterschreiten alle frequenzmäßig höher gelegenen Peaks, den Wert eines gültigen Maximums um einen relativen Grenzwert, so erhält dieser den Zuschlag und wird als aktuelle Pitchfrequenz angenommen.

Nachdem so die für einen bestimmten Zeitpunkt wahrscheinlichste Pitchfrequenz gefunden worden ist, wird danach versucht, die Informationen über vorhandene Partialtoneinträge aus der Gesamtheit der Haarzelleneinträge zu gewinnen. Als Nebenprodukt der Berechnung des Summen-Autokorrelations-Histogramms kann eine solche Darstellung als Einzelhistogramm für jede Haarzelle aufgebaut werden. Ähnliche Auswertungen der Spitzenwerte wie im übergeordneten Fall ordnen, unter Einbeziehung der analysierten Grundfrequenz, jeder Haarzelle eine lokale Schwingungsfrequenz zu. Danach werden anschließend Frequenzen bis zum siebten Partialton, die sich nicht mehr als einen Viertelton (Faktor $\leq$ 1,029) von der zur Pitchfrequenz idealen Harmonischen unterscheiden, verwertet. Jedem Partialton wird eine Wertigkeit zugeordnet, die sich aus der Summe der Amplituden der einzelnen Haarzellen mit teiltontauglichen Frequenzen aufaddiert. Die Beschränkung auf eine 7 Elemente umfassende Obertonreihe steht in Einklang mit den Breiten der Frequenzgruppen, wonach maximal 6 - 7 Teiltöne aufgelöst werden können, bevor diese innerhalb eines kritischen Bandes verschmiert werden.

Die bis hier beschriebene Zuordnung von Pitchwahrnehmungen und zugehörigen Obertonreihen, inklusive einer amplitudenbestimmten Wertigkeit, erfolgt im Zeittakt von 1 ms ($= \frac{1}{1000}$ s). Mit der gewählten Berechnungsrate kann sicher die maximale zeitliche Auflösung des Menschen, die bei höchstens 3 - 5 ms liegt, nachmodelliert werden.

Die aus der Pitchextraktion erhaltenen unbearbeiteten Zeitverläufe der Grundtonfrequenzen sind in Abbildung 5.7 (oben) dargestellt. Die grundsätzlichen Entwickungslinien der beiden Beispielmelodien sind in der Grafik bereits zu erkennen. Allerdings finden sich noch eine Reihe von Störungen insbesondere in den Grenzregionen zwischen zwei Noten.

5.1.2 Segmentierung

Die Aufgabe der nächsten, in diesem Kapitel vorgestellten, Verarbeitungsstufe ist es, die rohen, noch nicht weiter interpretierten Frequenzverläufe in

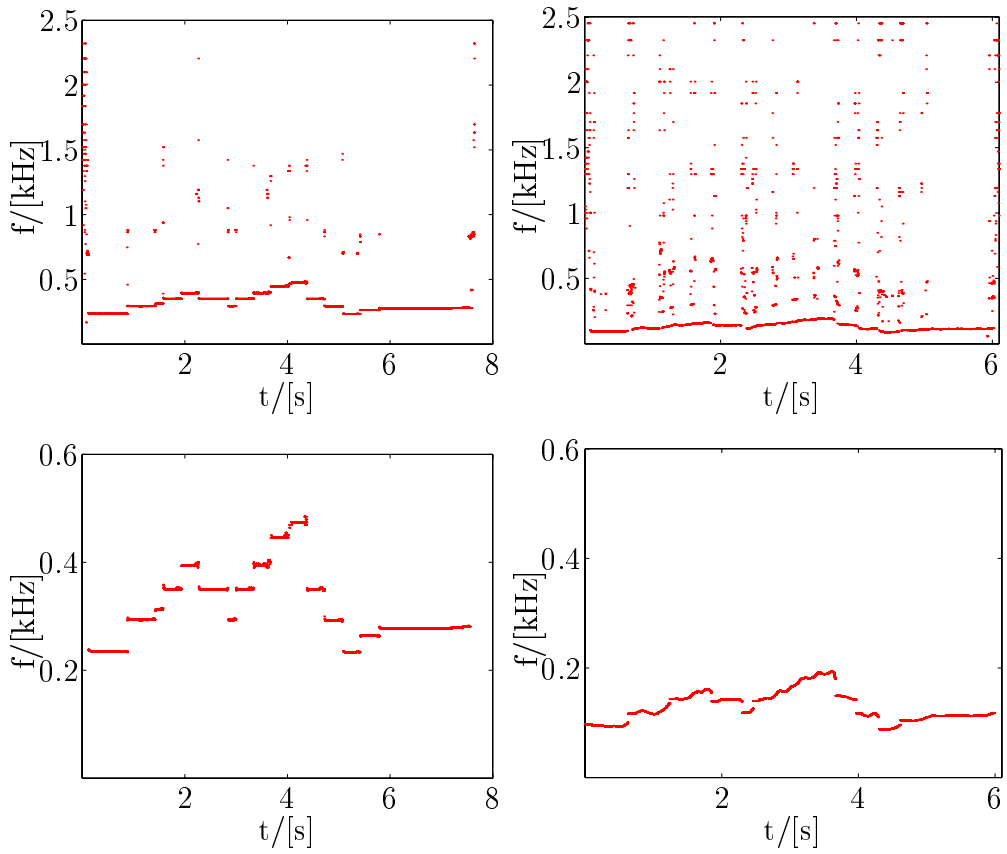

Abbildung 5.7: Pitchverläufe

einzelne Notenobjekte zu unterteilen, wie dies für eine Darstellung in musikalischer Schreibweise bzw. für die in Kapitel 6 erläuterte Datenbanksuche erforderlich ist. Wie schon vorher erwähnt, trennt sich gerade an der Fortgeschrittenheit der Lösung dieser Problematik die Qualität automatisierter Transkriptionssysteme. Während viele Implementierungen, gehörgerecht oder rein technisch, eine recht zuverlässige Pitchmitschrift erstellen können, ist eine Vernachlässigung des, nach Ansicht des Autors, wichtigeren und auch schwierigeren Schrittes der Segmentierung häufig zu beobachten.

In dieser Arbeit wird eine Anzahl verschiedener Ansätze kombiniert, um die Segmentierungsaufgabe zu lösen. Dabei wird wiederum Wert darauf gelegt, in den umgesetzten Verfahren möglichst die Strategien der menschlichen Signalverarbeitung und -interpretation zu berücksichtigen. Eine Anzahl psychoakustischer Erkenntnisse findet Berücksichtigung in der Wahl der Mittel.

In einer ersten Stufe werden aus den zu den Analysezeitpunkten vorliegenden Einzelfrequenzen sogenannte Pitchtrajektorien geformt. Entsprechend der in den Gestaltgesetzen (s. Kapitel 3.2) postulierten Kontinuitätskriterien werden benachbarte Pitcheinträge auf ihre gruppierenden Eigenschaften hin untersucht.

Im Einzelnen bedeutet dies, dass zunächst das relative Verhältnis zweier unmittelbar *nachfolgender* Einträge das Intervall eines Vierteltons (Faktor $\leq$ 1,029) nicht überschreiten darf. Eine Sequenz von Werten, die diese erste Gruppierungsbedingung erfüllen, wird in „Subtrajektorien“ zusammengefasst. Überschreitet die Ausdehnung solcher unterbrechungsfreier Einheiten eine Mindestlänge von minimal 10 ms, so werden diese in einer Liste gültiger Subtrajektorien für die weiteren Verarbeitungsschritte gespeichert.

Nachfolgend wird dann die Nähe zeitlich angrenzender Subtrajektorien auf ihre perzeptive Fusionierungsfähigkeit hin überprüft. Die drei Bedingungen zur Zusammenfassung von Subtrajektorien lauten:

1. Der Abstand darf einen Schwellenwert von 75 ms nicht überschreiten.
2. Ein relativer, abstandsabhängiger Frequenzunterschied von maximal 1,2 muss eingehalten werden.
3. Die Anregungsposition längs der Basilarmembran darf sich nicht um mehr als 20 Haarzellen unterscheiden.

Anschließend werden die resultierenden Trajektorien wieder einem Längenkriterium unterzogen. Das Mindestmaß für einen gültigen globalen Pitchverlauf beträgt 40 ms und liegt damit in der Größenordnung sehr kleiner vorkommender Notenlängen.

Die derart von rauschartigen Störungen „gereinigte“ Pitchdarstellung findet sich in Abbildung 5.7 (unten). Praktisch alle irrelevanten Anteile konnten somit durch die Methode der Trajektoriensuche eliminiert werden.

Zur Vorbereitung der weiteren Segmentierungsschritte werden dann die bis zu diesem Zeitpunkt extrahierten Daten durch verschiedene Manipulationen in eine günstige Form gebracht.

Zunächst werden die Datenlücken innerhalb einer Trajektorie, die durch Fusionierung nicht unmittelbar zusammenhängender Subtrajetorien entstanden sind, interpoliert. Aus den Lückenrandwerten werden über lineare Interpolation die Informationen bezüglich der Anregungsposition der fehlenden

Zwischenräume gewonnen. Aus den dort vorkommenden Einträgen wird dann jeweils die zugehörige Amplitude bestimmt. Die zu ergänzenden Frequenzen berechnen sich aus einer Kombination von Umgebungswerten und einer Neueinschätzung der vorher aufgebauten lokalen Summen-Autokorrelations-Histogramme. Diese, wie auch die weiteren Vorverarbeitungsalgorithmen werden für den Grundton sowie die Schar der betrachteten Obertöne vollzogen. Die beiden folgenden Schritte der Datenaufbereitung beziehen sich auf die den Trajektorien jeweils unmittelbar zeitlich vorherigen Bereiche. Während der Einschwingphasen von natürlichen Schallquellen kommt es zu charakteristischen Verläufen der einzelnen Partialtöne. Hierin finden sich unter anderem für die Klangfarben- bzw. Instrumentenerkennung des Menschen wichtige Zuordnungsindizien. Im transienten Startvorgang stark ausgeprägte Obertöne und auch kurzzeitig vorhandene Subharmonische können zu einer Reihe von Vertauschungen bei der lokalen Auswertung der Korrelation führen. Entsprechen die hier erhaltenen Pitchwerte (Sub-)Oktaven der Starteinträge der Trajektoriengrundfrequenzen, so werden diese Vorläuferbereiche für die komplette Obertonstruktur harmonisch auf die Nachfolgerwerte korrigiert.

In einem abschließenden Extrapolationsschritt wird dann die Trajektorienausdehnung grundsätzlich um 50 ms zeitlich nach vorn erweitert, um mögliche inharmonische Amplitudenschwankungen zu erfassen. Beispielsweise ensteht beim Anblasvorgang von Querflöten ein stark ausgeprägter rauschartiger Luftstrom, der deutliche perzeptive Hinweise auf den Beginn der gespielten Note erzeugt. Solche Amplitudenverläufe dienen, wie später gezeigt wird, dem Menschen auch dazu, die wahrgenommenen Pitchverläufe in Unterabschnitte zu zerlegen.

Nach dieser bisher hauptsächlich auf Frequenzinformationen basierten Aufbereitung der aus dem physiologischen Ohrmodell gewonnenen Daten werden die gefundenen Trajektorien im nächsten Schritt einer amplitudenorientierten Analyse unterzogen.

Besondere Beachtung findet hierbei die Berücksichtigung von Informationen über den gesamten Bereich der Obertonreihe. Dies steht im Einklang mit der Argumentation von Scheirer [Sch98], der eine Betrachtung der Gesamthüllkurve als Verschmierung der vorhandenen Informationen ausweist. Er stellt die Hypothese auf, dass das auditorische System zur Segmentierung eine spezielle Kreuzband-Integration über die Frequenzbänder ausführt.

Die Einführung eines absoluten Schwellenwertes, dessen Größe sich aus der internen Kalibrierung des inneren Haarzellenmodells ergibt, führt zur

Entfernung von Trajektorien, deren Amplituden in ihrem gesamten Verlauf die Größenordnung der Ruhekonzentrationen der Haarzellen nicht überschreiten. Wie schon in Kapitel 4.3 erläutert, weisen die inneren Haarzellen auch ohne äußere Anregung eine gewisse spontane Aktivität auf, die implizit mit einer bestimmten Konzentration von Transmittersubstanz gekoppelt ist.

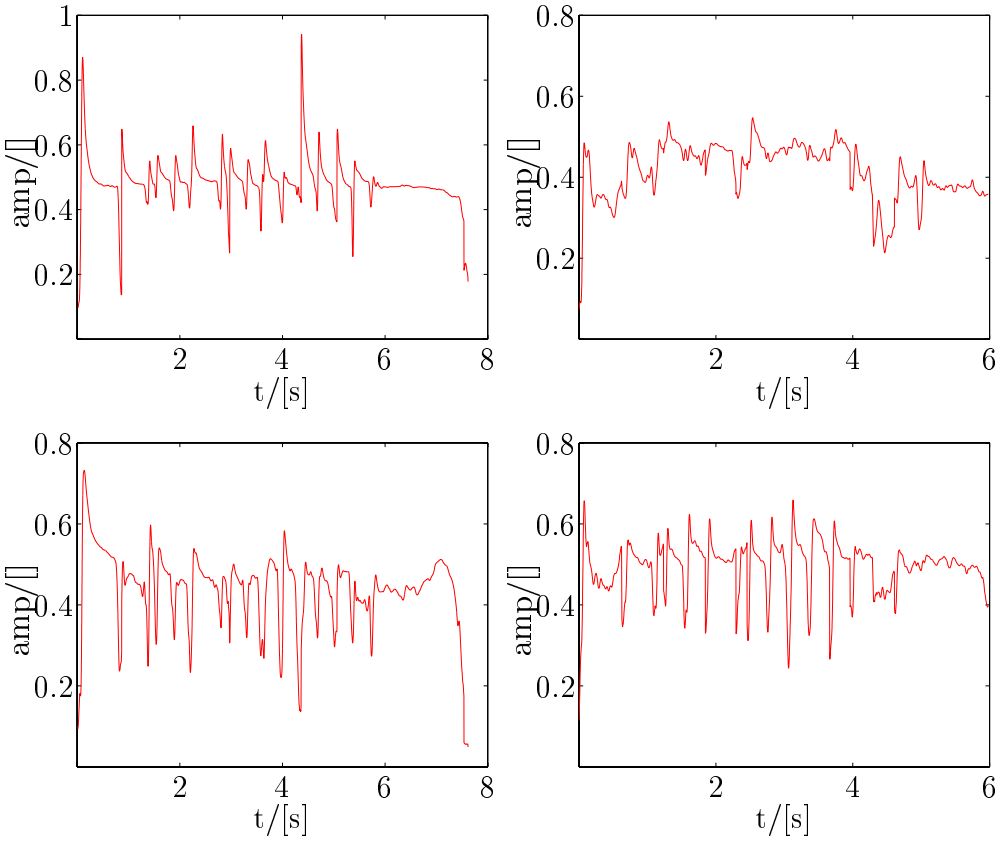

Abbildung 5.8: Hüllkurven der Partialtöne 1 und 4

Nachfolgend werden dann die Amplitudenhüllkurven der Partialtöne in Fenstern der Breite von 15 ms geglättet. Zweimalige Durchführung dieser Funktionalität führt praktisch zu tiefpassgefilterten Signalen, die von irrelevanten Spitzenausschlägen befreit sind. Die Beibehaltung dieser nur sehr kurzen Amplitudenmaxima würde der integrierenden Arbeitsweise der menschlichen Lautheitswahrnehmung widersprechen und in den weiteren Segmentierungsschritten zu einer häufigen Unterteilung der vorhandenen Noten in kleinere Einheiten führen.

In Abbildung 5.8 sind die Hüllkurvenverläufe der Transmitterwertigkeiten von Grundton und drittem Oberton dargestellt. Der zeitliche Verlauf der anderen Partialtöne zeigt sich je nach Ausgeprägtheit der Amplituden qualitativ ähnlich. Aufgrund bautechnisch bedingter Gegebenheiten besitzt beispielsweise der zweite Partialton der Klarinette nur eine geringe Stärke. Spektrale Verdeckungseffekte der ungeraden Teiltöne schränken die erzeugten Wertigkeiten weiter ein und führen in diesem Fall zu einem geringen Einfluss bzw. einem flachen Verlauf der zweiten Harmonischen.

Die Interpretation dieser Darstellungen wird im Folgenden als Grundlage zur weiteren Segmentierung dienen. Prinzipiell wird in den bereitgestellten Signalen nach steilen Wertigkeitsflanken gesucht, die in ihrer Kombination die Wahrnehmung von Notenanfängen verursachen können. Grundlegende Annahme ist, dass der Mensch immer dann, wenn sich in seiner sensorischen Umgebung etwas ändert, den Beginn einer neuen Aktivität, in diesem Fall eine neue Note, mutmaßt. Unterstützt wird diese These durch die Wirkungsweise der inneren Haarzellen, die bei neu einsetzender Anregungsaktivität, einem Alarmsystem gleich, eine besonders starke Reaktion zeigen, um anschließend auf ein stationäres Aktionsmaß zu adaptieren.

Praktisch wird dieser Ansatz umgesetzt durch die Einführung einer sogenannten „Onset-Map“. Hierunter versteht man die Darstellung ausgeprägter aufsteigender Wertigkeitsflanken für die Schar der Partialtöne über die Zeit. Abbildung 5.9 zeigt für die gewählten Melodien die zugehörigen Ergebnisse in perspektivischer Ansicht (oben) und Draufsicht (unten). Man erkennt, gerade in der unteren Bildzeile eine Struktur, die die rhythmischen Inhalte der Beispiele wiedergibt. Der zur Ermittlung dieser Darstellungen ausgeführte Algorithmus stellt sich wie folgt dar:

Analog zur Vorgehensweise bei der Berechnung von Trajektorien wird zunächst in den Amplitudenverläufen der einzelnen Partialtöne nach sogenannten „Subflanken“ gesucht. Darunter werden Bereiche verstanden, deren streng monotone Anstiege gewissen Kriterien genügen. Konkret muss die gemessene Steigung einen bestimmten Grenzwert überschreiten, sowie die erzeugte relative Amplitudenänderung bei mindestens 5 % liegen.

Anschließend werden dann wieder Kontinuitätskriterien aufgestellt, bei deren Erfüllung benachbarte Subflanken zu globalen Wertigkeitsflanken fusionieren. Im Einzelnen müssen dabei folgende Nebenbedingungen erfüllt werden:

1. Subflanken dürfen nicht weiter als 10 ms voneinander entfernt sein.

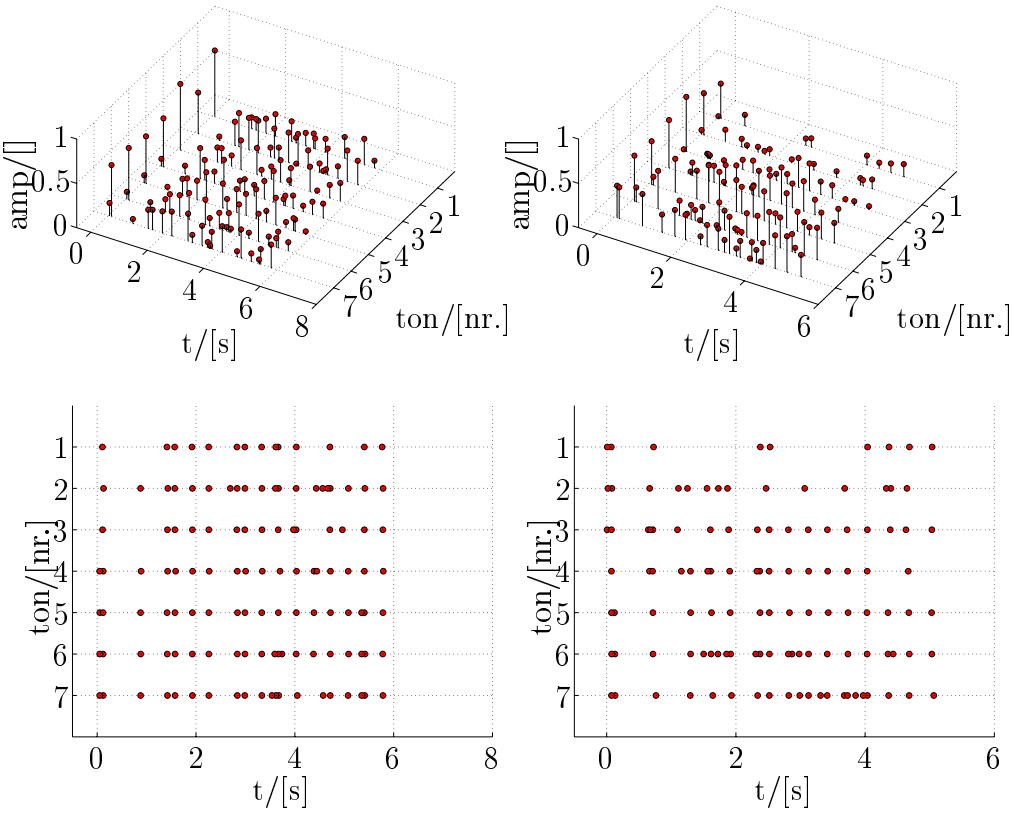

Abbildung 5.9: Onset-Maps

2. Die Startwertigkeiten der nachfolgenden Subflanke müssen mindestens bei 95 % der Vorgängerflanke liegen.

3. Die fusionierte Flanke genügt einer globalen Mindeststeigung und

4. besitzt eine relative Amplitudenänderung von minimal 7,5 %.

Sind die aufgezählten Voraussetzungen erfüllt, so werden die betreffenden Flanken der Onset-Map hinzugefügt (s. Abbildung 5.9).

Gerade bei gesungenen Passagen tritt aber hierbei eine zusätzliche Problematik auf. Durch die Aneinanderreihung von Phonemen, innerhalb von Einzeltönen bildenden Silben, kommt es zu starken Verschiebungen der spektralen Verteilung während einzelner Noten. Dies lässt sich vorteilhaft illustrieren anhand des gewählten Gesangbeispiels, welches auf die Silben „Na-Na-Na“ eingesungen wurde. In Abbildung 5.3 (rechte Seite) erkennt man, dass

der Konsonant „N“ als dentaler Nasallaut [Bün93b] zu Beginn über einige 10 ms seinen spektralen Schwerpunkt zu niedrigen Haarzellennummern, also tiefen Frequenzen, hin ausprägt. Hohe Frequenzen werden kaum angeregt. Bei Einsatz des langen Vokals „A“ zeigen dann insbesondere die neu angeregten höherfrequenten Bereiche ein starkes Aktivitätsmuster. Dies kann zu massiven Wertigkeitsflanken und damit verbundenen Einträgen in der Onset-Map führen, obwohl perzeptiv nur ein Notenanfang resultiert. Solche Phonemwechsel können bis zu einigen 100 ms betragen ohne dabei als zwei eigenständige Notensegmente wahrgenommen zu werden.

In Anlehnung an psychoakustische Erkenntnisse über Vor- und Nachverdeckungseffekte [ZF01] wird daher eine Strategie zur Fusionierung nicht relevanter Onsets eingeführt, die anhand Abbildung 5.10 näher erläutert werden soll. Zunächst werden grundsätzlich für jeden Partialtonverlauf alle globa-

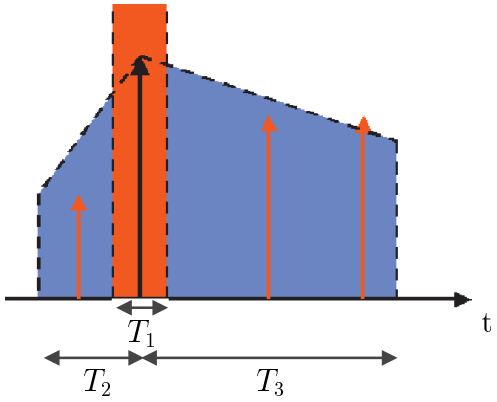

Abbildung 5.10: Onset-Fusion

len Flanken, die sich innerhalb eines Zeitintervalls von T_1 ($\pm$ 75 ms, roter Bereich) befinden, zu Gunsten der Höherwertigen zusammengefasst. Außerdem werden gleitende Schwellwerte benutzt, um weiter entfernte Flanken bezüglich ihrer Bedeutung zu beurteilen. Im Falle der Vorverdeckung beträgt das betrachtete Zeitintervall T_2 200 ms. Alle Onset-Wertigkeiten, die unter der in Abbildung 5.10 gestrichelten Linie liegen (linke blaue Fläche), werden verworfen. Die gleiche Argumentation trifft zu für die auf rechten Seite dargestellte Nachverdeckung, nur dass hier die vorgegebene Zeit T_3 mit 400 ms angenommen wird. In Abbildung 5.10 würde neben dem schwarz dargestellten Onset nur der ganz rechts befindliche rote Beitrag Berücksichtigung in der Onset-Map finden.

Das pinzipiell gleiche Vorgehen wie bei der Erkennung von Noteneinsätzen wird zur Detektion von Noten-Offsets gewählt. Die ähnliche Art und Weise des Aufbaus der zugehörigen Offset-Map soll hier nicht weiter erläutert werden. Zu erwähnen wären allerdings die etwas unterschiedlich zu wählenden Schwellenwerte. Aufgrund übertragungstechnischer Einflüsse, wie beispielsweise der vorhandenen Raumakustik, sind die transienten Ausschwingvorgänge in der Regel um einiges verschmierter als die Einschwingverläufe. Daher können hier nur extrem deutlich ausgeprägte Amplitudeneinbrüche zur Begrenzung globaler Trajektorien herangezogen werden.

Nachdem in der beschriebenen Form die Onset-Map aufgebaut worden ist, werden im Anschluss daran die den einzelnen Partialtönen zugeordneten lokalen Wertigkeitsflanken im sogenannten „Onset-Histogramm“ zusammengefasst, um aus einer Häufung ausgeprägter Onset-Einträge auf global wahrgenommene Notenanfänge zu schließen.

Dazu wird in überlappenden Fenstern einer Breite von 75 ms über die Onset-Map gescannt. Der gewählte Zeitabstand zwischen zwei Scan-Vorgängen beträgt 10 ms und resultiert somit in einer Überlappung von ungefähr 87 %. Die gewählte Zeitauflösung soll die Größenordung der integrierenden menschlichen Verarbeitung berücksichtigen, dabei aber eine möglichst exakte Positionierung der gefundenen Onsets gewährleisten. Je exakter die Startpunkte der einzelnen Noten analysiert werden können, desto besser kann das in Kapitel 6.1 beschriebene „Query-By-Humming“-System die rhythmischen Eigenschaften der untersuchten Melodien mit den Vorgaben in der Datenbank abgleichen.

Eine Reihe von Nebenbedingungen, von denen die wichtigsten im Folgenden aufgeführt werden, soll dabei sporadische, irrelevante Onset-Map-Einträge von perzeptiv gültigen Beiträgen trennen.

Zum Aufbau des Onset-Histogramms werden die in den einzelnen Fenstern gefundenen Teiltonwertigkeiten aufaddiert. Existieren pro Partialton mehrere Onsets, so wird nur der jeweils stärkere berücksichtigt. Der Summenwert wird allerdings nur dann in das Histogramm übernommen, wenn sich unter den Einzelbeiträgen mindestens einer der ersten 3 Partialtöne befindet. Andernfalls resultiert ein wertneutraler Eintrag.

Nach Aufbau des Onset-Histogramms über den kompletten Zeitverlauf erfolgt anschließend eine zweifache Glättung der Werte. Dabei wird pro Fenster unter Berücksichtigung der jeweils in beiden Richtungen nächsten 3 Nachbarwerten ein gewichteter Mittelwert berechnet. Motivation hierfür ist die hiermit implizit einhergehende weitere Eliminierung rauschhafter Onsetein-

träge zugunsten klar ausgeprägter Histogrammmaxima.

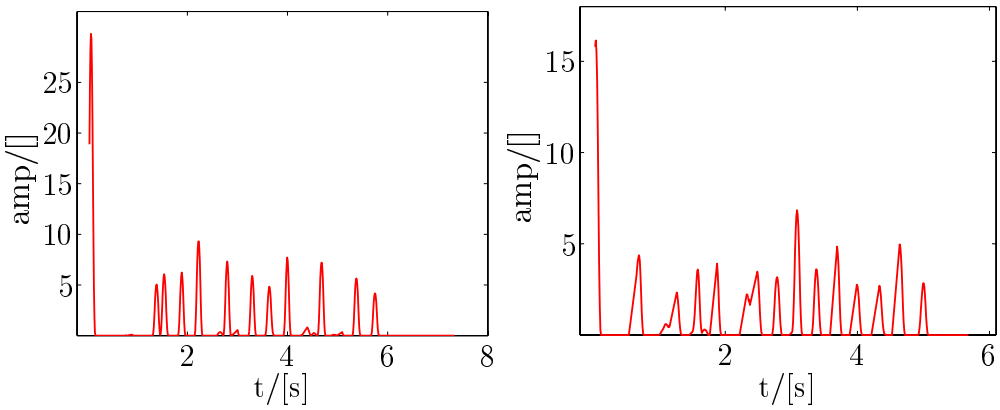

Abbildung 5.11: Onset-Histogramm

Die für die Beispielmelodien erhaltenen Darstellungsformen finden sich in Abbildung 5.11. Die beiden Grafiken können als repräsentative Vertreteter vorkommender Verläufe verstanden werden.

Der vorletzte Schritt bezüglich der amplitudenorientierten Segmentierung besteht schließlich in der Interpretation des derart vorbereiteten Onset-Histogramms. Durch die Anzahl der benutzten, aufwändigen Vorverarbeitungsschritte beschränkt sich die Beurteilung der erzielten Histogrammwerte allerdings auf ein einzelnes Kriterium. Übertreffen die Wertigkeitsmaxima einen empirisch festgelegten Grenzwert, so werden diese als gültige Notenanfänge gespeichert.

Den Anfängen globaler Trajektorien, denen über den Segmentierungsalgorithmus keine gültigen Flanken zugeordnet werden können, wird a-priori natürlich ein Notenstart zugesprochen.

Ein allerletzter „nachverarbeitender“ Prozess, der analog der oben beschriebenen Onset-Fusion funktioniert, versucht abschließend nochmals nah benachbarte Notenstartzeitpunkte mit zeitlich vor- bzw. nachverdeckender Wirkung zusammenzufassen.

In der Regel können die meisten Notenanfänge mittels der beschriebenen Segmentierung unter direkter Benutzung von Amplitudenwertigkeiten aus den Transmitterkonzentrationen der inneren Haarzellen bzw. der analysierten Partialtöne extrahiert werden. Mitunter können aber, gerade im Bereich stimmlich dargebotener Melodien, durch uneindeutig artikulierte Noteneinsätze nicht alle Onsets über die Amplitudensegmentierung erfasst wer-

den. In Abbildung 5.11 fällt beispielsweise der zweite Ton der Klarinettenmelodie im Onsethistogramm völlig aus. In diesem wie in verwandten Fällen greift aber eine weitere Segmentierungsstufe. Bei der sogenannten „Pitch-Segmentierung“ werden die Verläufe der globalen Trajektorien in Bezug auf ihre frequenzmäßige Entwicklung hin untersucht. Erneut finden sich eine auf hierarchischen Strukturen begründete Vorgehensweise.

Wie schon bei der Herausbildung von Pitchtrajektorien und der Suche nach Wertigkeitsflanken im Amplitudenverlauf beginnt auch die Pitch-Segmentierung mit der Einteilung der vorhandenen Daten in Subsegmente. In diesem Teil der Analyse wird allerdings nach konstanten Abschnitten bezogen auf die Frequenzen der vorherigen Segmentierungsergebnisse gesucht. Voraussetzung für diese „Pitch-Subsegmente“ sind eine Mindestlänge von 10 ms und eine relative Schwankungsbreite von $\Delta f_{max}/\Delta f_{min} \leq 1{,}029$ (Faktor Viertelton). Zur Bestimmung dieser Bereiche wird mit einem 1 Viertelton breiten Fenster vertikal durch die Zeit-Frequenzdarstellung gescannt. Gültige Abschnitte werden nach Zuordnung einer mittleren Frequenz in einem Segmentvektor gespeichert.

Die Aufgabe des nachfolgenden Fusionierungsschrittes ist es, überlappende Bereiche, deren Frequenzabstand weniger als ein Viertelton beträgt, zusammenzufassen. Die entstandenen Verläufe müssen als Gültigkeitsmaß eine Mindestlänge von 25 ms aufweisen. Ihnen wird abermals eine aus den Unterabschnitten gewichtete gemittelte Frequenz zugeordnet.

Ein zweites Fusionierungskriterium wertet wiederum Abstände bezüglich der Durchschnittsfrequenzen und der Zeiten aus. Die maximal zulässigen Abstände betragen für die Zeit 100 ms und für die relative Frequenz ein Verhältnis von 1,05, welches etwas weniger als einen Halbton darstellt.

Die abschließende Untersuchung der vorliegenden Pitchsegmente besteht in der Auswertung der vorhandenen Frequenzsteigungen. Im Groben wird dabei ausgewertet, ob der betrachtete Abschnitt bei Unterschreitung einer Mindestlänge einen zu steilen Verlauf aufweist. Sollte dies der Fall sein, so wird der betreffende Bereich als Durchgangssegment aufgefasst und aus der Ansammlung gültiger Noten entfernt. In Abbildung 5.12 repräsentieren die horizontalen Linien das typische Ergebnis einer pitchbasierten Segmentierung.

Nach Abschluss der zweiten Segmentierungsstufe kommt das Verfahren, im Vorgriff auf die in Kapitel 5.2 vorgestellten polyphonen Strategien, nochmals auf die Auswertung amplitudenbehafteter Kriterien zurück. Dazu wer-

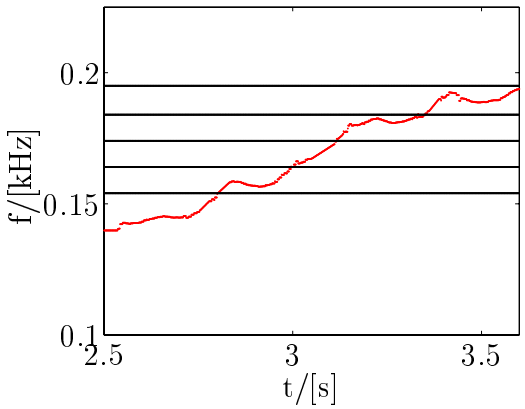

Abbildung 5.12: Pitch-Segmentierung

den zunächst als Maß der Lautheit der einzelnen Segmente die mittleren Amplitudenwerte berechnet. Noten mit, im Vergleich zu ihrer direkten Umgebung, zu geringen Lautstärkepegeln werden als perzeptiv nicht relevant angenommen. Die Unterschreitung eines Schwellenwertes von 0,6 bezogen auf den Amplitudenmittelwert führt zur Eliminierung des Eintrages.

Erfüllen die durch die Segmentierungsverfahren entstandenen Teiltrajektorien nicht das Kriterium einer Mindestlänge von 40 ms, werden diese aus dem Vektor gültiger Noten entfernt. Das Endergebnis der Segmentierung

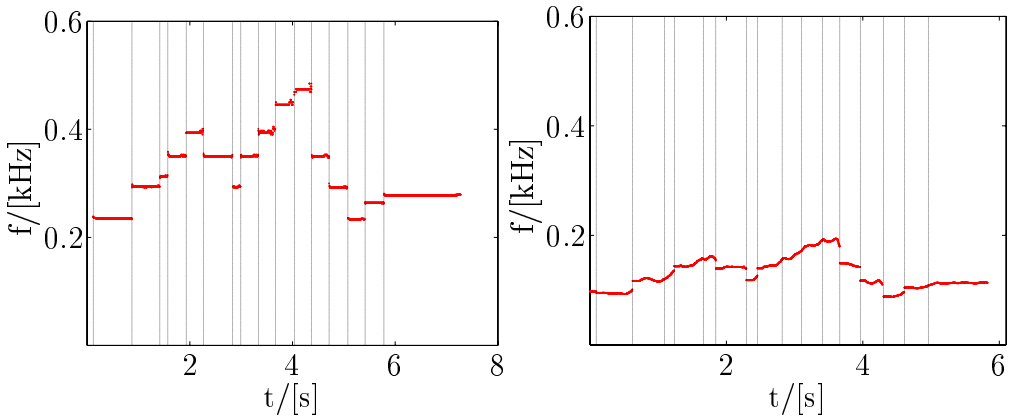

Abbildung 5.13: Pitch-Trajektorien

der globalen Pitchtrajektorien ist für beide Beispielmelodien in Abbildung 5.13 dargestellt. Die vertikalen Linien entsprechen dabei den gefundenen Notenanfängen. Die Grenzen aller gespielten bzw. gesungenen Noten sind mit

dem beschriebenen Verfahren korrekt analysiert worden.

Zusammenfassend lässt sich sagen, dass gerade die Kombination aus verschiedenen amplituden- und pitchmotivierten Verfahren zur Segmentierung vom monophonen Melodien ein probates Mittel darstellen und der komplexen menschlichen Wahrnehmung eher nahekommen als die Beschränkung auf einzelne Analysekriterien.

Ergänzend sind in Tabelle 5.1 exemplarisch für einen Klarinettenton sämtliche extrahierten Informationen bezogen auf den Notenanfang dargestellt. Neben den schon im vorherigen benutzten Parametern (Onset-Zeitpunkt, Frequenz und Amplitude) ist auch die Resonanzbreite der Teiltöne aufgeführt, die ein Maß für instrumentenspezifische Anregungscharakteristiken längs der Basilarmembran aufstellt. Diese Werte können unter anderem auch zur Klassifikation von Klangfarben benutzt werden [HB03].

Partialton $[Nr.]$	Onset $[s]$	Frequenz $[Hz]$	Amplitude []	Haarzelle $[Nr.]$	Resonanzbreite $[Nr.]$
1	0.187	236.4	0.00878	27	30
2	0.218	467.8	0.00555	48	14
3	0.192	703.5	0.00862	73	36
4	0.219	878.2	0.00735	94	2
5	0.201	1170.6	0.00733	102	12
6	0.202	1405.3	0.00776	113	12
7	0.206	1640.8	0.00783	123	16

Tabelle 5.1: Notenanfang, Klarinette Bb3

5.1.3 Interpretative Nachbearbeitung

In einer Kette von Nachbearbeitungsschritten sollen die im vorherigen Kapitel 5.1.2 erhaltenen Ergebnisse im Hinblick auf wesentliche, melodiebestimmende Elemente reduziert werden. Zum einen werden hierfür wiederum perzeptiv motivierte Ansätze benutzt, die an die in Kapitel 5.2 über polyphone Strategien eingehender benutzte auditorische Szenenanalyse [Bre90] angelehnt sind. Weiterhin kommen kulturell basierte Annahmen zur Anwendung, die bestimmte charakteristische Eigenschaften der im Bereich der westlichen

Musikwelt benutzten Kompositionsprinzipien voraussetzen. Schließlich soll auch der schon erwähnten Anwendung des Verfahrens in einem „Query-By-Humming“-System Rechnung getragen werden. Dazu werden die Analyse-Parameter im Sinne einer Optimierung der Ergebnisse des zugehörigen Suchalgorithmus kalibriert.

Der Eingabedatenraum des in diesem Abschnitt beschriebenen Verfahrens setzt sich zusammen aus dem im Vorherigen extrahierten Notenvektor. Dessen Elemente beinhalten Informationen über die mittlere Frequenz des Grundtons, die in der Regel auch ein gutes Maß für die wahrgenommene Tonhöhe (Pitch) darstellt. Desweiteren ist jeder Note ein exakter Startzeitpunkt und eine zugehörige Tondauer in der Auflösung von 1 ms zugeordnet.

Ein erster Interpretationsschritt besteht in der Detektion längerer Pausen, die eine Melodie beenden bzw. zwei Melodien voneinander trennen. Bei einer Pausenlänge von mehr als 2,5 s wird angenommen, dass es sich um eine signifikante Aufteilung des Eingangssignals in zwei nicht mehr eindeutig zusammengehörige Teile handelt.

In Anlehnung an die Gestaltgesetze (s. Kapitel 3.2) werden dann Sequenzen von Tonintervallen gesucht, die aus zwei nachfolgend und in Gegenrichtung ausgeführten Tonsprüngen von in beiden Fällen mehr als 12 Halbtönen, also einer Oktave, bestehen. In solchen Fällen führen die Prinzipien von fehlender Nähe und nicht vorhandener guter Fortsetzung zu einem Auseinanderfallen der Elemente im Gruppierungsprozess. Diese Vorgehensweise lässt sich auch im Hinblick auf die musikalische Realität rechtfertigen. Eine solch extreme Intervallabfolge findet sich nur in einer verschwindenden Anzahl von Melodien. Die in Abbildung 5.14 zu findende Notation einer Phantasieme-

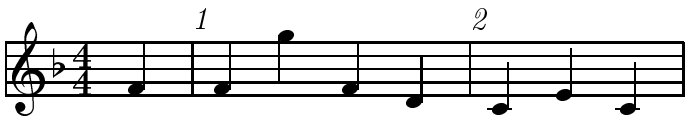

Abbildung 5.14: Phantasiemelodie - Tonintervalle

lodie veranschaulicht auf visuelle Weise die Vorgänge. Die dritte Note wird auch bei Betrachtung der Darstellung kaum zu einer Einheit mit den anderen Noten gruppiert werden. Als Konsequenz würde dieser Noteneintrag aus der aktuellen Melodiesequenz entfernt.

Aufgrund der komplexen transienten Abläufe während des Einschwingvorganges einzelner Töne kommt es im Segmentierungsverfahren mitunter

zu einer zusätzlichen Unterteilung am Anfang von Frequenztrajektorien. Die Strategie zur Eliminierung dieser Fehleinschätzung benutzt folgende Kriterien, um rückwirkend eine Fusionierung der beiden irrtümlich voneinander getrennten Anteile vorzunehmen:

1. Die Noten liegen auf der gleichen Tonhöhe.

2. Der Abstand beträgt weniger als 50 ms.

3. Die erste Note erstreckt sich um weniger als 100 ms.

Gerade im Bereich gesungener Melodien, die den Hauptanteil der in einem „Query-By-Humming"-System untersuchten Eingaben darstellen, ist eine solche bewusste Artikulation äußerst unwahrscheinlich und insbesondere durch Laien kaum durchführbar.

Die im weiteren beschriebenen interpretativen Nachbearbeitungsschritte beziehen sich im wesentlichen auf die im westlichen Kulturkreis benutzte, zwölf Einträge umfassende, Tonleiter in wohltemperierter Stimmung. Die realen, in der Regel von den musiktheoretischen Vorgaben abweichenden, Frequenzeinträge werden auf ihre Distanzen zu den idealen Tonleitern hin untersucht.

Zur Vorbereitung dieser Methodik werden zunächst in Art eines gleitenden Mittelwertes aufeinderfolgende, quasi gleichfrequente Noten auf *eine* exakt gleiche Tonhöhe gebracht. Chronologisch wird jeweils zwei Tönen, die einen frequenzmäßigen Abstand von weniger als einem Achtelton aufweisen, eine mittlere Frequenz zugewiesen. Diese Durchschnittsfrequenz wird anschließend als Ausgangspunkt zur Untersuchung späterer ähnlicher Frequenzen benutzt. Findet sich wieder eine Übereinstimmung, so werden auch alle vorher involvierten Noten auf die neue mittlere Frequenz gesetzt. Dies führt zu einer ersten Quantisierung der Notenwerte, die sich aber noch unabhängig von einem vorgegebenen Tonsystem darstellt.

Im nächsten Schritt werden dann die bis hierhin festgelegten Melodiesegmente in untergeordnete Einheiten aufgeteilt. Die Vorgehensweise lässt sich anhand der einprägsamen Anfangsmelodie der italienischen Nationalhymne illustrieren. In Abbildung 5.15 findet man die ersten 8 Noten. Jeweils 4 Noten bilden musikalisch, wie auch optisch, eine Einheit, die im Folgenden als Phrase bezeichnet werden soll. Als Parameter zur Indizierung solcher Phrasenumbrüche werden Eingabepausen einer Mindestlänge von 300 ms sowie Intervallsprünge von wenigstens 6 Halbtönen festgelegt.

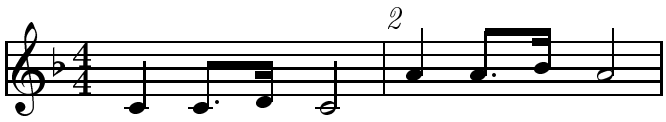

Abbildung 5.15: „Italienische Nationalhymne“ - Musikalische Phrasen

Begründung für die Einführung dieser Unterteilung ist die Tatsache, dass, insbesondere bei von Laien gesungenen Eingaben, häufig das große Intervall zur Einleitung der neuen Phrase unsauber wiedergegeben wird, während die Noten innerhalb des Subsegmentes meist einen stimmungstechnisch sauberen Verlauf aufweisen. Die weiteren Analyseschritte beziehen sich dann jeweils auf die Noten innerhalb einer solchen Einheit und werden auch in dieser Form dem Suchalgorithmus des „Query-By-Humming“-Systems übergeben.

Die so segmentierten Untereinheiten werden nun erstmals auf ein konkretes Tonsystem bezogen. Die übliche wohltemperierte Stimmung des Notenintervallsystems führt zu einer gleichmäßigen Aufteilung einer Oktave in 12 Halbtonschritte, die ein relatives Frequenzverhältnis von $\sqrt[12]{2}$ ($(\sqrt[12]{2})^{12} = 2 =$ 1 Oktave) aufweisen. Als Referenzstimmung hat sich in der musikalischen Praxis die Zuordnung einer Frequenz von 440 Hz für die Note „A4“ bewährt (Kammerton A), auf die sich, entsprechend der obigen Halbtondefinition, die übrigen Notenfrequenzen beziehen. In der Regel trifft aber ein Sänger ohne *absolutes Gehör*, dessen Gabe nur einer sehr kleinen Gruppe von Menschen zuteil wird, nicht diese standardisierte Idealstimmung. Daher wird im nächsten Verfahrensschritt versucht, die Töne innerhalb einer Phraseneinheit auf eine wahrscheinlichste Intonation zu beziehen. Mittels der Methode der kleinsten Fehlerquadrate wird nach der Stimmungsfrequenz gesucht, bei der die Abweichung für die Gesamtheit der einbezogenen Einzelnoten einen möglichst kleinen Wert aufweist. Nach der frequenzmäßigen Festlegung des Notenrasters wird dann jedem Ton ein auf die gefundene Skala bezogener Notenname zugeordnet und so erstmals der Schritt von physikalischer auf musikalische Schreibweise vollzogen.

In der abschließenden Verarbeitungsstufe wird dann von der Annahme Gebrauch gemacht, dass die in der heutigen Musikwelt verbreiteten Kompositionen praktisch ausschließlich auf den gewöhnlichen Dur- und Molltonleitern aufbauen. Wiederum über die Methode kleinster Fehlersummen wird dann den untersuchten Eingaben eine wahrscheinlichste Tonleiter zugeordnet. Die Molltonarten werden dabei jeweils als Paralleltonart der Durtonleiter

aufgefasst und somit die Suche auf 12 Durtonarten beschränkt.

Töne, die sich nach Festlegung einer Referenztonart, gemäß ihres zugeordneten Notennamens nicht in diese einfügen lassen, werden daraufhin einer gesonderten Untersuchung unterzogen. Damit soll überprüft werden, ob sie aufgrund intonationstechnischer Unsauberkeiten aus dem Raster gefallen sind oder als tonartfremde Melodieteile interpretiert werden müssen. Konkret wird die physikalische Frequenz des Tons mit der quantisierten Frequenz des vermuteten Notenrasters verglichen. Befindet sich die untersuchte Note in einem „Achteltonschlauch“ (relativer Frequenzfaktor = $\sqrt[48]{2} = 1{,}015$) um eine Tonleiterfrequenz, so wird sie als schlecht intoniert angenommen und korrigiert. Andererseits behält Sie ihre von der Tonleiter abweichende Notenbezeichnung. Beispiel hierfür ist die letzte Note in der Beispielmelodie aus Prokoffiews „Peter und der Wolf“ (s. Abb. 5.1). Das vorliegende „Gis4“ *passt* nicht in die vorzeichenlose Ausgangstonart. Vielmehr liegt an dieser Stelle eine Modulation der Harmonik in der Komposition vor, die durch diesen Ton mit eingeleitet wird.

Als zusammenfassendes Ergebnis der monophonen Melodietranskription sind in Tabelle 5.2 die Parameter der extrahierten Noten der Beispielmelodien aus „Peter und der Wolf“ zusammengefasst. Im Vergleich zum musikalischen Ausgangsmaterial (s. Abb. 5.1) erkennt man, dass alle Noten im Sinne gegenseitiger relativer zeitlicher und frequenzmäßiger Beziehungen richtig erkannt worden sind. Bedingt durch persönliche artikulatorische Unterschiede von Sänger und Instrumentalist ergeben sich natürlich Abweichungen im Bezug auf den wiedergegebenen Frequenzbereich sowie das gewählte Tempo. Diese Differenzen erweisen sich aber als invariant bei der Einspeisung der extrahierten Melodien in das in Kapitel 6.1 beschriebene „Query-By-Humming“-System, da die dort vollzogene Suche in einer Referenzdatenbank ausschließlich auf relativen Verhältnissen der Noten innerhalb der untersuchten Sequenzen beruht.

5.2 Polyphone Strategien

Nachdem im bisherigen Verlauf dieses Kapitels ausführlich auf die Verwendung des physiologischen Ohrmodells bei der monophonen Melodieextraktion eingegangen worden ist, sollen nun grundlegende Strategien zur Analyse und Interpretation polyphoner Musikstücke hergeleitet und erläutert werden.

	Klarinette			Gesang		
Note [Nr.]	Onset t/[s]	Dauer t/[s]	f f/[Hz]	Onset t/[s]	Dauer t/[s]	f f/[Hz]
1	0.115	0.757	236	0.090	0.531	98
2	0.873	0.543	294	0.626	0.437	117
3	1.417	0.144	312	1.094	0.14	123
4	1.573	0.353	350	1.236	0.284	144
5	1.927	0.330	394	1.661	0.183	155
6	2.267	0.562	350	1.846	0.446	142
7	2.837	0.154	294	2.294	0.160	124
8	2.993	0.339	350	2.456	0.321	142
9	3.334	0.329	396	2.816	0.148	156
10	3.671	0.340	445	3.100	0.251	181
11	4.034	0.337	474	3.409	0.250	188
12	4.376	0.336	350	3.663	0.297	145
13	4.713	0.352	292	3.962	0.325	115
14	5.079	0.325	233	4.304	0.304	97
15	5.414	0.365	264	4.610	0.319	106
16	5.781	1.772	277	4.959	1.023	113

Tabelle 5.2: „Peter und der Wolf“ - Extrahierte Melodien

Ausgangspunkt für die weiteren Überlegungen ist das in Kapitel 4.5 vorgestellte Hierarchiemodell, das sich in seinen wesentlichen Annahmen auf die Theorie der Gestaltpsychologie [Wer25][And01] stützt. Wie schon in Kapitel 3.2 eingehend dargestellt, geht man bei diesem Ansatz davon aus, dass sich die Informationsverarbeitung des Menschen in hierarchisch aufgebauten Strukturen vollzieht. Diese im Laufe der kognitiven Prozesse innerhalb des Gehirns zunehmend abstrakteren und semantisch bedeutungsgeladeneren Stufen ergeben sich jeweils als Interpretation einzelner Elemente der vorhergehenden Wahrnehmungsschichten, die in Entscheidungsprozessen zu übergeordneten Einheiten zusammengefasst werden. Wesentlich ist dabei die hypothetische Ergänzung mitunter fehlender Teilelemente als Konsequenz aus der Verwendung der sogenannten Gestaltgestetze, die Kriterien zur Fusionierung dieser Fragmente zur Verfügung stellen.

Wie schon bei der Beschreibung der monophonen Vorgehensweise soll auch in diesem Abschnitt die Implementierung anhand eines konkreten musikalischen Beipieles illustriert werden. Die Wahl fiel dabei auf eine zweistim-

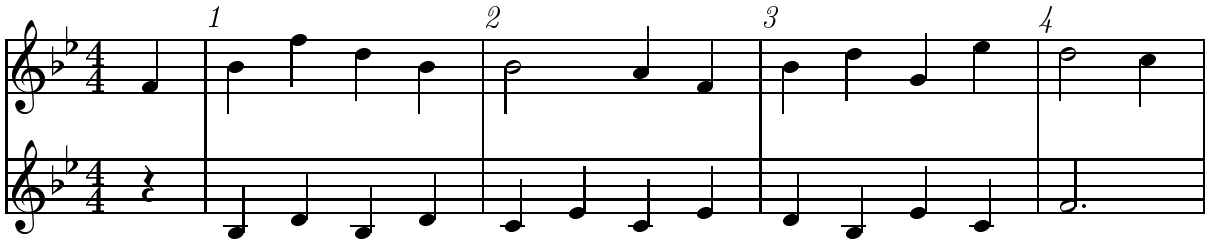

Abbildung 5.16: „Klarinettenduett“ - Musikalische Notation

mige Holzbläsermelodie, anhand derer die grundlegenden Strategien erläutert werden können, ohne durch die, polyphonen Darbietungen innewohnende, Komplexität der Vorgänge unübersichtlich zu erscheinen.

In Abbildung 5.16 findet sich die musikalische Notation eines Ausschnittes aus dem Duo Nr. 3, Satz 1 „Andante con variazioni“ aus „6 konzertante Duos für Klarinetten“ von Joseph Haydn (1732-1809) [Hay00]. Das untersuchte zweistimmige Melodiefragment ist gut geeignet, um die wesentlichen Ansätze zur Transkription polyphoner Inhalte übersichtlich darzustellen.

Abbildung 5.17 zeigt die zugehörige Schallwellenform, Auch hier kann man im Zeitverlauf noch gewisse Segmentgrenzen erkennen, die aber schon wesentlich weniger ausgeprägt sind als in den monophonen Beispielen und für

eine zuverlässige Segmentierung bei weitem nicht ausreichen. Mit zunehmen-

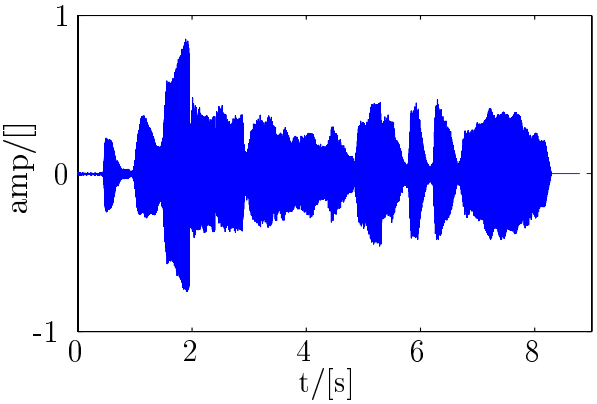

Abbildung 5.17: „Klarinettenduett“ - Schallwellenform

der musikalischer Komplexität (Mehrstimmigkeit, etc.) nimmt grundsätzlich die Tendenz einer „verschmierten“ Zeitdarstellung zu und Verfahren im Zeitbereich erweisen sich, wie schon in Kapitel 2 erläutert, in diesem Anwendungsfeld als nicht adäquat.

Die zweite Stufe im vorgestellten Hierarchiemodell (s. Abb. 4.3) bezieht sich auf die Ergebnisse der signalzerlegenden Prozesse der auditorischen Peripherie, die im weiteren Verlauf als Grundlage für die kognitiven Vorgänge zur (hypothetischen) Interpretation der innewohnenden musikalischen Informationen verwendet werden. Abbildung 5.18 zeigt in der Hüllkurvendarstellung die zur Auslösung von neuronalen Impulsen verantwortliche Konzentration von Transmittersubstanz im synaptischen Spalt zwischen inneren Haarzellen und den Fasern des Hörnervs. Die Darstellung ist limitiert auf die ersten 200 Sektionen von inneren Haarzellen entlang der Basilarmembran, da das zur Frequenzextraktion benutzte Verfahren des „Phase-Locking“ nur für spektrale Anteile bis maximal 5 - 6 kHz Gültigkeit besitzt. Die im Modell verwendeten zusätzlichen 51 Haarzellensektionen besitzen charakteristische Resonanzen oberhalb dieser Frequenzgrenze und sind somit nicht für eine entsprechende Auswertung geeignet. In weiterführenden Untersuchungen sollen aber über übliche Orts-Frequenz-Transformationen auch aus diesen Bereichen Informationen hochfrequenter Partialtonanteile gewonnen werden.

Allerdings erkennt man sowohl in der perspektivischen Darstellung als auch in der Draufsicht schon eine wesentlich strukturiertere Anordnung der im Signal enthaltenen Anteile als dies im Zeitsignal der Fall war. Insbeson-

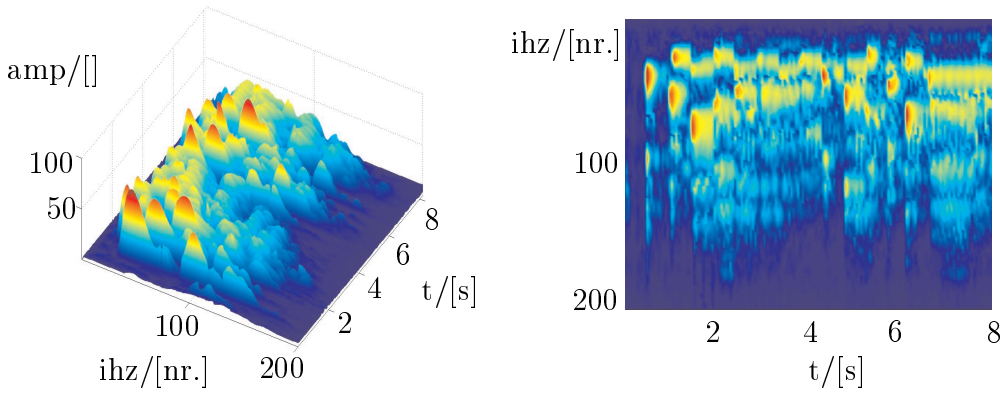

Abbildung 5.18: Transmitterkonzentration der inneren Haarzellen

dere die Grafik auf der rechten Seite gibt schon einige Hinweise auf die zu erwartende Partialtonanordnung. Wie bereits im monophonen Teil reicht aber auch hier diese Betrachtungsweise nicht aus, und eine erfolgreiche Analyse bedingt die Untersuchung der Feinstruktur der Spaltkonzentrationen.

In Abbildung 5.19 sind zur vorbereitenden Erläuterung der benutzten Strategien die theoretisch idealen Obertonreihen für das betrachtete Beispiel wiedergegeben. In roter Farbe sind jeweils die den Grundfrequenzen der

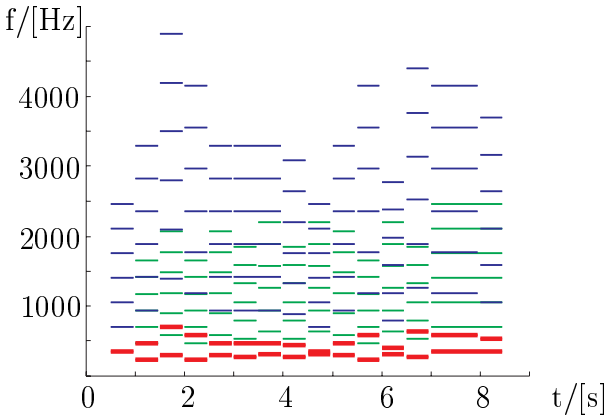

Abbildung 5.19: Ideale Obertonstruktur

Klarinettentönen zugeordneten Notenverläufe dargestellt, während in blau bzw. grün die harmonischen Obertoneinträge der beiden Stimmen zu sehen sind. Primäres Ziel der nachfolgenden Untersuchungen ist es, eben solche

Anordnungen von Teiltönen aufzuspüren und daraus, über bestimmte gestaltbasierte Prinzipien, Hypothesen über Einzelnoten aufzustellen.

Noch analog zur rein monophonen Vorgehensweise wird über den Ansatz des „Phase-Locking“ (s. Kap. 4.4) aus dem zeitlichen Verlauf der Feinstruktur der Transmitterkonzentrationen die am Ausgang des Innenohres vorliegende spektrale Information ausgewertet. Im Takt der Signalanalyse (1 ms) wird über diese Zeitkodierung jeder Haarzellensektion eine anregende Frequenz zugeordnet.

Da die Schwingungen der Basilarmembran als mechanisch kontinuierliches Konstrukt immer über einen gewissen räumlichen Bereich ausgedehnt verlaufen, werden im Sinne einer hierarchischen Informationsverarbeitung, aber auch zur Reduktion gleichartiger Dateninhalte zu jedem Analysezeitpunkt die sich aus dem „Phase-Locking“ ergebenden Frequenzen zusammengefasst. Dazu werden Haarzelleneinträge, die den Gestaltgesetzen der Nähe und Ähnlichkeit genügen, in sogenannte Resonanzbereiche fusioniert, d.h. Frequenzen, die sowohl in ihrem Hz-Wert als auch zusätzlich in ihrer Lokalisierung auf der Basilarmembran nahe beieinander angeordnet sind, werden zu einer übergeordneten Einheit verschmolzen. Dabei werden auch, ganz in der Tradition der Gestalttheorie, Lückenwerte akzeptiert, die zwar der Bildung solcher Einheiten entgegenstehen, die starken Vereinigungstendenzen aber nicht aufheben können. Die Anwendung dieser gestaltorientierten Interpretationsweise der Transmitterkonzentrationen und ihrer abgeleiteten Größen erweist sich, wie schon vorher angedeutet, in der hier vorgestellten polyphonen Verarbeitungsstufe in Zeit- und Frequenzbereich als wesentliche Vorgehensweise und wird in den nachfolgenden Schritten zur Umsetzung des Hierarchiemodells ein ums andere Mal zur Anwendung kommen. Den resultierenden Resonanzbereichen wird eine Wertigkeit zugeordnet, die sich aus der Summe der einzelnen Hüllkurvenamplituden ergibt.

Ähnlich der Extraktion der Pitchtrajektorien bei der monophonen Aufgabenstellung zeigt sich die nächste polyphone Stufe. Aus den im vorherigen Schritt gewonnenen Resonanzbereichen sollen Verläufe bestimmt werden, die dem zeitlichen Verlauf der Partialtontrajektorien entsprechen, d.h. in Frequenz und Zeit benachbarte Haarzellensektionen werden wiederum in übergeordnete Einheiten zusammengefasst. Die hierfür verwendeten Strategien folgen den bei der monophonen Vorgehensweise benutzten (s. Kap. 5.1.1). Nach Herausbildung sogenannter Subtrajektorien, die strengen Kriterien bei der Kontinuitätsvorgabe genügen müssen, werden diese Kernblöcke verwendet, um aus ihnen globale Trajektorien zu gewinnen. Bei Überschreitung

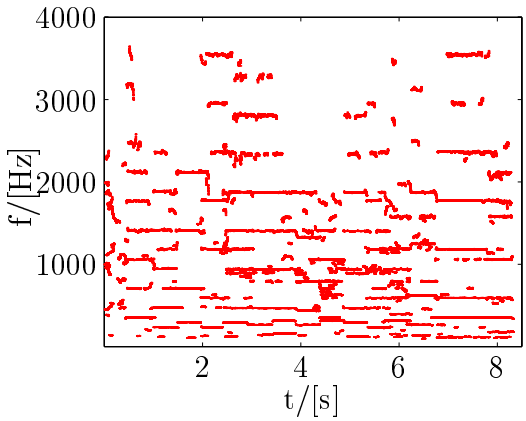

Abbildung 5.20: Transmitterkonzentration der inneren Haarzellen

einer Mindestlänge von 40 ms werden diese in den Vektor potentiell wichtiger Teiltonlinien übernommen. Gestaltorientierte Größen wie Abstand, gute Fortsetzung, etc. prägen auch in diesem Fall die Vorgehensweise.

In Abbildung 5.20 ist das Ergebnis der Suche nach Partialtontrajektorien, die sich letztendlich ausschließlich aus dem Prinzip des „Phase-Locking“ ergeben, zusammengefasst. Man kann hier bereits einige der Grundzüge der idealen Obertonstruktur aus Abbildung 5.19 erkennen. Wie sich aber in den eigenen Untersuchungen herausgestellt hat, ist die alleinige Verwendung der „Phase-Locking“-Frequenzen nicht ausreichend für eine weiterführende polyphone Verwendung.

5.2.1 Partialtoninterferenzen

Zur Gewinnung weiterer Informationen werden wiederum Erkenntnisse aus der Psychoakustik zur Motivation zusätzlicher Strategien herangezogen. In der Literatur ([ZF01], etc.) sind eine Reihe spektraler Verdeckungseffekte beschrieben, deren Ursachen sich in den Hüllkurven der physiologischen Vorgänge im Innenohr wiederfinden lassen. Hauptproblem für die polyphone Anwendung ist, dass anregende Frequenzen großer Amplituden ausgeprägte Resonanzbereiche erzeugen, die solche, verursacht von Frequenzen niedriger Intensitäten, dominant überlagern. Als Konsequenz schwingen dann diese Bereiche in der Regel mit „Phase-Locking“-Frequenzen, die den amplitudenstärkeren Vibrationen entsprechen. Die Einführung eines neuen Ansatzes, in Abbildung 4.3 und im folgenden als „AMDF“-Trajektorien bezeichnet,

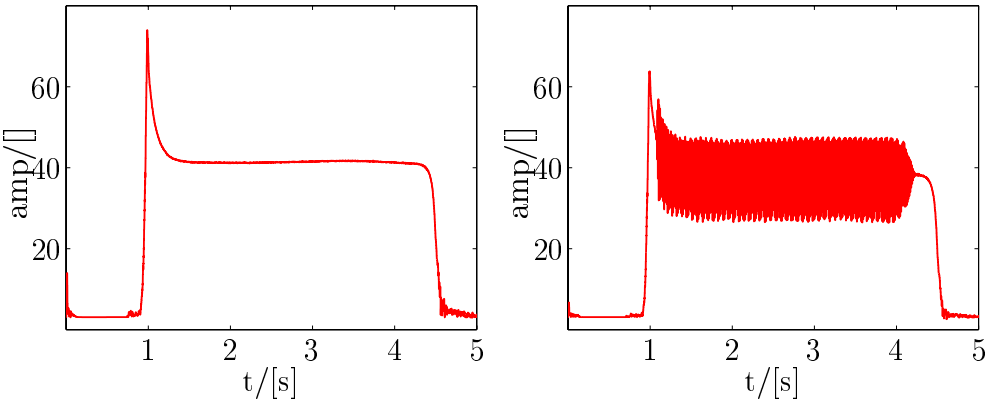

Abbildung 5.21: Partialtoninterferenz

zeigt, dass sich auch in den Bereichen verdeckter Spektralelemente Informationen über eben diese finden lassen. Die Bezeichnung „AMDF“ bezieht sich auf das gewählte Verfahren der „Average Magnitude Difference Function“, das auch schon im Stand der Technik (Kap. 2.1) als probates Mittel der Pitchbestimmung über Autokorrelationsbetrachtungen beschrieben worden ist.

In Abbildung 5.21 ist in einem Beispiel der zeitliche Verlauf der Hüllkurve der Transmitterkonzentration der Haarzelle Nr. 41 dargestellt. Auf der linken Seite findet sich das Signal für einen einstimmigen Klarinettenton der Frequenz 350 Hz. Der Verlauf entspricht weitgehend der theoretisch vorhergesagten idealen Form, d.h. bei Beginn der Note wird das „Alarmsystem“ der inneren Haarzellen aktiviert, und es kommt zu einem signifikanten Anstieg des Spaltinhaltes. Nach einer Phase der Adaptation pegelt sich die Hüllkurve schließlich auf einem quasi-stationären Wert ein, um dann bei Beendigung der Note auf die Ruhekonzentration abzufallen.

Abweichend das Verhalten in der Grafik auf der rechten Seite: kurz nach Beginn der ersten Note wird ein zweiter Ton mit einer Frequenz von 440 Hz zugeschaltet. Die Idealform geht verloren, und an ihre Stelle tritt ein unruhiger Verlauf des resultierenden Schwingungsmusters.

Die Verhältnisse werden etwas klarer, wenn man sich zu einem bestimmten Zeitpunkt die Verteilung der Spaltinhalte der ersten 200 Haarzellensektionen entlang der Basilarmembran anschaut. In Abbildung 5.22 sind die Konzentrationen wieder für einen Einzelton und die Kombination zweier relativ nah benchbarter Noten illustriert. In der Einzeltondarstellung auf der

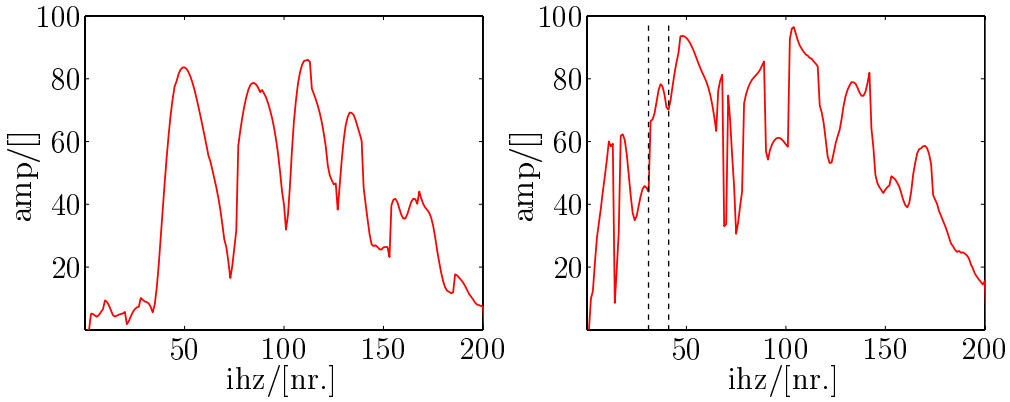

Abbildung 5.22: Transmitterkonzentration bei Interferenz

linken Seite kann man die Ausprägung der, den jeweiligen Partialtönen zugeordneten, Resonanzbereiche gut erkennen. Hingegen resultiert die kombinierte Darbietung zweier Töne in einem weniger klar strukturierten Bild. Die einzelnen charakteristischen Anregungsbereiche sind nicht mehr eindeutig nachvollziehbar. Im gestrichelten Bereich befindet sich der für das gewählte Beispiel interessante Bereich. Hier ergibt sich eine ausgeprägte Interferenz der Grundtöne der beiden involvierten Anregungsfrequenzen. Jeder Basiston würde für sich in den Haarzellen dieses Abschnittes einen klar definierten, wie in Abbildung 5.21 auf der linken Seite zu findenden, idealen Verlauf aufweisen. Aufgrund der dominierenden Amplituden des oberen Frequenzbeitrages schwingt aber dieser Bereich in der „Phase-Locking“-Betrachtung mit einer Frequenz von 440 Hz. Der eigentlich ebenfalls hier vorhandene Beitrag von 350 Hz würde so verloren gehen.

Eine Vergrößerung des Zeitverlaufes der überlagerten Schwingungen aus Abbildung 5.21 (rechte Seite) führt zur Darstellung der Feinstruktur im Bereich von 2,4 - 2,5 s in Abbildung 5.23 für die Haarzelle 41. Die blauen Linien geben die tatsächlichen Schwankungen der Transmitterkonzentrationen an. Auswertung der Abstände der einzelnen Maxima würde, wie schon vorher erwähnt, zu einer „Phase-Locking“-Frequenz von 440 Hz führen. Auffällig in der Grafik ist allerdings die offensichtlich regelmäßige Amplitudenmodulation der blauen Linie, deren Verlauf durch die rote Hüllkurve verdeutlicht werden soll. Dieser Verlauf ist das Ergebnis der Interferenz der beiden beteiligten anregenden Frequenzen. Vor der ausführlichen Diskussion bezüglich möglicher Interpretationen des vorliegenden Schwingungsmuster soll zunächst kurz das Verfahren erläutert werden, mit dem die Periode der vermuteten Modulation

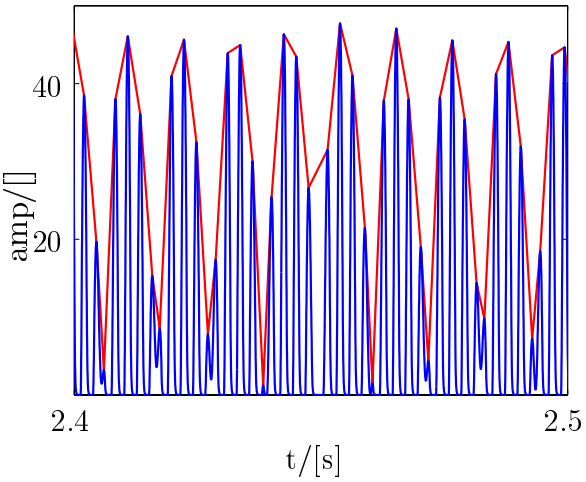

Abbildung 5.23: Amplitudenmodulation

festgestellt werden kann. Ähnlich wie bei verschiedenen Pitchbestimmungsverfahren wird dazu im vorliegenden Fall nach Autokorrelationen in den Zeitverläufen der einzelnen Haarzellensektionen gesucht. Die Wahl fiel dabei auf das Verfahren der „Average Magnitude Difference Function“ (AMDF). Die Berechnung erfolgt über Differenz- und Betragsbildung des mit sich selbst verschobenen Signales.

$$AMDF(t_i, \tau) = \sum_{i=1}^{N} |s(t_i) - s(t_i + \tau)| \,. \tag{5.1}$$

Damit entfällt die Benutzung der bei der „normalen“ Autokorrelation anfallenden Multiplikationsschritte und macht das Verfahren aus Effizienzgründen bei der Implementierung auf einem handelsüblichen PC vorteilhaft. Die resultierende AMDF-Funktion, welche sich aus der in Abbildung 5.23 dargestellten Amplitudenkurve ergibt, ist in Abbildung 5.24 zu sehen. Dargestellt ist der Verlauf in Abhängigkeit vom, in diesem Fall normierten, Verschiebungsparameter τ. Ähnlich wie bei der Auswertung der Autokorrelationsmaxima [Rab77b] sucht man in diesem Fall nach dem ersten signifikanten Minimum im Kurvenverlauf. Umrechnung der gefundenen normierten Zeitverschiebung in SI-Einheiten ergibt eine Periodiziät der Amplitudenmodulation der Haarzellenkonzentrationen von $\Delta t_{mod} = 0{,}011$ s. Dies entpricht einer zugehörigen Frequenz von ungefähr $f_{amdf} = \frac{1}{\Delta t_{mod}} = 90$ Hz. Die Interpretation dieser Amplitudenmodulation ist einfach und einleuchtend. Man erkennt, dass sich die

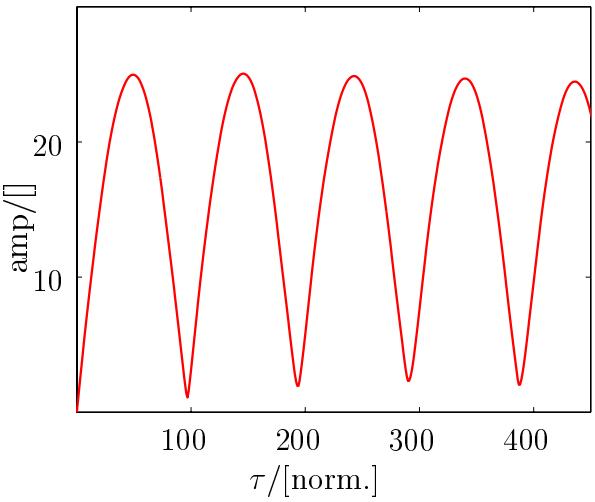

Abbildung 5.24: AMDF-Funktion

gefundene Modulationsfrequenz als Differenz der beiden interferierenden Anregungsfrequenzen im betrachteten Basilarmembranbereich darstellen lässt:

$$f_{amdf} = f_2 - f_1 = 440Hz - 350Hz = 90Hz. \tag{5.2}$$

Prinzipiell lassen sich zwei miteinander wechselwirkende Einzelfrequenzen und die daraus resultierende Modulationsfrequenz über folgende Formel allgemeingültig miteinander verknüpfen:

$$f_{mod} = m \cdot f_2 \pm n \cdot f_1. \tag{5.3}$$

Die Periode der Amplitudenmodulation ergibt sich somit aus der Linearkombination der Summe bzw. Differenz von Vielfachen der Ausgangsschwingungen.

Aufgrund der physiologischen Vorgaben wurde dieser Ansatz für den vorliegenden Fall allerdings auf den Spezialfall der direkt beteiligten Teiltonfrequenzen reduziert. Da durch die Resonanzeigenschaften des Innenohres entlang der Basilarmembran nur frequenzmäßig nahe benachbarte Anregungen miteinander in „Kontakt“ treten, wurde die Auswertung der „AMDF“-Funktion auf die direkte Kombination aus den „Phase-Locking“-Frequenzen der Resonanzbereiche und den Korrelationswerten beschränkt.

Dies führt zu folgender Bedingung für die Aufstellung von Partialton-Hypothesen als Ergebnis der Überlagerung zweier in den Anregungssignalen

enthaltener Partialtöne:

$$f_{Partial,hyp} = f_{PL,i} \pm f_{amdf,j} \tag{5.4}$$

($f_{PL,i}$ = „Phase-Locking"-Frequenz, $f_{Partial,hyp}$ = hypothetische Partialtonfrequenz).

Um die Verwendung nur sporadisch auftretender Modulationswerte zu vermeiden, wurden verschiedene Nebenbedingungen eingeführt, die als Konsequenz der nachfolgenden Analyse ausschließlich perzeptuell bedeutsame Verläufe zuführen. Hier seien nur die zwei wichtigsten Kriterien genannt, die zur Elimierung rauschhaftiger Einträge dienen. Zum einen benötigen die Minima im „AMDF"-Verlauf einen deutlich ausgeprägten Extrempunkt, um als signifikant anerkannt zu werden. Der hierfür benutzte Parameter legt als Mindestverhältnis zwischen globalem Maximum und einem möglichen Minimumwert einen relativen Faktor von 0,5 fest. Der zweite wesentliche Bestandteil der Akzeptanzbedingung für mögliche Partialton-Modulationsfrequenzen besteht, analog zur Vorgehensweise bei den „gewöhnlichen" Partialtontrajektorien, in einer zeitlichen Mindestausdehnung der gefundenen Werte. Über verschiedene gestaltbasierte Kontinuitätskriterien (s. Kapitel 5.1.2) werden derart sogenannte Subtrajektorien der AMDF-Einträge extrahiert.

Anschließend kommt für jeden Zeitpunkt des Signalverlaufes eine vorsortierende Strategie zur Einschränkung der Anzahl möglicher vorkommender Teiltöne zum Einsatz. Über offensichtlich korrelierende Frequenzkombinationen gemäß Gleichung 5.4 wird versucht, den vorkommenden „AMDF"-Ergebnissen Eindeutigkeiten zuzuordnen. Die komplementären Kombinationsthesen werden verworfen und aus der Liste möglicher Partialtöne entfernt.

Dazu wird in zeitgleichen Einträgen nach folgenden zu erfüllenden „Matching"-Bedingungen gesucht:

$$\begin{array}{ll} & f_{PL,1} = f_{PL,2} - f_{amdf,2} \\ \vee & f_{PL,1} = f_{PL,2} + f_{amdf,2} \\ \vee & \left(\begin{array}{rcl} f_{PL,1} & = & f_{PL,2} - f_{amdf} \\ \wedge \; f_{PL,2} & = & f_{PL,1} + f_{amdf} \end{array} \right) \\ \vee & f = f_{PL,1} + f_{amdf,1} = f_{PL,2} - f_{amdf,2} \end{array} \tag{5.5}$$

Bei Erfüllung der Gleichung 5.5, d.h. bei sich ergänzenden Kombinationen zweier „Phase-Locking"-Frequenzen mit den Modulationseinträgen, wird angenommen, dass die gefundenen Korrelationswerte eindeutig interpretierbar sind. Dies kann wie in den ersten drei Fällen zu einer direkten Ergänzung

der Summen- und Differenzbildungen führen. In der Praxis ergeben sich aber auch Kombinationen, die eine gemeinsame, in der Analyse nicht entdeckte, „virtuelle“ Frequenz ergeben. Bei Vorkommen eines der fünf Fälle aus Gleichung 5.5 wird die entgegengesetzte Frequenzvermutung aus Gleichung 5.4 verworfen.

Im nächsten, mit dem Begriff „Partialton-Fusion“ bezeichneten Schritt werden die bis hierhin erhaltenen Einträge der Liste wahrnehmungssignifikanter Frequenzen, die die Nebenbedingung der Gleichung 5.6 erfüllen, zusammengefasst. Dazu werden jeweils die Amplitudenwertigkeiten aufaddiert und in einer gemeinsamen Frequenz verschmolzen.

$$\begin{aligned} & f_{PL,1} &=& f_{PL,2} - f_{amdf,2} \\ \vee \quad & f_{PL,2} &=& f_{PL,1} + f_{amdf,1} \end{aligned} \tag{5.6}$$

Als Resultat der Interpretation der vorkommenden „Phase-Locking“- und Modulationsfrequenzen erhält man schließlich als Summe die Gesamtheit der tatsächlichen und hypothetischen Partialtonfrequenzen. Diese werden wiederum in den nachfolgenden Abschnitten den semantisch höherwertigen Stufen der Pitch- bzw. Notenzuordnung und letztendlich der sequentiellen Integration zur Bestimmung von Melodieverläufen zugeführt.

5.2.2 Pitchhypothesen

Im bisherigen Verlauf waren die physiologischen Schwingungseigenschaften des Innenohres und die daraus resultierenden Transmitterkonzentrationen in den synaptischen Spalten der inneren Haarzellen als hauptsächliche Grundlage der Untersuchungen herangezogen worden. Das Augenmerk der weiteren Analyse besteht nun in stärker kognitiv orientierten Interpretationen der vorher gewonnenen Resultate. Hierbei spielen sowohl angeborene wie auch durch Erfahrung erworbene Strategien der auditiven Perzeption eine Rolle. Ein gutes Beispiel hierfür ist die Aufstellung der in diesem Kapitel beschriebenen Pitchhypothesen. Zum einen sorgt eine Obertonreihe für eine umso bestimmtere Pitchwahrnehmung je mehr ihre Teiltöne der idealen harmonischen Abfolge entsprechen. Zum anderen akzeptiert ein Zuhörer bis zu einem gewissen Grad auch Abweichungen von diesem Verhalten, insbesondere dann, wenn er dies durch praktische Erfahrung bei in der Praxis vorkommenden Instrumenten gelernt hat, wie dies zum Beispiel bei den nach hohen Partialtönen abfallenden Frequenzen der Streichinstrumente der Fall ist [FR98].

Wie schon angedeutet liegt der Schwerpunkt dieses Abschnittes in der Aufstellung von Pitchhypothesen bezüglich der vorkommenden Teiltöne. Dazu wird aus den vorher extrahierten signifikanten Frequenzen für alle Zeitpunkte der Analyseschrittweite für jeden involvierten Partialton eine Pitchwertigkeit W_{Pitch} aufgestellt, die sich ergibt aus den Amplituden einer aufzustellenden quasi-harmonischen Obertonreihe.

Ausgangspunkt für die Hypothesen sind jeweils die zum betrachteten Zeitpunkt vorhandenen Teiltöne, d.h. jede Frequenz f_i wird zunächst uneingeschränkt als möglicher Grundtonkandidat aufgefasst. Fundamentale Annahme für die Wahrnehmung einer stark pitchhaften Klangstruktur ist, wie gesagt, die Existenz einer nahezu idealen harmonischen Partialtonanordnung. Die Bedingung für die Eignung eines Partialtones f_j als Obertoneintrag einer vorgegebenen Grundfrequenz f_i muss somit lauten:

$$n \cdot \frac{f_i}{IsOctave} \leq f_j \leq n \cdot f_i \cdot IsOctave \quad (n \epsilon \mathcal{N}_{2..MNP}) \tag{5.7}$$

Die in diesem Modell akzeptierte maximale Abweichung vom idealen Oktavverhältnis („IsOctave“) beträgt höchstens einen Viertelton. Bei Erfüllung dieser Nebenbedingung kann man dann dem betreffenden Oberton einen Partialtonindex

$$j = int\left(\frac{f_j}{f_i}\right) \tag{5.8}$$

zuordnen, der sich also jeweils bezieht auf das relative Verhältnis zu dem potentiellen aktuellen Grundtoneintrag f_i.

Über die gewichtete Summation der Teiltonamplituden Amp_i wird dann für alle vorhandenen Frequenzen die Pitchwertigkeit berechnet. Hierzu findet eine empirisch entwickelte Berechnungsvorschrift Anwendung, die in Übereinstimmung mit verschiedenen in der Literatur beschriebenen Theorien ([PG79] [Her88][Sla90][MO98]) steht:

$$\begin{aligned} W_{Pitch}(Partialton_i) &= (MNP)^{GewExp} \cdot (Amp_i)^{FundExp} \\ &+ \sum_{j=2}^{MNP} (MNP - j + 1)^{GewExp} \cdot (Amp_j)^{PartialExp}. \end{aligned} \tag{5.9}$$

In ihrer Grundaussage bedeutet diese Formel, dass mit zunehmendem Partialtonindex der Einfluss eines Obertons auf den wahrgenommenen Grundtonpitch sukzessive abnimmt. Über die verschiedenen Koeffizienten kann explizit die Bedeutung der Teiltöne eingestellt werden. Es wird maximal die zwölfte

MNP	Maximaler Partialtonindex	12
GewExp	Gewichtungsexponent	2
FundExp	Grundtonexponent	4
PartialExp	Partialtonexponent	3
IsOctave	Oktavfaktor	1.029 (Viertelton)

Tabelle 5.3: Koeffizienten der Pitchhypothesen

Harmonische in den Aufbau einer Teiltonreihe mit einbezogen. Die, wie schon erwähnt, empirisch gefundenen Parameter sind in Tabelle 5.3 zusammengefasst.

Nachdem nun derart eine Anzahl von Pitchhypothesen für jeden Zeitpunkt aufgestellt worden ist, werden diese anschließend über verschiedene Ansätze von irrelevanten Einträgen bereinigt. Wertigkeiten, die unterhalb einen relativen Schwellwert bezogen auf den gleichzeitig bedeutsamsten Pitch fallen, werden verworfen. Desweiteren werden die üblichen Kontinuitäts- und Ausdehnungsbetrachtungen zur Herausbildung von Pitchtrajektorien herangezogen, die wiederum sporadische, rauschhafte Verläufe von perzeptuell wichtigen Inhalten trennen.

Ein Korrekturschritt soll nicht unerwähnt bleiben, der die turbulenten Vorgänge in den transienten Einschwingsequenzen nachverarbeitet. Bei verschiedenen Instrumenten kommt es zu einer stark ausgeprägten Verzögerung des Fundamentaltons zu Beginn einer neuen Note. Wenn es sich noch zusätzlich um einen ohnehin amplitudenschwachen Grundtoneintrag handelt, wie dies beispielsweise beim Fagott der Fall ist, so kann es zu einem signifikanten Verlust der „korrekten“ Pitchwerte zu Beginn einer neuen Trajektorie kommen, obwohl trotz der lückenhaften Obertonstruktur durchaus der richtige Grundton wahrgenommen werden würde. Zu Beginn einer Note wird daher untersucht, ob eine solche vorausschreitende Anordnung existiert. Bei positivem Ergebnis wird dann die jeweilige Pitchtrajektorie um eben diesen Zeitraum nach vorne ausgeweitet. Diese Vorgehensweise steht im Einklang mit Erkenntnissen aus der kognitiven Psychologie [And01], wonach der Mensch unter Verwendung eines Kurzzeitspeichers vorherige Sinneseindrücke im Sinne eines Gesamtkonzeptes nachträglich uminterpretieren kann.

Die nachfolgende Segmentierung, die im wesentlichen der in Kapitel 5.1.2 beschriebenen Vorgehensweise folgt, und Glättung der Frequenzverläufe führt zu den in Abbildung 5.25 wiedergegebenen Pitchtrajektorien, die somit letztendlich die erkannten Einzelnoten darstellen. Man erkennt bei Vergleich mit dem Notenverlauf aus Abbildung 5.16, dass bis auf den letzten Eintrag der Oberstimme bis hierhin soweit alle Noten richtig erkannt worden sind.

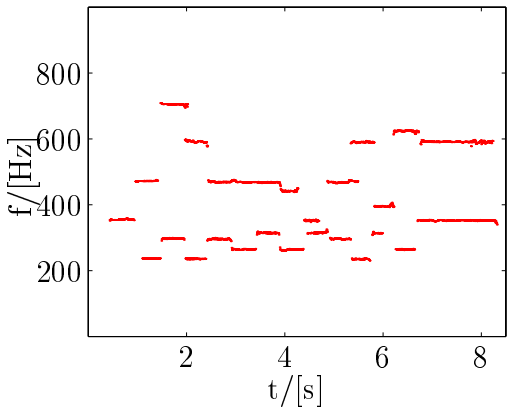

Abbildung 5.25: Haydn - Pitchtrajektorien

5.2.3 Oktaverkennung

Auf eine spezielle Schwierigkeit bei der richtigen Interpretation der aufgestellten Pitchhypothesen soll in diesem Abschnitt kurz gesondert eingegangen werden. Ein von den im Stand der Technik (Kapitel 2) beschriebenen Verfahren ungelöstes Problem besteht in der zuverlässigen Detektion von oktavverwandten Tonhöhenverläufen. Eine eigene Strategie, die sich bewusst an den für diese Arbeit maßgeblichen gestaltpsychologisch basierten Ansätzen orientiert, soll in diesem Zusammenhang erläutert werden. Im gewählten polyphonen Beispiel erkennt man in der musikalischen Notation in Abbildung 5.16, dass die beiden Klarinettenstimmen auf der ersten Viertelnote des ersten Taktes im Oktavabstand auseinanderliegen. Es handelt sich hierbei um die Noten Bb3 und Bb4. Die zugehörigen Frequenzen lassen sich in den gefundenen Pitchtrajektorien gemäß Abbildung 5.25 mit $f_1 = 236,9$ Hz (Bb3) bzw. $f_1 = 472,2$ Hz (Bb4) nachvollziehen.

Das zu lösende Problem besteht nun in der Unterscheidung zwischen Partialtonpaaren und voneinander unabhängigen Einzelnoten. Zur Veranschau-

lichung des hier vorgeschlagenen Lösungsansatzes sind in Abbildung 5.26 die Pitchwertigkeiten der beiden zur Diskussion stehenden Noten aufgetragen. Praktisch die gesamte Palette der Gestaltgesetze (Nähe, Ähnlichkeit, etc.)

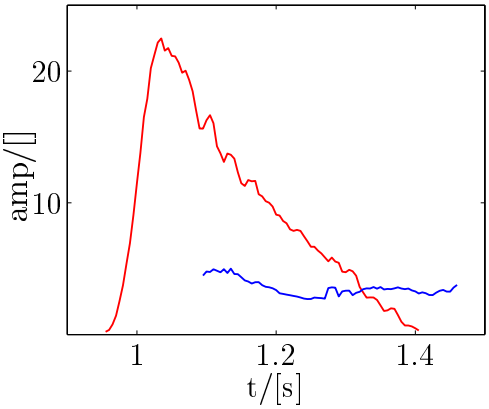

Abbildung 5.26: Haydn - Oktaven

lässt sich auf den dargestellten Fall anwenden, um die eigenständige Bedeutung der beiden Noten zu untermauern. Zunächst fällt die Asynchronität im Einsatz der beiden Wertigkeitsverläufe auf, die das perzeptiv stärkste Indiz zur Separierung darstellt. Aber auch der in sich stark unterschiedliche Amplitudenverlauf bestimmt den wahrnehmungsbezogenen Zerfall der Pitchverläufe. Der wesentlich ausgeprägtere Pitcheindruck der zuerst einsetzenden Note, in der Grafik wiedergegeben durch die deutlich höhere Wertigkeitsamplitude, lässt sich in der subjektiven Hörbewertung des Klangbeispieles verifizieren. Ein weiteres Entscheidungskriterium zur Separierung von Oktaven besteht mitunter in der Asynchronität im zeitlichen Verlauf der Partialtonfrequenzen. Allerdings konnte dieses Hilfsmittel zur Entscheidungsfindung im vorgestellten Beispiel nicht eindeutig nachgewiesen werden. In anderen mehrstimmigen Stücken ist dies aber hilfreich, insbesondere wenn starke Vibratos als musikalisches Stilmittel eingesetzt werden.

5.2.4 Sequentielle Integration

Die letzte und semantisch abstrakteste in dieser Arbeit verwendete Verarbeitungsebene beschäftigt sich mit der sequentiellen Integration („Streaming“) der bis hierhin extrahierten Einzelnoten. Darunter versteht man die Zusammenfassung der Notenobjekte in auditorische Ströme, die Bregman [Bre90]

sinngemäß wie folgt definiert:

> *„Perzeptuell übergeordnete Einheit als Summe zusammengehörig empfundener einzelner Soundobjekte“.*

Die Anwendung dieses Ansatzes auf die Fusion der bislang noch separierten Noten mündet schließlich in der Bereitstellung von im Musiksignal wahrgenommenen Melodien. Die besondere Herausforderung bei der Analyse polyphoner Musik besteht in der Möglichkeit gleichzeitig vorhandener eigenständiger Melodielinien.

Die untersuchten Notenobjekte lassen sich mit den folgenden Parametern semantisch beschreiben:

1. Startzeit
2. Endzeit
3. Pitch (Wahrgenommene Tonhöhe)
4. Lautheit
5. Klangfarbe

Auf die Benutzung und Beschreibung der Klangfarbeninformationen wurde hier verzichtet, da zum einen für das gewählte Beispiel die diesbezüglichen Unterschiede der zwei Klarinetten als marginal zu bezeichnen sind. Zum anderen weisen die Untersuchungen von Eichler [Eic03] und Hartmann [Har03] zwar nach, dass das benutzte Ohrmodell für monophone Anwendungen zuverlässige Klangfarbenzuordnungen treffen kann. Die Verfahren sind allerdings für die Benutzung in einem polyphonen Umfeld noch nicht ausgereift und bedürfen der Weiterentwicklung.

In Abbildung 5.27 ist der prinzipielle Verlauf der hier vorgeschlagenen Methode zur sequentiellen Integration illustriert. Den Kernpunkt des Algorithmus stellt die Berechnung von Kontinuitätshypothesen zwischen den einzelnen Notenobjekten dar. Entsprechend den für diese, der menschlichen Wahrnehmung durch Anlehnung an die beschriebenen Gestaltgesetze angepassten, erzielten Hypothesenwertigkeiten werden dann perzeptiv relevante Melodielinien extrahiert.

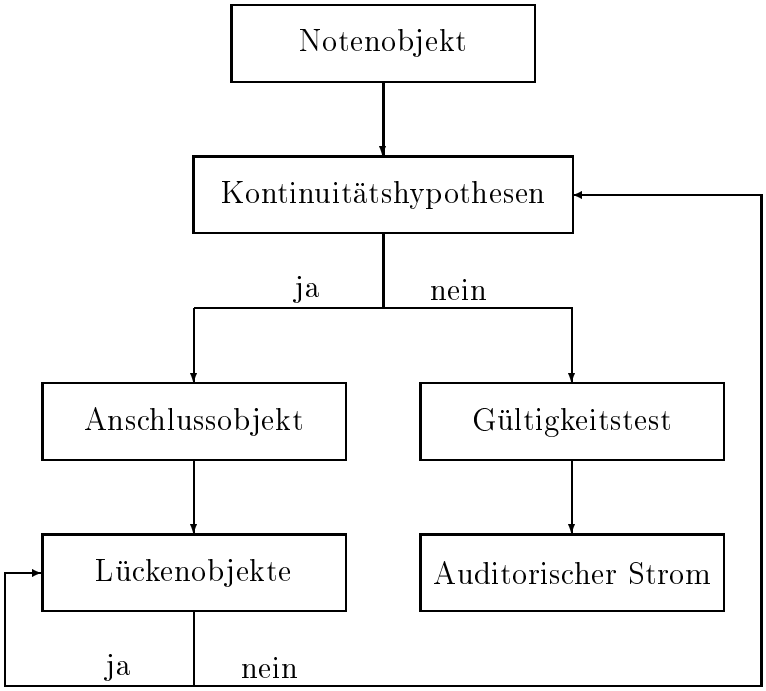

Abbildung 5.27: Sequentielle Integration

Zur Erstellung der auftretenden Nachbarschaftsbeziehungen werden für jedes Paar von Einzelnoten zunächst drei Abstandsmaße ausgewertet:

$$\begin{aligned} Zeit: \quad \Delta t &= t_2 - t_1 \\ Frequenz: \quad \Delta f &= f_2 / f_1 \\ Lautheit: \quad \Delta L &= L_2 - L_1 \end{aligned} \tag{5.10}$$

Es sei nochmals erwähnt, dass sich die hier benutzten Lautheiten aus den Amplitudenwerten der Spaltinhalte ergeben und nicht mit den aus der Psychoakustik verwendeten Größen (Sone bzw. Phon) verwechselt werden dürfen.

Die erhaltenen Werte für Zeit, Frequenz und Lautheit werden anschließend in normierte Abstände $HypVal_i$ überführt, um über die Gewichtsfaktoren g_i die Parameter explizit an die Randbedingungen des Ohrmodells anpassen zu können. Die hierfür benutzten Formeln, die jeweils zwischen den

g_1	Zeit	1.5
g_2	Frequenz	100.0
g_3	Lautheit	50.0

Tabelle 5.4: Gewichtsfaktoren

sich aus der Normierung ergebenden Randwerten linear interpolieren, lauten:

$$HypVal_i = \Delta i_{min} + m_i \cdot (\Delta i - \Delta i_{min}) \tag{5.11}$$

$$m_i = \frac{HypVal_{max} - HypValmin}{\Delta i_{max} - \Delta i_{min}} \tag{5.12}$$

(Der Index i steht jeweils für Zeit, Frequenz oder Lautheit).

Der resultierende Gesamtabstand $HypVal_{ges}$, der letztendlich als Kontinuitätshypothese zwischen zwei Noten herangezogen wird, ergibt sich schließlich aus der gewichteten Summe der 3 Einzelmaße:

$$HypVal_{ges} = \sum_{i=1}^{3}(g_i \cdot HypVal_i). \tag{5.13}$$

Die im Modell verwendeten Gewichtsfaktoren g_i sind in Tabelle 5.4 aufgeführt.

Nach der Berechnung der Kontinuitätshypothesen für alle möglichen Notenkombinationen werden die vorhandenen Melodieverläufe gemäß Abbildung 5.27 aus dem Vektor der Einzelnoten herausgebildet. Zunächst werden die plausibelsten Anschlussobjekte aneinandergehängt und so eine Kette von Noten erzeugt. Lassen sich Einträge finden, die zwar eine schlechtere Hypothese aufweisen, aber im Rahmen der Akzeptanz in eine vorhandene Linie einfügbar sind, so können diese bei ausreichend großen Zeitlücken in die entsprechenden Zwischenräume integriert werden.

Abschließend müssen die gefundenen Melodien noch einem Gültigkeitstest genügen, der unter anderem die Anzahl der Noten und die zeitliche Ausdehnung der Melodie als Kriterien heranzieht.

Auf die Darstellung der gefundenen Melodielinien für das vorgestellte Beispiel wird an dieser Stelle verzichtet, da die zwei extrahierten Verläufe im wesentlichen der Notation aus Abbildung 5.16 entsprechen. Lediglich der letzte

Ton der Oberstimme ist verloren gegangen. Aufgrund des Ausschwingvorganges des Vorgängertons wird die Resonanzstruktur des nicht detektierten recht kurzen Tons soweit verschmiert, dass er durch das verwendete Ohrmodell nicht mehr aufgelöst werden konnte.

Die grundsätzliche Herangehensweise zur Analyse von polyphonen Signalen konnte aber anhand des vorgestellten Beispiels durch die richtig bestimmten Melodieverläufe als vielversprechend gekennzeichnet werden. Eine Anzahl von moderaten zweistimmigen Bläser- und Streicherduetten wurde untersucht und ähnlich gute Transkriptionsergebnisse konnten hierfür erzielt werden. Das Verfahren soll nun in weiteren Implementierungsschritten auf die Anwendbarkeit bezüglich 3 - 4 stimmiger Kammermusik sowie einfacher Popularmusik erweitert werden.

Kapitel 6

Evaluierung

Die Beurteilung der Qualität eines Melodiextraktionssystems ist aufgrund der semantisch bedeutungsgeladenen Signalinhalte einer direkten Messung nur eingeschränkt zugänglich, da die „Richtigkeit“ der Analyseergebnisse interpretativen Unschärfen unterworfen ist. Die statistische Auswertung der umgesetzten Algorithmen im Rahmen eines Query-By-Humming-Systems (QbH-System), dessen Implementierung die Motivation dieser Arbeit darstellt, kann die auftretenden Variabilitäten in eindeutiger bewertbare Versuchsergebnisse leiten.

Bei dem in den Testreihen verwendeten System handelt es sich um die am „Fraunhofer - Institut für digitale Medientechnologie“ (IDMT) [idm04] entwickelte QbH-Software, die als integralen Bestandteil die in dieser Arbeit vorgestellte Melodieextraktion enthält.

Weiterhin werden in den am ELIS-Institut der Universität Ghent, Belgien, durchgeführten externen Untersuchungen die Transkriptionsergebnisse des eigenen sowie einer Reihe konkurrierender Verfahren mit den subjektiven Referenztranskriptionen eines musikalischen Experten verglichen.

6.1 Testumgebung „Query-by-Humming“

Die grundlegende Funktionsweise eines Qbh-Systems ist bereits in Kapitel 1 erläutert worden. Über eine benutzerfreundliche, intuitiv bedienbare Eingabe, der Anwender singt oder spielt mit einem Instrument in ein Mikrofon, wird die QbH-Anwendung mit einer Sucheingabe versehen. Das in den vorherigen Kapiteln beschriebene Extraktionsverfahren konvertiert die digitalisier-

te Schallwellenform in musikalische Noten. Nach der Abfrage der gefundenen Notenfolge in einer Melodiedatenbank werden dem Musikinteressierten die Metadaten der ähnlichsten Einträge einer vorhandenen Melodiedatenbank in Form von Liedtitel, Sänger, Komponist, etc. dargeboten.

Das vom Fraunhofer Institut implementierte QbH-System basiert auf dem MPEG7-Standard [mpe03], d.h. die gefundenen Metadaten werden in einem zum Standard konformen Format verwaltet und sind somit kompatibel zu anderen weiterführenden Anwendungen, soweit sich diese ebenfalls an die genormte Datenschnittstelle halten.

Zum Zeitpunkt der Drucklegung dieser Arbeit sind drei verschiedene Applikationen implementiert. Innerhalb eines Einzelplatzsystems findet die komplette Verarbeitung auf einem PC statt. Die zugehörige Bedienoberfläche ist in Abbildung 6.1 dargestellt. Server-basiert zeigt sich die Anwendung

Abbildung 6.1: QbH - Bedienoberfläche (Quelle: [idm04])

als Internetdienst [mus04]. Über ein „Java Applet“ erfolgt die Audioaufnahme im Browser des Benutzers. Auf einem Server werden die Daten verarbeitet und das Ergebnis auf der Benutzeroberfläche wiedergegeben. Analog dazu funktioniert die mobile Anwendung mittels eines GSM-Dienstes im wirtschaftlich wichtigen Markt der Mobilfunknetze. Hierbei kann der Nutzer in sein Mobiltelefon einsingen und erhält vom auswertenden Server in Form einer SMS („Short Message Service“) die gewünschte Information. Der bereits durchgeführte Feldtest verifiziert als Nebeneffekt die Praxis-

tauglichkeit des implementierten Melodieextraktionsverfahren. Die in Kapitel 6.1.3 aufgeführten Testergebnisse beziehen sich in diesem Zusammenhang unter anderem auf die Robustheit des Verfahrens bezüglich verschiedener Sprachübertragungsverfahren („Codecs“).

6.1.1 Datenbanken

Für die durchgeführten Tests zur Evaluierung des vorgestellten Verfahrens mit Hilfe des beschriebenen QbH-Systems wurden eine Reihe verschiedener Referenzdatenbanken aufgebaut.

In einer auf dem MIDI-Standard [MA01] basierten Suchdatenbank wurden in einem repräsentativen Querschnitt Melodien aus den Genres Klassik, Pop- und Rockmusik, inklusive aktueller Hits, zusammengefasst. Diese Datenbasis umfasst eine Kollektion von insgesamt 1024 Einträgen, aus denen schließlich zwei Testsätze kreiert wurden. Zum einen besteht die Auswahl aus der kompletten Anzahl der Titel, während im zweiten Fall eine Untergruppe von 200 Stücken verwendet wird, um Unterschiede der getesteten Verfahren in Bezug auf verschieden große Datenbanken zu dokumentieren.

Die zweite Datenbank mit den zu suchenden Anwendereingaben besteht aus einer Gesamtheit von 1152 gesungenen oder mit Instrumenten gespielten Liedern, die sich auch in den Suchdatenbanken befinden. Bei diesen Beispielen handelt es sich weitgehend um von Laien dargebotene Melodien, die im Rahmen einer Messepräsentation in lärmerfüllter Umgebung aufgenommen worden sind. Die Eingabe erfolgte über ein handelsübliches Mikrofon in „Consumer“-Qualität, welches an den Mikrofoneingang eines Notebooks angeschlossen war.

Zum Nachweis der Robustheit des Verfahrens gegenüber Verzerrungen und anderen Störungen wurden die Einträge der Anwenderdatenbank mit verschiedenen Verfahren der im Mobilfunk verwendeten GSM-Technologie kodiert und dekodiert. Zentrale Aufgabe dieser „Codecs“ ist es, mit möglichst wenig Verbrauch an Frequenzen für möglichst viele Kunden eine möglichst große Menge an Information zu übertragen. Dabei kommt es je nach Kodierungsalgorithmus zu Einbußen in der übertragenen Signalqualität. In den durchgeführten Testreihen wurden die Verfahren „Halfrate“, „Enhanced Fullrate“ und „Fullrate“ benutzt.

Mit der Verwendung von Datenbanken, deren Größenordnungen jeweils im Bereich von über 1000 Einträgen liegen, ist weiterhin die statistische Re-

levanz der vorgestellten Ergebnisse gewährleistet.

6.1.2 Dynamische Programmierung

Für eine effiziente und zuverlässige Suche der extrahierten Melodien in der aufbereiteten Suchdatenbank wurde ein Verfahren aus der Bioinformatik implementiert, das ursprünglich zur Identifikation von DNA-Strings entwickelt worden war [Wat95][Gus97]. Die wesentlichen Grundideen für die Suche nach charakteristischen Mustern in musikalischen Anwendungen werden beschrieben in der Arbeit von Hockel [Hoc02].

Die vorliegenden Melodieverläufe können aufgefasst werden als Suchstrings, deren Alphabet aus der Gesamtheit der Noteneinträge der verwendeten Zwölfton-Skala über alle Oktaven besteht. Zunächst werden die im MIDI-Format vorliegenden Sequenzen absoluter Notenhöhen transformiert in eine relative Intervalldarstellung, um Invarianz gegenüber abweichenden Stimmlagen zu gewährleisten.

Als optimales Verfahren hat sich die Verwendung des sogenannten paarweisen „Local-Alignment“ herausgestellt. Mit diesem wird ein Ähnlichkeitsmaß zwischen zwei zu vergleichenden Strings berechnet. Die benutzte Methode ist bis zu einem gewissen Grad in der Lage, Abweichungen von den in der Suchdatenbank gespeicherten Melodien zu akzeptieren. Diese ergeben sich teilweise aus musikalischen Variationen, wie z. B. durch die Verwendung von Durchgangstönen, Vorschlägen oder anderen Verzierungen. Andererseits kommen aber auch gerade bei musikalischen Laien Einsingfehler bezüglich der verwendeten Intervalle vor.

Neben den reinen Tonhöheninformationen wertet der Suchalgorithmus zusätzlich noch die vorliegenden Rhythmusinformationen aus, die sich aus den Einsatzzeitpunkten und Dauern der einzelnen Noten ergeben. Empirisch hat sich eine Gewichtung der Beiträge zum Gesamtähnlichkeitsmaß von 70 % für die Intervall- bzw. Tonhöheninformation zu 30 % Rhythmuseinfluss als sinnvoll herausgestellt.

Eine Reihe weiterer, unveröffentlichter Modifikationen, die auf musiktheoretischen Erkenntnissen basieren, wurden in das Suchverfahren integriert, können aber hier leider nicht näher erläutert werden.

Das beschriebene Verfahren wird auf alle Kombinationen der extrahierten Melodie mit den Inhalten der Suchdatenbank angewendet und als Ergebnis eine nach Ähnlichkeit sortierte Liste aufgebaut.

6.1.3 Ergebnisse

Schließlich sollen nun die Testergebnisse im Rahmen der vorgestellten QbH-Anwendung diskutiert werden. Um die erhaltenen Erkennungsraten mit denen anderer Verfahren vergleichbar zu machen, wurden eine Anzahl weiterer Melodieextraktionssysteme in der gleichen Testumgebung evaluiert. Die im folgenden mit „Extreme“ bezeichnete Anwendung wertet in erster Instanz lokale Maxima im Zeitsignal zur Detektion von Pitchverläufen aus. Es handelt sich hierbei um ein in der Fraunhofer-Arbeitsgruppe entwickeltes, undokumentiertes Alternativverfahren. Weiterhin wurde die bereits in Kapitel 2.1 vorgestellte Methode mittels Hough-Transformation [Ric01] ausgewertet. Externe Implementierungen standen mit dem „WIDI Recognition System“ [wid03] und der an der Universität Ghent am ELIS-Institut entwickelten Software [Lem02] zur Verfügung. Das in dieser Arbeit entwickelte System wird im weiteren als „EarAnalyzer“ bezeichnet. Die gefundenen Melodieverläufe wurden jeweils für alle untersuchten Verfahren dem gleichen Suchalgorithmus übergeben.

In Abbildung 6.2 sind die an Position 1 gefundenen Suchanfragen der unterschiedlichen Extraktionsverfahren für die unveränderten Originalsignale illustriert. Die Ergebnisse bezüglich der zwei verschieden großen Datenbanken sind hier einander gegenübergestellt. Analog dazu finden sich in Abbildung 6.3 die Anteile der unter den ersten 10 ähnlichsten Liedeinträgen klassifizierten Anfragen. In Tabelle 6.1 sind die zugehörigen Zahlenwerte detailliert aufgelistet.

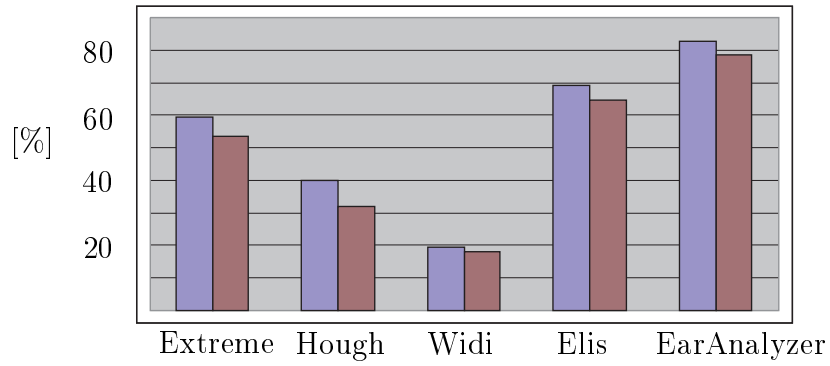

Abbildung 6.2: QbH-Testergebnisse Top 1
(Links [blau]: Datenbank 1;
Rechts [rot]: Datenbank 2)

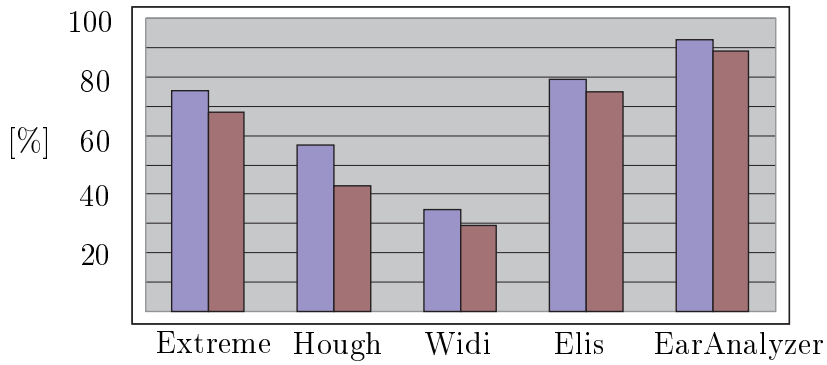

Abbildung 6.3: QbH-Testergebnisse Top 10
(Links [blau]: Datenbank 1;
Rechts [rot]: Datenbank 2)

	Datenbank 1 (200 Lieder)		Datenbank 2 (1024 Lieder)	
	Position 1 [%]	Top 10 [%]	Position 1 [%]	Top 10 [%]
Extreme	59.5	75.1	53.5	67.9
Hough	39.9	56.9	32.0	42.9
Widi	19.3	34.6	17.9	29.3
Elis	69.3	79.3	64.8	75.0
EarAnalyzer	**82.8**	**92.5**	**78.5**	**88.9**

Tabelle 6.1: QbH-Testergebnisse (1152 Suchabfragen)

Die Anzahl der richtig an Position 1 bzw. unter den ersten 10 ähnlichsten Liedeinträgen gefundenen Suchresultate gibt Aufschluss über die Qualität der untersuchten Verfahren.

Aufgrund der deutlich differierenden Zahlen erkennt man, dass die Leistung der Systeme auf stark unterschiedlichem Niveau anzusiedeln ist. Mit signifikantem Abstand erweist sich das vom Autor implementierte Verfahren

als überlegen in der Nutzung innerhalb des QbH-Testsystems. Die Leistungsunterschiede rangieren dabei in der Größenordnung von ca. 13 % gegenüber ELIS bis zu über 50 % im Vergleich mit dem WIDI-System.

Prinzipiell zeigt sich für die Anwendungen die Tendenz geringerer Erkennungsraten bei Verwendung einer größeren Datenbank. Die Ergebnisse des „EarAnalyzer"-Verfahrens verschlechtern sich in diesem Fall um ungefähr 4 %. Da für praktisch alle verglichenen Anwendungen die Einbrüche bei der Anzahl der gefundenen Melodien in einer ähnlichen Größenordnung zu finden sind, muss das Leistungsdefizit zumindest teilweise in der Datenbankabfrage begründet sein. So nimmt bei Vergrößerung des Datenbankumfangs die Selbstähnlichkeit der Inhalte zu, was wiederum Vertauschungen in den Suchresultaten zur Folge hat. Ungeachtet dieser „äußeren" Einflüsse erweist sich das auf dem Ohrmodell basierende Extraktionsverfahren als überlegen. Die außerordentlichen Erkennungsraten von ungefähr 80 % für die „Top 1"-Position bzw. ca. 90 % für die „Top 10" zeigen, dass die Implementierung auch für einen kommerziellen Einsatz ausgereift ist.

Gleichartige Ergebnisse wurden erzielt im Test mit GSM-verzerrten Originaldaten. Dazu wurden diese zur Simulation einer Mobilfunkübertragung mit verschiedenen Sprachübertragungsverfahren kodiert und dekodiert. Mit den erhaltenen Testdaten soll die Robustheit der untersuchten Implementierungen bezüglich minderer Datenqualität diskutiert werden. Bei den ver-

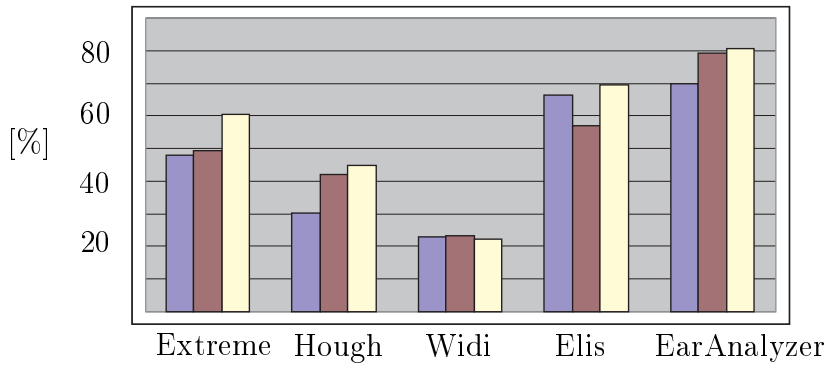

Abbildung 6.4: QbH-Testergebnisse GSM Top 1
(Links [blau]: Halfrate;
Mitte [rot]: Enhanced Fullrate;
Rechts [gelb]: Fullrate)

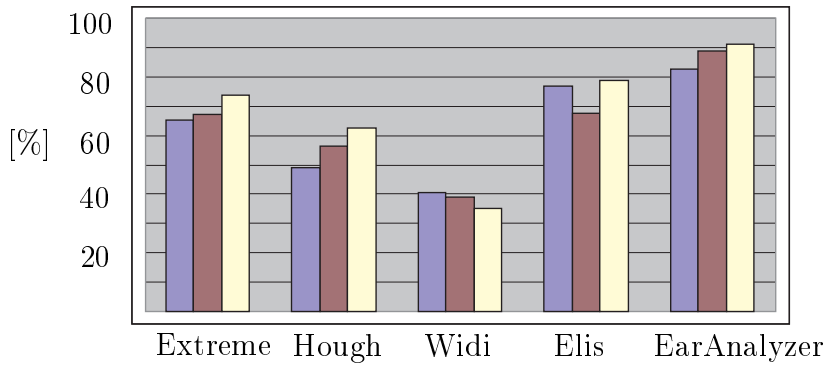

Abbildung 6.5: QbH-Testergebnisse GSM Top 10
(Links [blau]: Halfrate;
Mitte [rot]: Enhanced Fullrate;
Rechts [gelb]: Fullrate)

wendeten „Codecs“ handelt es sich um die „Halfrate“-, „Enhanced Fullrate“- und „Fullrate“-Kodierungsverfahren. Analog zur Darstellung der Ausgangsdaten sind die Testergebnisse der unterschiedlichen Algorithmen bezüglich der an Position 1 bzw. unter den „Top 10“ gefundenen Melodieanfragen in den Abbildungen 6.4 und 6.5 wiedergegeben. Tabelle 6.2 gibt Auskunft über die entsprechenden Prozentzahlen.

	Halfrate		Enhanced Fullrate		Fullrate	
	Position 1 [%]	Top 10 [%]	Position 1 [%]	Top 10 [%]	Position 1 [%]	Top 10 [%]
Extreme	48.4	65.4	49.4	67.1	60.4	73.9
Hough	30.2	49.0	42.0	56.3	45.0	62.7
Widi	22.8	40.6	23.4	39.1	22.1	35.1
Elis	66.2	76.9	57.1	67.4	69.4	78.7
EarAnalyzer	**69.9**	**82.5**	**79.3**	**88.9**	**80.6**	**91.2**

Tabelle 6.2: QbH-Testergebnisse inkl. GSM-Kodierung

Wie schon in den vorherigen Untersuchungen zeigt sich auch hier das Ohrmodell als überlegen gegenüber den anderen Anwendungen. Gerade bei der „Enhanced Fullrate“-Kodierung ergibt sich ein Abstand der erkannten Melodien von mehr als 20 % zu dem nächstbesten System (ELIS). Die Rangliste der Verfahren bleibt im Vergleich zur ersten Datenreihe erhalten. Grundsätzlich erleiden aber fast alle Methoden Einbrüche bei der Analyse GSM-kodierter Daten. Nur das WIDI-System schneidet geringfügig besser ab als in der ersten Studie. Auch die Abstände zwischen „Top 1“- und „Top 10“-Wert bleiben für die Gesamtheit der Systeme ungefähr gleich (ca. 10 % für EarAnalyzer). Speziell ergibt sich bezüglich der „Fullrate“-Kodierung für das Ohrmodell-Verfahren nur ein minimaler Rückgang der Erkennungsrate von 1 - 2 %. Während bei der „Halfrate“-Übertragung ein Verlust von mehr als 10 % auftritt, so bleibt doch die Anzahl der richtig in den „Top 10“ klassifizierten Melodien weiterhin bei über 80 %. Insgesamt erweist sich das EarAnalyzer-Verfahren somit als weitestgehend robust gegenüber den durch die verschiedenen Kodierungsmethoden hinzugefügten Verzerrungen. Diese Robustheit ist bereits implizit nachgewiesen worden durch die Benutzung der in lärmerfüllter Umgebung (Messepublikum) aufgenommenen Gesangs- und Instrumentaldatenbank.

Zusammenfassend lässt sich für die Auswertung der GSM-kodierten Signale feststellen, dass sich, aufgrund der hohen Erkennungsraten, das vorgestellte Melodieextraktionsverfahren in Kombination mit den beschriebenen Datenbankalgorithmen für eine QbH-Anwendung im Sektor der Mobilfunktechnologie als geeignet erweist. Diese Interpretation der Testreihen konnte ebenfalls bereits in umfangreichen Feldtests der in Kapitel 6.1 vorgestellten Mobiltelefonanwendung praktisch nachgewiesen werden.

Weitere Informationen über Testreihen bezüglich des Fraunhofer-Systems sind in den technischen Reports von Kátai [Kát01] und Hofmann [Hof02] beschrieben.

6.2 Referenztest

Um die Objektivität der Evaluierungsergebnisse zu erhöhen, wurde das „EarAnalyzer“-Verfahren nachträglich einer am ELIS-Institut der Universität Ghent, Belgien, durchgeführten Reihenuntersuchung zugeführt. Dort wurde eine Anzahl von insgesamt 10 Melodieextraktionsverfahren einem Corpus von 18 gesungenen Melodien gegenübergestellt. Im Vergleich zu den eigenen

Untersuchungen (1152 Melodieeingaben) stellt der Umfang des Testdatensatzes statistisch gesehen eine unzureichende Gesamtheit dar. Nichtsdestotrotz lassen sich aber gewisse Tendenzen der Testergebnisse erkennen und eine vorsichtige Aussage über die Leistungsfähigkeit der einzelnen Systeme machen.

Insgesamt umfasst die von 5 Männern und 6 Frauen eingesungene Datenbank eine Anzahl von ungefähr 300 Einzelnoten. Sieben Melodien erfolgten in textueller Form, die anderen 11 ohne Worte. Die Berechnung der Bewertungsmaße der Verfahren erfolgt über einen Vergleich mit der Referenztranskription eines erfahrenen musikalischen Experten. Dieser versieht die Wellenformen der Zeitsignale nach optischer und akustischer Auswertung unter Zuhilfenahme eines Audioeditierprogramms mit Zeitindizes, die Notenanfänge und -enden sowie Pausen markieren. Der Vergleich mit den automatischen Transkriptionssystemen erfolgt wie in der QbH-Umgebung über die Verwendung von Algorithmen aus der dynamischen Programmierung. Nach bestimmten, empirisch aufgestellten, Abweichungskriterien werden den Mitschriften der Testprogramme jeweils Prozentzahlen bezüglich der richtig oder falsch extrahierten Noten zugewiesen. Bei den untersuchten Bewertungsgrößen handelt es sich um die relativen Anteile zu viel entfernter bzw. zu wenig detektierter Noten. Weiterhin werden die im Vergleich zur Referenz korrekt bestimmten Tonhöhen im Viertel- und Halbtonabstand ausgewertet. Auch die Zeitabweichungen der Onsetangaben sowie der Notenlängen ($\Delta t < 50$ ms bzw. $\Delta t < 150$ ms) finden Berücksichtigung. Diese Längenangaben sollen aber hier nicht weiter besprochen werden, da aufgrund raumakustischer Einflüsse das Ende von Schallsignalen in der Regel verschmiert erscheint und gerade die gewählte akzeptierte Abweichung von weniger als 150 ms daher nicht sinnvoll erscheint.

Die detaillierte Beschreibung des ELIS-Verfahrens sowie der durchgeführten Testreihen ist beschrieben bei Clarisse et al. [Lem02]. Bei den weiteren untersuchten Verfahren handelt es sich um die Systeme von Akoff [ako03], AudioWorks [aud03], Autoscore [aut03], DigitalEar [dig03], Intelliscore [int03], Meldex [mel03], SoloExplorer [sol03], WIDI [wid03] sowie Haus und Pollastri [HP01]. Die genauen Testergebnisse können in den Tabellen 6.4, 6.5 und 6.6 nachgeschlagen werden.

Trotz des, wie schon erwähnt, statistisch unzureichenden Umfangs der Testdatensätze sollen nun einige grundsätzliche qualitative Tendenzen der Ergebnisse diskutiert werden. Zur Interpretation der Güte der einzelnen Systeme sollte erwähnt werden, dass die zur Evaluierung benutzten Inhalte der Datenbank nur den Entwicklerteams der ELIS- und der „Haus-Pollastri"-Software bekannt waren. Eine Optimierung dieser Verfahren auf die relativ

kleine Anzahl ausgewerteter Melodien kann zumindest nicht ausgeschlossen werden.

Zur übersichtlicheren Darstellung der untersuchten Verfahren werden diese in den nachfolgenden Schaubildern gemäß der Zuordnung von Tablle 6.3 aufgelistet.

Akoff	V1
Audioworks	V2
Autoscore	V3
Digital Ear	V4
Intelliscore	V5
Meldex	V6
Soloexplorer	V7
Widi	V8
Pollastri	V9
Elis	V10
Ear Analyzer	V11

Tabelle 6.3: Referenztest - Verfahrensindizes

In Abbildung 6.6 sind die Anteile der richtig zugeordneten Notengrundfrequenzen abgebildet. Der jeweils linke (blaue) Balken repräsentiert die innerhalb eines Viertelton-Intervalls erhaltene Erfolgsquote (bezogen auf die Referenztranskription des musikalischen Experten). Beim rechten (roten) Balken ist die Nebenbedingung des akzeptierten Intervalls auf Halbtonbreite erweitert worden. Die Abbildungen 6.7 und 6.8 veranschaulichen den gleichen Sachverhalt getrennt nach ohne und mit Text gesungenen Melodien. Eine bemerkenswerte Diskrepanz ergibt sich bei den Ergebnissen des physiologischen Ohrmodells (V11). Während der Algorithmus bei der Viertelton-Bedingung nur an vorletzter Stelle rangiert, liegt er bei der Halbton-Betrachtung auf der Spitzenposition.

Betrachtet man die typischen Pitchverläufe gesungener Eingaben, die in der beschriebenen Untersuchung den Testdatenvektor aufspannen, so erkennt man, dass die Beschränkung auf ein Viertelton-Intervall nicht unbedingt sinnvoll ist (vergleiche hierzu Abbildung 5.7). Vielmehr finden sich stark ausgeprägte Tonhöhenschwankungen, die bei der Interpretation der gefundenen Frequenzinformationen berücksichtigt werden sollten.

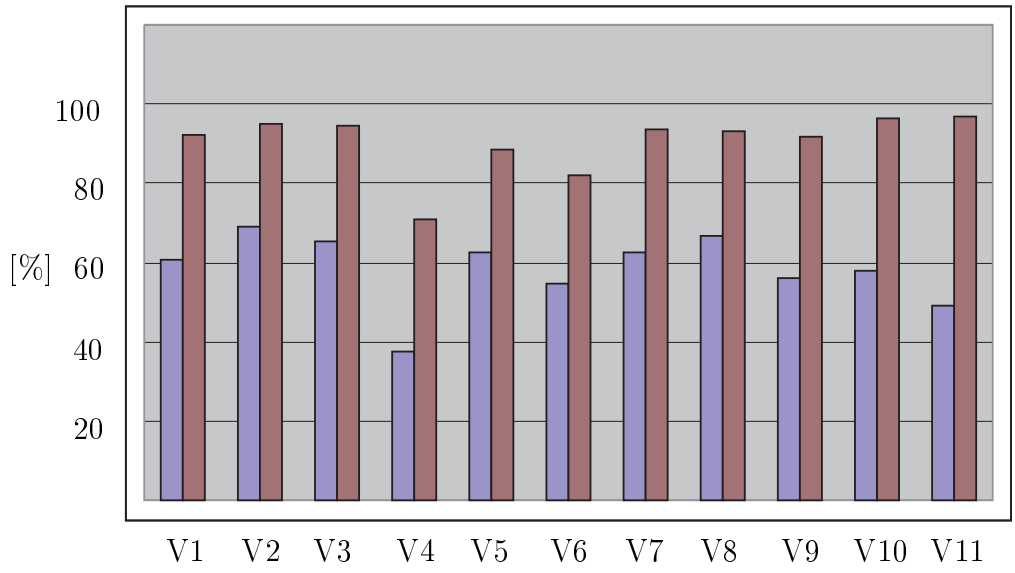

Abbildung 6.6: Referenztest - F0 - Gesamtergebnis
(Links [blau]: $F_0 \leq \frac{1}{4}HT$;
Rechts [rot]: $F_0 \leq HT$)

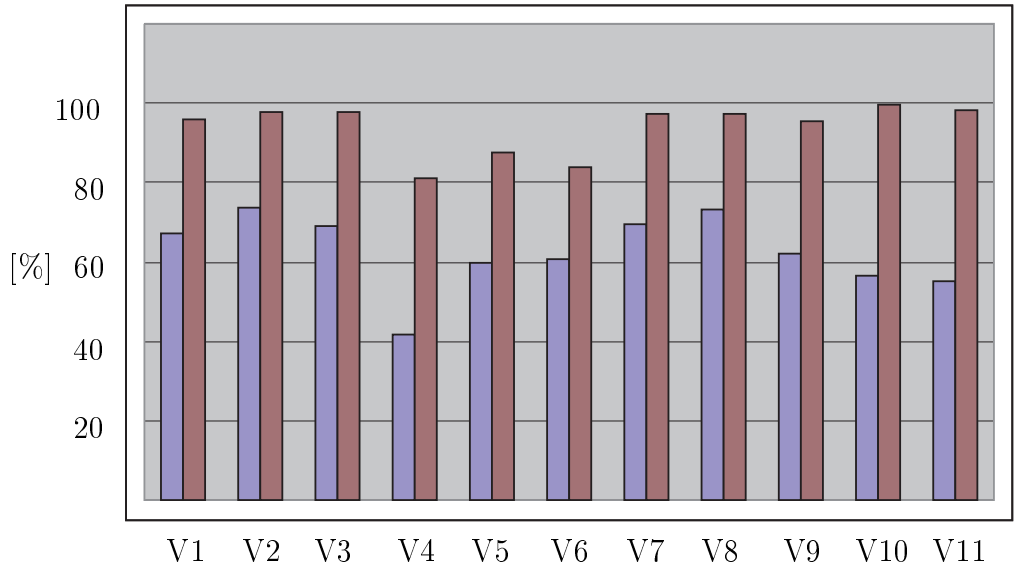

Abbildung 6.7: Referenztest - F0 - ohne Text
(Links [blau]: $F_0 \leq \frac{1}{4}HT$;
Rechts [rot]: $F_0 \leq HT$)

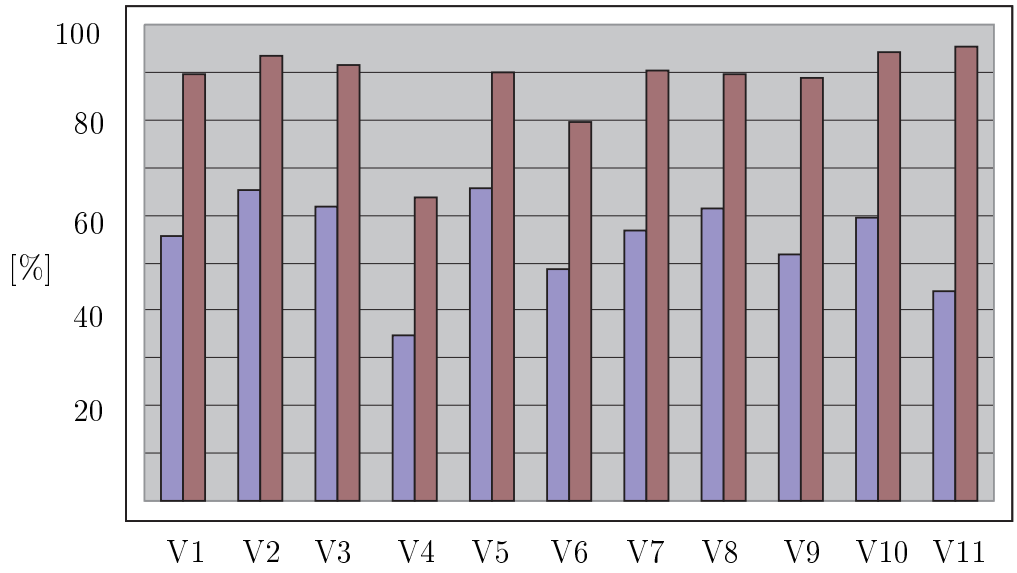

Abbildung 6.8: Referenztest - F0 - mit Text
(Links [blau]: $F_0 \leq \frac{1}{4}HT$;
Rechts [rot]: $F_0 \leq HT$)

Die Parameter des „EarAnalyzer“-Verfahren garantieren somit eine zuverlässige Bestimmung der vorhandenen Tonhöhen im Rahmen einer (bezüglich Gesangseingaben) angemessenen Genauigkeit.

Insgesamt lässt sich der überwiegenden Mehrzahl der Verfahren eine akzeptable Qualität bei der Pitchextraktion zuerkennen. Diese steht allerdings bei den meisten Algorithmen in Kontrast zur bereitgestellten Segmentierungsleistung. Zur Verdeutlichung dieser Aussage sind in Abbildung 6.9 die für die Segmentierung maßgeblichen Parameter veranschaulicht. Der jeweils gelbe (rechte) Balken repräsentiert die innerhalb eines Zeitintervalls von 150 ms korrekt erkannten Notenanfänge. Der mittlere (rote) Balken illustriert die Anzahl der im Original nicht vorhandenen, von den Verfahren irrtümlich eingefügten Onsets. Die entsprechend nicht gefundenen Noten finden sich in den linken (blauen) Balken. Wie bei der Betrachtung der Frequenzparameter sind in den Abbildungen 6.10 und 6.11 die Ergebnisse aufgespalten in nach ohne bzw. mit Text dargebotenen Eingaben.

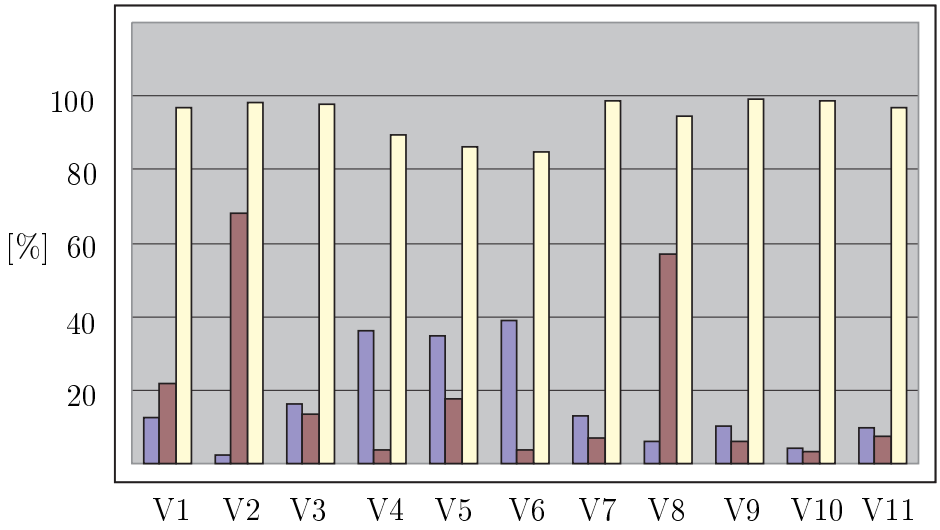

Abbildung 6.9: Referenztest - Gesamtergebnis
(Links [blau]: Entfernte Noten;
Mitte [rot]: Eingefügte Noten;
Rechts [gelb]: Onsets)

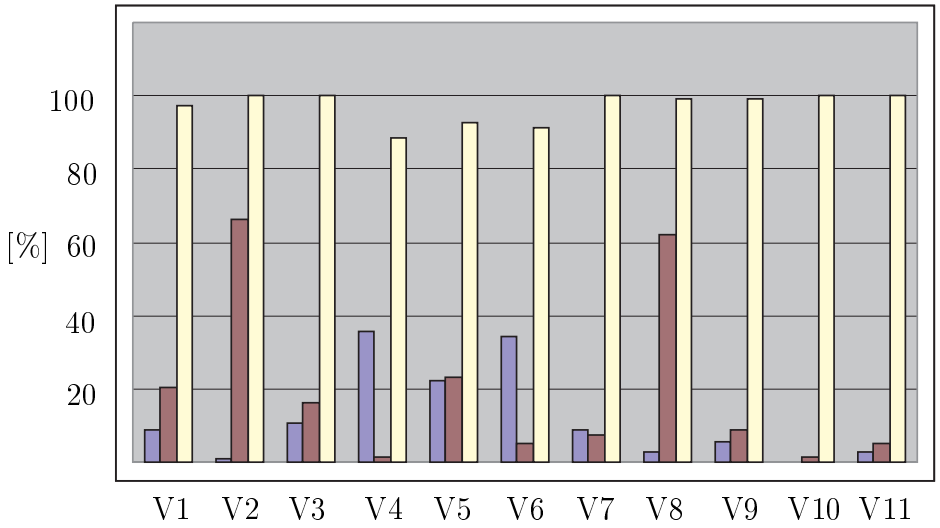

Abbildung 6.10: Referenztest - ohne Text
(Links [blau]: Entfernte Noten;
Mitte [rot]: Eingefügte Noten;
Rechts [gelb]: Onsets)

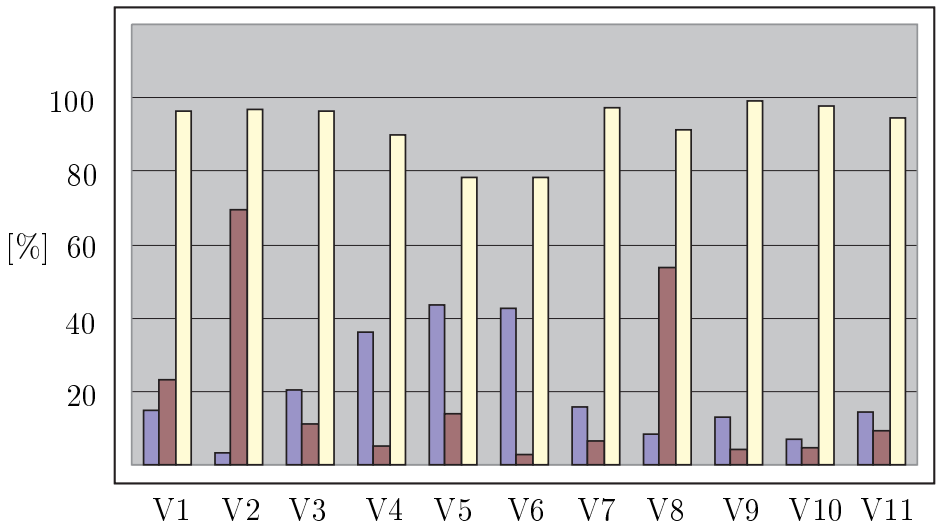

Abbildung 6.11: Referenztest - mit Text
(Links [blau]: Entfernte Noten;
Mitte [rot]: Eingefügte Noten;
Rechts [gelb]: Onsets)

Aus den Grafiken erkennt man, dass die meisten Verfahren innerhalb der gewählten Zeittoleranz einen hohen Prozentsatz der tatsächlich vorhandenen Notenanfänge korrekt analysieren. Dieser Wert steht aber in direkter Verbindung zu den richtig bzw. zu viel detektierten Noten. So erkennen die Programme von Audioworks (V2) und Widi (V8) zwar den Großteil der vorhandenen Onsets, fügen aber so viele zusätzliche Notenanfänge ein, sodass ein beträchtliches Missverhältnis zwischen den tatsächlich vorhandenen und irrtümlich vermuteten Noten entsteht. Gegensätzliche Tendenz zeigen die Verfahren Digital Ear (V4), Intelliscore (V5) und Meldex (V6), die eine große Zahl der vorhandenen Onsets nicht finden. Dies deutet auf eine unzureichende Zeitauflösung hin. Obwohl die Ergebnisse bezüglich der „150 ms“-Bedingung überzeugen können, so sind offensichtlich mehrfache Onsets, die innerhalb dieses Zeitintervalls auftreten, für diese Algorithmen nicht auflösbar.

Die „EarAnalyzer“-Software (V11) schneidet bei diesem speziellen Kriterium wie auch bei der Auswertung über den gesamten Datenraum (s. Tabellen 6.4, 6.5 und 6.6) neben den Systemen von ELIS (V10) sowie Haus und Pollastri (V9) als zuverlässigster Vertreter ab und stellt somit das beste System

dar, dem die Inhalte der Datenbank nicht bekannt waren.

Vergleicht man zusätzlich jeweils die Abbildungen 6.7 und 6.8 bzw. 6.10 und 6.11, so lässt sich aus den Daten schlussfolgern, dass bei Gesang die Verwendung von Worten zur Verschlechterung der Transkriptionsergebnisse führt. Unter anderem ist dies zurückzuführen auf die in diesem Fall deutliche Verkürzung der stimmhaften Anteile, die die Detektion der vorhandenen Pitchsequenzen wesentlich erschwert.

Zusammenfassend lässt sich feststellen, dass die vom ELIS-Institut durchgeführte Testreihe einige Informationen bereitstellt, aber wegen der erläuterten statistischen Probleme nicht überbewertet werden sollte. Bei vielen Verfahren ist die korrekte Segmentierung der Zeitverläufe das offensichtlich vordergründige Problem, während die Erkennung der richtigen Tonhöhen für monophone Daten im wesentlichen besser umgesetzt werden kann. Das diesbezüglich implementierte Segmentierungsverfahren des Ohrmodells kann auch bei der Unterteilung der Pitchtrajektorien überzeugen. Aus Sicht des Autors ist dies zurückzuführen auf die im Innenohr bereitgestellte optimale Kombination aus Zeit- und Frequenzauflösung. Das vom EarAnalyzer-Programm erzielte gute Ergebnis ist umso positiver zu bewerten, da die Segmentierung des Ohrmodells optimiert ist auf die Detektion der stimmhaften Anteile. Die stimmlosen Anteile, die zur melodiebestimmenden Pitchwirkung nicht beitragen, werden bewusst verworfen.

In Tabelle 6.4 findet sich die Gesamtheit der getesten Parameter als über alle Eingaben gemittelte Prozentzahlen. Die fett gedruckten Zahlen repräsentieren das im jeweiligen Fall optimale Verfahren. Die Tabellen 6.5 und 6.6 geben die Ergebnisse aufgespalten in Melodien mit bzw. ohne Wortinhalte wieder.

Als Fazit der vorgestellten eigenen und externen Testreihen ist das implementierte automatisierte Melodietranskriptionsverfahren als leistungsfähig und zuverlässig evaluiert. Sowohl die absoluten Erkennungsraten als auch der Vergleich mit anderen konkurrierenden Anwendungen verdeutlichen ebenso wie die praktischen Erfahrungen innerhalb des Fraunhofer-QbH-Systems das immense Potential des Verfahrens auch im Hinblick auf kommerzielle Anwendungen.

	Entf. Noten [%]	Eing. Noten [%]	F_0 ($\leq \frac{1}{4}HT$) [%]	F_0 ($\leq HT$) [%]	Onsets ($\leq 0,05s$) [%]	Onsets ($\leq 0,15s$) [%]	Länge ($\leq 0,05s$) [%]	Länge ($\leq 0,15s$) [%]
Akoff	12,3	21,8	60,5	92,3	87,7	96,9	57,9	74,7
Audioworks	**2,2**	68,1	**68,9**	95,2	88,1	98,3	44,0	61,4
Autoscore	16,1	13,3	65,1	94,4	88,8	97,6	61,0	77,5
Digital Ear	36,0	3,5	37,6	71,0	78,0	89,3	49,5	64,0
Intelliscore	34,7	17,7	62,6	88,7	52,6	86,0	20,5	43,2
Meldex	39,1	3,8	54,7	81,8	45,9	84,9	23,9	54,1
Soloexplorer	12,9	6,9	62,6	93,4	84,9	98,5	63,7	81,9
Widi	6,0	57,1	66,6	92,9	70,1	94,7	31,3	61,9
Pollastri	10,0	6,2	56,1	91,8	90,0	**99,2**	65,7	84,0
Elis	4,0	**3,2**	58,0	96,6	**90,3**	98,8	**80,8**	**93,3**
EarAnalyzer	9,5	7,6	49,3	**96,7**	81,1	97,0	67,4	82,6

Tabelle 6.4: Referenztest - Gesamtergebnis

	Entf. Noten [%]	Eing. Noten [%]	F_0 ($\leq \frac{1}{4}HT$) [%]	F_0 ($\leq HT$) [%]	Onsets ($\leq 0,05s$) [%]	Onsets ($\leq 0,15s$) [%]	Länge ($\leq 0,05s$) [%]	Länge ($\leq 0,15s$) [%]
Akoff	9,0	20,2	67,0	96,1	92,2	97,4	66,1	77,4
Audioworks	0,8	66,4	**73,8**	97,6	96,8	**100,0**	52,4	64,3
Autoscore	10,5	16,4	69,0	97,8	99,1	**100,0**	73,5	81,4
Digital Ear	35,8	1,5	41,8	81,0	84,8	88,6	62,0	70,9
Intelliscore	22,4	23,1	59,8	87,6	53,6	92,8	20,6	51,6
Meldex	34,3	5,2	60,5	84,0	46,9	91,4	32,1	66,7
Soloexplorer	9,0	7,5	69,6	97,4	95,8	**100,0**	80,9	88,7
Widi	3,0	61,9	73,2	97,2	80,5	99,2	43,9	66,7
Pollastri	5,6	8,7	62,0	95,6	99,1	99,1	83,2	85,8
Elis	**0,0**	**1,3**	56,7	**99,6**	**99,5**	**100,0**	**97,4**	**99,2**
EarAnalyzer	3,0	5,2	55,3	98,4	87,0	**100,0**	82,1	90,2

Tabelle 6.5: Referenztest - ohne Text

	Entf. Noten [%]	Eing. Noten [%]	F_0 ($\leq \frac{1}{4}HT$) [%]	F_0 ($\leq HT$) [%]	Onsets ($\leq 0,05s$) [%]	Onsets ($\leq 0,15s$) [%]	Länge ($\leq 0,05s$) [%]	Länge ($\leq 0,15s$) [%]
Akoff	14,8	23,0	55,5	89,4	**84,3**	96,6	51,4	72,6
Audioworks	**3,3**	69,4	65,3	93,4	81,4	97,0	37,7	59,3
Autoscore	20,2	10,9	61,8	91,5	80,2	96,6	50,7	74,3
Digital Ear	36,1	4,9	34,6	63,6	72,9	89,7	40,2	58,9
Intelliscore	43,7	13,7	**65,6**	89,8	51,6	78,5	20,4	34,4
Meldex	42,6	2,7	48,7	79,5	44,9	78,2	15,4	41,0
Soloexplorer	15,9	6,6	56,9	90,3	77,1	97,2	50,0	76,4
Widi	8,2	53,6	61,4	89,6	62,0	91,1	21,5	58,2
Pollastri	13,1	**4,4**	51,7	88,9	83,2	**99,3**	52,4	82,6
Elis	6,9	4,6	59,3	94,2	82,9	97,9	**67,6**	**88,6**
EarAnalyzer	14,2	9,3	44,2	**95,2**	76,2	94,6	55,1	76,2

Tabelle 6.6: Referenztest - mit Text

Kapitel 7

Zusammenfassung und Ausblick

Stetig wachsende multimediale Datenmengen sorgen für einen Bedarf an effizienten und intuitiven Eingabemethoden zur Indexierung der gewünschten Informationen. Motivation für die Durchführung der vorliegenden Arbeit war die Implementierung eines Query-By-Humming-Systems zur Suche eingesungener oder instrumental dargebotener Melodien in Metadatenbanken.

Das vorgestellte Verfahren dient der Transkription melodischer Inhalte in abstrakte musikalische Notendarstellungen. Diese können nachfolgend intelligenten Suchalgorithmen zugeführt und auf ihre Ähnlichkeit bezüglich der vorhandenen Datenbankinhalte untersucht werden.

Die im ersten Teil der Arbeit diskutierten aktuellen Verfahren der semantischen Audioanalyse nähern sich über eine Reihe unterschiedlich komplexer Ansätze der Thematik. Sowohl Pitchextraktions- als auch Segmentierungsalgorithmen können entweder direkt in Zeit- bzw. Frequenzbereich angesiedelt werden oder ergeben sich ebenso wie das eigene Modell als Kombination der beiden Signaldomänen. Gesamtsysteme zur monophonen Melodietranskription zeigen vielversprechende, aber häufig zu eng gefasste, Ansätze. Insbesondere bei den Verfahren zur Analyse polyphoner Musik gelingt es nicht, zuverlässige bzw. praxistaugliche Ergebnisse bereitzustellen. Die von den besprochenen Anwendungen eingeführten Randbedingungen erweisen sich als zu limitiert, um ausreichende Allgemeingültigkeit zu gewährleisten.

In wissenschaftlichen und kommerziellen Applikationen erwächst zunehmend die Tendenz, auf die in technischen Anwendungen bisher unerreichte menschliche Fähigkeit der Umweltwahrnehmung aufzubauen. Im Audio-

Sektor basiert dies derzeit hauptsächlich auf psychoakustischen Erkenntnissen.

Ziel dieser Arbeit ist es, so weit wie möglich, konsequent die Ergebnisse *physiologischer* Messungen auszunutzen, um sich explizit an den Analyseeigenschaften der menschlichen auditorischen Peripherie zu orientieren. Hierbei sollen die durch die Evolutionsgeschichte optimierten Gegebenheiten zur Reduktion auf die für den Menschen relevanten Schallanteile genutzt werden.

Da in den kognitiven Wahrnehmungsebenen die physiologischen Vorgänge noch nicht ausreichend untersucht sind, werden zusätzlich psychoakustische und gestaltpsychologische Modelle zur Weiterverarbeitung der Informationen verwendet.

Zur Nachbildung des komplexen psychophysiologischen Wahrnehmungsapparates des Menschen werden eine Reihe sich ergänzender Erklärungsmodelle miteinander kombiniert. Die auditorische Peripherie (Außen-, Mittel- und Innenohr) wird als zentrales Element der Schallvorverarbeitung durch experimentell bestätigte, ausschließlich physiologisch basierte, Ansätze implementiert. Das „erweiterte Analogmodell“ beschreibt die hydromechanischen Vorgänge der Cochlea, aufgrund derer die Frequenzanteile der Schallsignale an charakteristischen Abschnitten der Basilarmembran in ortsabhängige Resonanzschwingungen transformiert werden. Die Transduktion der mechanischen Bewegungsabläufe des Innenohres in elektrische Impulse auf benachbarten Nervenfaserpopulationen wird mittels des Modells der inneren Haarzellen von Meddis beschrieben. Die hiermit verbundene Zeitkodierung der neuronalen Aktionspotentialmuster wird über den Ansatz des „Phase-Locking“ zur Frequenzanalyse herangezogen.
Die nachfolgenden Perzeptionsschritte werden ausschließlich auf psychoakustischen und psychologischen Modellen basiert. Bei der Segmentierung zeigt sich die Notwendigkeit der Einführung von Vor- und Nachverdeckungseffekten in Kombination mit gehörgerechter Amplitudenbewertung („Weber-Bruch“).
Um die implementierte Applikation auf die Verwendung innerhalb eines QbH-Systems zu optimieren, kommen im Sinne einer interpretativen Nachbearbeitung eine Anzahl musiktheoretischer und kulturell bedingter Vorgaben zur Anwendung.
Die gestaltpsychologisch fundierte Aufstellung eines eigenen Hierarchiemodells zur Analyse polyphoner Musik beschreibt die in kognitiven Prozessen zunehmend abstraktere teils hypothetische Fusion einzelner perzeptionsbe-

stimmender Elemente. Die Kompatibilität zur monophonen Herangehensweise bleibt erhalten.

Der hohe Implementierungsaufwand des gewählten Transkriptionsverfahrens erweist sich aufgrund der guten Testergebnisse als gerechtfertigt.

Im Rahmen des die Arbeit motivierenden QbH-Systems kann das vom Autor umgesetzte *physiologische* Modell gegenüber anderen Verfahren als überlegen eingestuft werden. Die Erkennungsraten der richtig einsortierten Suchmelodien liegen signifikant über denen der untersuchten Mitbewerber. Die Tauglichkeit für kommerzielle Applikationen (auch in Mobilfunkanwendungen) kann nachgewiesen werden.

Die wesentliche Aussage bezüglich der Qualität der Verfahren zur automatisierten Transkription musikalischer Inhalte ergibt sich aus dem Vergleich der notwendigen Analyseschritte. Während die Mehrheit der Verfahren eine alles in allem zufriedenstellende Pitchextraktion zur Verfügung stellen kann, so findet sich häufig eine mangelhafte Segmentierung der Pitchverläufe in Einzelnoten. Diesbezüglich stellt das ausgezeichnete Verhältnis der implizit vom Ohrmodell bereitgestellten Frequenz-Zeit-Auflösung eine zuverlässige Unterteilung in musikalisch logische Abschnitte bereit.

In Rahmen der polyphonen Untersuchungen werden grundlegende Strategien gefunden, die Ansätze zur musikalischen auditiven Szenenanalyse bereitstellen, ohne die Allgemeingültigkeit aufgrund zu eng gesetzter Rahmendaten zu gefährden. So erlaubt die Untersuchung von Partialtoninterferenzen als typische Schwingungsmuster an charakteristischen Positionen der Basilarmembran Aussagen zur Präsenz primär spektral verdeckter Teiltonfrequenzen. Weiterhin gibt die konsequente Anwendung der Gestaltgesetze Hinweise auf die Herangehensweise zur Unterscheidung von oktavverwandten Noten und Partialtoneinträgen. Die von Bregman [Bre90] eingeführten kognitiven Vorgänge zur sequentiellen Integration werden benutzt, um aus den analysierten Noten auditorische Ströme bzw. im vorliegenden Fall parallel verlaufende Melodielinien zu extrahieren.

Das gewählte physiologische Ohrmodell stellt eine leistungsfähige Nachbildung der auditorischen Peripherie dar. Ebenso weisen die kognitiven Nachverarbeitungsschritte eine Reihe erfolgversprechender Resultate vor. Der komplexe Wahrnehmungsapparat des Menschen bietet aber für die Verallgemeinerung der Ergebnisse, insbesondere bezüglich polyphoner musikalischer Inhalte, Potential zur weiteren Optimierung des Verfahrens.

Die Hinzunahme der Frequenz-Orts-Theorie, wonach bestimmte charakteristische Frequenzen mehr oder weniger eindeutigen Orten auf der Basilarmembran zugeordnet werden können, sollte die Detektion hoher Partialtöne weiter verbessern, und zu einer noch robusteren Pitchextraktion führen.

Die Auswertung klangfarbenspezifischer Eigenschaften von Schallquellen stellt einen wichtigen Faktor der menschlichen Perzeption der akustischen Umwelt dar und kann bei der auditiven Szenenanalyse einen wesentlichen Beitrag zur Verbesserung der Extraktion auditorischer Ströme bereitstellen.

In aktuellen Untersuchungen zur Implementierung einer Phonemerkennung wird auf die Verbesserung der Erkennungsraten innerhalb des QbH-Systems abgezielt. Bei gesungenen Eingaben soll im Falle textueller Suchanfragen die Sprachinformation genutzt werden.

Abkürzungen

AGC	**A**utomatic **G**ain **C**ontrol
AKF	**A**uto**k**orrelations**f**unktion
AMDF	**A**verage **M**agnitude **D**ifference **F**unction
AO	**A**ußen**o**hr
BM	**B**asilar**m**embran
BU	**B**ottom-**U**p
CD	**C**ompact **D**isc
CMDF	**C**umulative **M**ean **D**ifference **F**unction
CO	**C**ortisches **O**rgan
CODEC	**Co**der/**Dec**oder
CT	**C**ochleäre **T**rennwand
DB	**D**ezi**b**el
DOF	**D**egree-of-**O**nset-**F**unction
DFT	**D**iskrete **F**ourier-**T**ransformation
EM	**E**xpectation-**M**aximization
ESRPD	**E**nhanced **S**uper-**R**esolution **P**itch **D**eterminator
FIS	**F**uzzy **I**nference **S**ystem
GSM	**G**lobal **S**ystem for **M**obile Communications
HC	**H**eli**c**otrema
HFC	**H**igh **F**requency **C**ontent
HFCF	**H**igh **F**requency **C**ontent Detection **F**unction
HM	**H**ydro**m**echanik
HMM	**H**idden-**M**arkov-**M**odell
HRTF	**H**ead **R**elated **T**ransfer **F**unction
HZ	**H**ert**z**
IF	**I**nstantaneous **F**requency
IO	**I**nnen**o**hr
IOI	**I**nter-**O**nset-**I**nterval

JND	**J**ust-**N**oticable-**D**ifference
KHZ	**K**ilo**h**ert**z**
KS	**K**nowledge **S**ource
LSF	**L**east-**S**quare-**F**itting
LOO	**L**eave-**O**ne-**O**ut
MIDI	**M**usical **I**nstrument **D**igital **I**nterface
MO	**M**ittel**o**hr
MPEG	**M**oving **P**icture **E**xperts **G**roup
MS	**M**illi**s**ekunde
NAP	**N**erven**a**ktions**p**otential
NN	**N**euronales **N**etz
ODFT	**O**dd **DFT**
OF	**O**vales **F**enster
PDF	**P**robability **D**ensity **F**unction
PL	**P**hase-**L**ocking
QBH	**Q**uery-**B**y-**H**umming
RM	**R**eissnersche **M**embran
RMS	**R**oot-**M**ean-**S**quare
S	**S**ekunde
SACF	**S**um **A**uto**c**orrelation **F**unction
SC	**S**pectral **C**entroid
SF	**S**pectral **F**lux
SI	**S**equentielle **I**ntegration
SMS	**S**hort-**M**essage-**S**ervice
SNR	**S**ignal-to-**N**oise-**R**atio
SPL	**S**ound **P**ressure **L**evel
ST	**S**cala **T**ympani
STFT	**S**hort **T**ime **F**ourier **T**ransform
SV	**S**cala **V**estibuli
TD	**T**op-**D**own
TM	**T**ektorial**m**embran
TWM	**T**wo-**W**ay-**M**ismatch
VC	**V**oicing-**C**oefficent
WDF	**W**ellen**d**igital**f**ilter
ZCR	**Z**ero-**C**rossing-**R**ate

Literaturverzeichnis

[ako03] *„Akoff music composer 2.0“. Akoff Sound Lab.* http://www.akoff.com, 2003.

[AKZ02] Arfib, D., F. Keiler und U. Zölzer: *DAFX-Digital audio effects.* John Wiley & Sons, 2002.

[And01] Anderson, J.R.: *Kognitive Psychologie.* Spektrum Akademischer Verlag, 3. Auflage, 2001.

[aud03] *„Audioworks 2.15“.* http://www.audioworks.com, 2003.

[aut03] *„Autoscore Deluxe 2.0“. Wildcat Canyon Software.* http://www.wildcat.com, 2003.

[Bag94] Bagshaw, P.: *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching.* Doktorarbeit, University of Edinburgh, 1994.

[Bau95] Baumann, U.: *Ein Verfahren zur Erkennung und Trennung multipler akustischer Objekte.* Doktorarbeit, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 1995.

[Bau00] Baumgarte, F.: *Ein psychophysiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung.* Doktorarbeit, Universität Hannover, 2000.

[BCVA91] Barnard, E., R.A. Cole, M.P. Vea und F.A. Alleva: *Pitch detection with a neural-net classifier.* IEEE Trans. Sig. Proc., 39:298–307, 1991.

[BDS02] Bello, J.P., L. Daudet und M. Sandler: *Time-domain Polyphonic Transcription using Self-Generating Databases.* AES 112th Convention, München, Deutschland, 2002.

[Beg94] Begault, D.R.: *3-D Sound for Virtual Reality and Multimedia.* Morgan Kaufmann Pub, 1994.

[Ben90] Benade, A.H.: *Fundamentals of Musical Acoustics.* Dover edition, 1990.

[BFRR95] Bräunl, Th., St. Feyer, W. Rapf und M. Reinhardt: *Parallele Bildverarbeitung.* Addison-Wesley, Bonn, 1995.

[Bla96] Blauert, J.: *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1. Auflage, 1996.

[BMS00] Bello, J., G. Monti und M. Sandler: *Techniques for Automatic Music Transcription.* International Symposium on Music Information Retrieval, 2000.

[BN93a] Basseville, M. und I.V. Nikiforov: *Detection of abrupt changes.* PTR Prentice-Hall, 1993.

[Bün93b] Bünting, K.-D.: *Einführung in die Linguistik.* Frankfurt: Anton Hain, 14. Auflage, 1993.

[BP89] Brown, J.C. und M.S. Puckette: *Calculation of a Narrowed Autocorrelation Function.* J. Acoust. Soc. Am., 85:1595–1601, 1989.

[BP92] Brown, J.C. und M.S. Puckette: *An efficient algorithm for the calculation of a constant Q transform.* J. Acoust. Soc. Am., 92(5):2698–2701, 1992.

[BP93] Brown, J.C. und M.S. Puckette: *A high resolution fundamental frequency determination based on phase changes of the fourier transform.* J. Acoust. Soc. Am., 94(2):662–667, 1993.

[Bra99] Brandenburg, K.: *MP3 and AAC explained.* AES 17th International Conference on High Quality Audio Coding, Florenz, Italien, 1999.

[Bre90] Bregman, A.: *Auditory Scene Analysis.* MIT Press, 1990.

[Bro92] Brown, J.C.: *Musical Fundamental Frequency Tracking Using a Pattern Recognition Method.* J. Acoust. Soc. Am., 92(3):1394–1402, 1992.

[BS00] Bello, J.P. und M. Sandler: *Blackboard System and Top-Down Processing for the Transcription of Simple Polyphonic Music.* COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italien, 2000.

[But98] Butz, T.: *Fouriertransformation für Fussgänger.* B.G. Teubner Stuttgart, 1998.

[BZ91] Brown, J.C. und B. Zhang: *Musical Frequency tracking using the methods of conventional and „narrowed" autocorrelation.* J. Acoust. Soc. Am., 89(5):2346–2354, 1991.

[Can98] Cano, P.: *Fundamental Frequency Estimation in the SMS Analysis.* Proceedings of the Digital Audio Effects Workshop (DAFX98), 1998.

[Car02] Carré, M.: *Systèmes de Recherche de Documents Musicaux par Chantonnement.* Doktorarbeit, Ecole Nationale Supérieure des Télécommunications, 2002.

[Cha01] Chai, W.: *Melody Retrieval On The Web.* Diplomarbeit, Massachusets Institute of Technology, September 2001.

[Cho97] Choi, A.: *Real-Time Fundamental Frequency Estimation by Least-Square Fitting.* IEEE Transactions on Speech and Audio Processing, 5(2):201–205, 1997.

[cso03] *„CSound".* http://www.csounds.com, 2003.

[Dau96] Daubechies, I.: *Where do wavelets come from? - a personal point of view.* In Proceedings of the IEEE, Wavelets, 84(4):510–513, 1996.

[DB00] Demuth, H. und M. Beale: *Neural network toolbox user's guide. Version 4.* The mathwork Inc., 2000.

[DBS77] Davis, R., B. Buchanan und E. Shortliffe: *Production Rules as a Representation for a Knowledge-Based Consultation Program.* Artificial Intelligence, 8:15–45, 1977.

[dC02] Cheveigné, A. de: *YIN, a fundamental frequency estimator for speech and music.* J. Acoust. Soc. Am., 111(4):1917–1930, 2002.

[dCK02] Cheveigné, A. de und H. Kawahara: *Multiple period estimation and pitch perception model.* Speech Communication, 27:175–185, 2002.

[DDS01] Duxburry, C., M. Davies und M. Sandler: *Separation of transient information in musical audio using multiresolution analysis techniques.* In Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Irland, 2001.

[DH97] Depalle, Ph. und T. Hélie: *Extraction of Spectral Peak Parameters Using a Short-time Fourier transform and no Sidelobe Windows.* IEEE 97 Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 1997.

[DH00] Duda, R. und P. Hart: *Pattern classification and Scene Analysis.* John Wiley and sons, 2. Auflage, 2000.

[dig03] *„Digital Ear". Epinoisis Software.* http://www.digital-ear.com, 2003.

[Dix01] Dixon, S.: *Learning to Detect Onsets of Acoustic Piano Tones.* MOSART Workshop on Current Research Directions in Computer Music, Barcelona, Spanien, November 2001.

[DLR77] Dempster, A.P., N.M. Laird und D.B. Rubin: *Maximum likelihood from incomplete data via the EM algorithm.* J. Roy. Stat. Soc. B, 39(1):1–38, 1977.

[DPF96] Dallos, P., A. Popper und R. Fay: *The Cochlea.* Springer, 1996.

[DR93] Doval, B. und X. Rodet: *Fundamental Frequency Estimation and Tracking Using Maximum Likelihood Harmonic Matching and HMMs.* Proceedings of ICASSP, 1:221–224, 1993.

[DRS01] Dudel, J., R.Menzel und R. Schmidt: *Neurowissenschaft.* Springer, 2. Auflage, 2001.

[DS94] Deetjen, P. und E. Speckmann: *Physiologie.* Urban & Schwarzenberg, 2. Auflage, 1994.

[Eic03] Eichler, B.: *Analyse existierender Sprechererkennungsalgorithmen und Adaption auf ein Cochlea-Modell zur Sängeridentifizierung.* Diplomarbeit, Technische Universität Ilmenau, Deutschland, 2003.

[Ell96] Ellis, D.P.W.: *Prediction-driven computational auditory scene analysis.* Doktorarbeit, M.I.T., Cambridge, MA, 1996.

[Fel84] Fellbaum, K.: *Sprachverarbeitung und Sprachübertragung.* Springer-Verlag, 1984.

[Fer98] Ferreira, A.J.S.: *Spectral Coding and Post-Processing of High Quality Audio.* Doktorarbeit, Faculdade de Engenharia da Universidade do Porto, Portugal, 1998.

[Fet86] Fettweis, A.: *Wave digital filters: theory and practice.* Proc. IEEE, 74:207–327, 1986.

[FR98] Fletcher, N.H. und T.D. Rossing: *The Physics of Musical Instruments.* Springer-Verlag, 2. Auflage, 1998.

[FS99] Fitch, J. und F. Shabana: *A Wavelet-Based Pitch Detector For Musical Signals.* Proc. 2nd COSTG6 Workshop on Digital Audio Effects (DAFx99), 1999.

[GH99] Goto, M. und S. Hayamizu: *A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals.* Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis, Seiten 31–40, 1999.

[GLCS95] Ghias, A., J. Logan, D. Chamberlin und B.C. Smith: *Query By Humming - Musical Information Retrieval in An Audio Database.* Proc. of ACM Multimedia'95, San Francisco, Kalifornien, USA, November, 1995.

[Gol97] Goldstein, E.B.: *Wahrnehmungspsychologie.* Spektrum Akademischer Verlag, 1997.

[Got00] Goto, M.: *A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings.* Proc. IEEE International Conference on Acoust., Speech and Signal Processing, Istanbul, Türkei, 2000.

[Got01a] Goto, M.: *An Audio-based Real-time Beat Tracking System for Music With or Without Drum-Sounds.* Journal of New Music Research, 30(2):159–171, 2001.

[Got01b] Goto, M.: *A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models.* IEEE International Conference on Acoustics, Speech and Signal Processing, 5:3365–3368, 2001.

[GR69] Gold, B. und L. Rabiner: *Parallel processing techniques for estimatimg pitch periods of speech in the time domain.* J. Acoust. Soc. Am., 46(2):442–448, 1969.

[Gus97] Gusfield, D.: *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, 1997.

[HA01] Howard, D. und J. Angus: *Acoustics and Psychoacoustics.* Focal Press, 2. Auflage, 2001.

[Har03] Hartmann, D.: *Implementierung von Algorithmen zur automatisierten Erkennung von Holzblasinstrumenten basierend auf einem Cochleamodell.* Diplomarbeit, Technische Universität Ilmenau, Deutschland, 2003.

[Hay00] Haydn, J.: *Six Duos concertantes pour deux clarinettes.* Sieber, 1800.

[HB03] Heinz, Th. und A. Brückmann: *Using a Physiological Ear Model for Automatic Melody Transcription and Sound Source Recognition.* AES 114th Convention, Amsterdam, Niederlande, 2003.

[Her88] Hermes, D.J.: *Measurement of pitch by subharmonic summation.* J. Acoust. Soc. Am., 83(1):257–264, 1988.

[HM91] Hewitt, M.J. und R. Meddis: *An evaluation of eight computer models of mammalian inner hair-cell function.* J. Acoust. Soc. Am., 90(2):904–917, 1991.

[HMPF95] Hawkins, H., T. McMullen, A. Popper und R. Fay: *Auditory Computation.* Springer, 2. Auflage, 1995.

[Hoc02] Hockel, K.: *Ähnlichkeitssuche in Audiodatenbanken.* Institut für theoretische und technische Informatik, Technische Universität Ilmenau, Studienarbeit, 2002.

[Hof02] Hofmann, Ch.: *Auswertung der Testreihen für die Melodieerkennungssoftware: Query by Humming.* Technischer Bericht 11, Fhg IIS-AEMT Ilmenau, Abt. Metadaten, 2002.

[HP00] Haus, G. und E. Pollastri: *A multimodal framework for music inputs.* Proc. of ACM Multimedia 2000, Los Angeles, Kalifornien, USA, November 2000.

[HP01] Haus, G. und E. Pollastri: *An Audio Front End for Query-by-Humming Systems.* International Symposium on Music Information Retrieval 2001 (Ismir 2001), Bloomington, Indiana, USA, Oktober 2001.

[idm04] *Fraunhofer „Institut für digitale Medientechnologie“ (IDMT)*, 2004. http://www.idmt.fhg.de.

[IM92] Immerseel, L. Van und J.P. Martens: *Pitch and voiced/unvoiced determination with an auditory model.* J. Acoust. Soc. Am., 91(6):3511–3526, 1992.

[int03] *„Intelliscore 4.0“. Innovative Music Systems Inc.* http://www.intelliscore.net, 2003.

[JM01] Jensen, K. und D. Murphy: *Segmenting Melodies into Notes.* Proceedings of the DSAGM, Kopenhagen, Dänemark, 2001.

[Kát01] Kátai, A.: *Testbericht Query By Playing System.* Technischer Bericht 10, Fhg IIS-AEMT Ilmenau, Abt. Metadaten, 2001.

[Kla99] KLAPURI, A.: *Sound Onset Detection by Applying Psychoacoustic Knowledge.* Proc. of the IEEE ICASSP, Phoenix, Arizona, 1999.

[Kla01a] KLAPURI, A.: *Automatic Transcription of Music.* Symposium on „Stochastic Modelling of Music", 14th Meeting of the FWO Research Society on Foundations of Music Research, Ghent, Belgien, 2001.

[Kla01b] KLAPURI, A.: *Multipitch Estimation and Sound Separation By the Spectral Smoothness Principle.* IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, USA, 5, 2001.

[Kli87] KLINKE, R.: *Die Verarbeitung von Schallreizen im Innenohr.* HNO, 35:139–148, 1987.

[KNKT98] KASHINO, K., K. NAKADAI, T. KINOSHITA und H. TANAKA: *Application of the Bayesian probability network to music scene analysis.* in Computational Auditory Scene Analysis, D. F. Rosenthal and H. Okuno (Eds.), Mahwah, NJ: Lawrence Erlbaum, Seiten 115–137, 1998.

[Kuh90] KUHN, W.B.: *A real-time pitch recognition algorithm for music applications.* Computer Music Journal, 14(3):60–71, 1990.

[KVES01] KLAPURI, A., T. VIRTANEN, A. ERONEN und J. SEPPÄNEN: *Automatic Transcription of Musical Recordings.* Consistent & Reliable Acoustic Cues Workshop. CRAC-01, Aalborg, Dänemark, 2001.

[KVH00] KLAPURI, A., T. VIRTANEN und J.-M. HOLM: *Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals.* Proc. COST-G6 Conference on Digital Audio Effects, Verona, Italien, 2000.

[Lan90] LANE, J.E.: *Pitch detection using a tunable IIR filter.* Computer Music Journal, 14(3):46–59, 1990.

[Lar77] Larrson, B.: *Pitch tracking in music signals.* Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Schweden), 1977.

[Lem02] Leman, L.P. Clarisse; J.P. Martens; M. Lesaffre; B. De Baets; H. De Meyer; M.: *An Auditory Model Based Transcriber of Singing Sequences.* ISMIR 2002, 3rd International Conference on Music Information Retrieval, IRCAM - Centre Pompidou Paris, Frankreich, Seiten 116–123, Oktober, 2002.

[LYZ] Lu, L., H. You und H.-J. Zhang: *A New Approach To Query By Humming In Music Retrieval.* Microsoft Research, Beijing, China.

[MA01] Manufacturers Association, MIDI: *The Complete MIDI 1.0 Detailed Specification.* 2. Auflage, 2001.

[Mar96a] Martin, K.D.: *Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing.* Technischer Bericht 399, MIT Media Laboratory Perceptual Computing Section, Dezember 1996.

[Mar96b] Martin, K.D.: *A blackboard system for automatic transcription of simple polyphonic music.* Technischer Bericht 385, MIT Media Laboratory Perceptual Computing Section, Juli 1996.

[Mar01] Marolt, M.: *SONIC: Transcription of Polyphonic Piano Music with Neural Networks.* Proceedings of Workshop on Current Research Directions in Computer Music, Barcelona, 2001.

[MB94] Maher, R.C. und J.W. Beauchamp: *Fundamental Frequency Estimation of Musical Signals Using a Two-way Mismatch Procedure.* J. Acoust. Soc. Am., 95(4):2254–2263, 1994.

[MB96] Masri, P. und A. Bateman: *Improved Modelling of Attack Transients in Music Analysis-Resynthesis.* In Proceedings of the International Computer Music Conference, Hong Kong, Seiten 100–104, 1996.

[Med86] Meddis, R.: *Simulation of mechanical to neural transduction in the auditory receptor.* J. Acoust. Soc. Am., 79(3):702–711, 1986.

[Med88] Meddis, R.: *Simulation of auditory-neural transductions: Further studies.* J. Acoust. Soc. Am., 83(3):1056–1063, 1988.

[mel03] *Meldex is part of the New Zealand Digital Library project.* http://www.nzdl.org, 2003.

[MF02] Martins, L.G.P.M. und A.J.S. Ferreira: *PCM to MIDI Transposition.* AES 112th Convention, München, Deutschland, 2002.

[MHS90] Meddis, R., M.J. Hewitt und T.M. Shackleton: *Implementation details of the inner hair-cell/auditory-nerve synapse.* J. Acoust. Soc. Am., 87(4):1813–1816, 1990.

[MO98] Meddis, R. und L. O'Mard: *Psychophysically Faithful Methods for Extracting Pitch.* In: Rosenthal, D.F. und H.G. Okuno (Herausgeber): *Computational Auditory Scene Analysis*, Seiten 43–58. Lawrence Erlbaum Associates, 1998.

[Moo95] Moore, B.: *"Hearing". Handbook of Perception and Cognition, 2nd Edition.* Academic Press, 1995.

[mpe03] *MPEG-7 „Multimedia Content Description Interface" Documentation*, 2003. http://ipsi.fraunhofer.de/delite/Projects/MPEG7.

[MR97] Moelants, D. und C. Rampazzo: *A Computer System for the Automatic Detection of Perceptual Onsets in a Musical Signal.* In A. Camurri (Ed.) "KANSEI, The Technology of Emotion", Genova, Seiten 140–146, 1997.

[MS90] Mirollo, R.E. und S.H. Strogatz: *Synchronization of pulse-coupled biological oscillators.* SIAM J. Appl. Math., 50(6), 1990.

[MS00] Monti, G. und M. Sandler: *Monophonic Transcription with Autocorrelation.* Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italien, 2000.

[MS02] Monti, G. und M. Sandler: *Automatic Polyphonic Piano Note Extraction Using Fuzzy Logic in a Blackboard System.* Proc. of the 5^{th} Int. Conference on Digital Audio Effects (DAFx-02), Hamburg, Deutschland, 2002.

[MSW96] McNab, R.J., L.A. Smith und I.H. Witten: *Signal Processing for Melody Transcription.* Proceedings of the 19th Australian Computer Science Conference, Melbourne, Australien, Januar, 1996.

[MSWH00] McNab, R.J., L.A. Smith, I.H. Witten und C.L. Henderson: *Tune Retrieval in the Multimedia Library.* Multimedia Tools and Applications, 10(2/3):113–132, 2000.

[mus04] *Phononet GmbH*, 2004. http://www.musicline.de.

[Ney82] Ney, H.: *A time warping approach to fundamental period estimation.* IEEE Trans. on Systems, Man and Cybernetics, SMC-12(3):383–388, 1982.

[NI02] Nakatani, T. und T. Irino: *Robust Fundamental Frequency Estimation Against Background Noise and Spectral Distortion.* ICSLP-2002, 3:1733–1736, 2002.

[Nol70] Noll, A.W.: *Cepstrum Pitch Determination.* J. Acoust. Soc. Am., 41(2):293–309, 1970.

[OBCQ02] Ortiz-Berenguer, L.I. und F.J. Casajus-Quiros: *Polyphonic Transcription Using Piano Modelling For Spectral Pattern Recognition.* Proc. of the 5^{th} Int. Conference on Digital Audio Effects (DAFx-02), Hamburg, Deutschland, 2002.

[OW87] Opolko, F. und J. Wapnick: *McGill University Master Samples [Compact disc].* McGill University, Montreal, Quebec, 1987.

[Pau02] Pauws, S.: *CubyHum: A Fully Operational Query by Humming System.* ISMIR 2002, 3rd International Conference on Music Information Retrieval, Ircam Paris, Frankreich, Seiten 187–196, 2002.

[Pea86] Pearl, J.: *Fusion, Propagation and Structuring in Belief Networks.* Artificial Intelligence, 29(3):241–288, 1986.

[PG79] Piszczalski, M. und B. Galler: *Predicting musical pitch from component frequency ratios.* J. Acoust. Soc. Am., 66(3):710–720, Sept. 1979.

[PH90] Patterson, R. und J. Holdsworth: *An Introduction to Auditory Sensation Processing.* in AAM HAP, 1(1), 1990.

[PJ82] Pollard, H.F. und E.V. Jansson: *A tristimuls method for the specification of musical timbre.* Acustica, 51:161–171, 1982.

[Rab77a] Rabiner, L.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of the IEEE, Seiten 257–286, 1977.

[Rab77b] Rabiner, L.R.: *On the use of autocorrelation analysis for pitch determination.* IEEE Trans. on ASSP, 25:22–33, 1977.

[Rap99] Raphael, C.: *Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(4):360–370, 1999.

[Rap02] Raphael, C.: *Automatic Transcription of Piano Music.* ISMIR 2002, 3rd International Conference on Music Information Retrieval, IRCAM - Centre Pompidou Paris, Frankreich, Seiten 15–19, 2002.

[RG02] Rosier, J. und Y. Grenier: *Pitch Estimation for the Separation of Musical Sounds.* AES 112th Convention, München, Deutschland, 2002.

[Ric01] Richter, Ch.: *Pitchbestimmung von harmonic sustained Musikinstrumenten und Amplitudenhüllkurvenbestimmung.* Technischer Bericht 9, Fhg IIS-AEMT Ilmenau, Abt. Metadaten, 2001.

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

[RR02] Rao, P. und M. Anand Raju: *Building A Melody Retrieval System.* Department of Electrical Engineering, Indian Institute of Technology, Bombay, Indien, 2002.

[RRS+99] Rossignol, S., X. Rodet, J. Soumagne, J.-L. Colette und P. Depalle: *Automatic characterisation of musical signals: feature extraction and temporal segmentation.* Journal of New Music Research, 28(4):281–295, 1999.

[RSC+74] Ross, M.J., H.J. Shaffer, A. Cohen, R. Freudberg und H.J. Manley: *Average Magnitude Difference Function Pitch Extractor*. IEEE Trans. on Acoustics, Speech and Signal Processing, 22(5):353–362, 1974.

[Sch85] Schloss, W.A.: *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High Level Analysis*. Doktorarbeit, Stanford University, CCRMA, 1985.

[Sch96] Schürmann, J.: *Pattern classification*. John Wiley and sons, 1996.

[Sch98] Scheirer, E.D.: *Tempo and beat analysis of acoustic musical signals*. J. Acoust. Soc. Am., 103(1):588–601, 1998.

[Sla90] Slaney, M.: *A perceptual pitch detector*. Proc. ICASSP, Seiten 357–360, 1990.

[Smi96] Smith, L.S.: *Onset-based Sound Segmentation*. Advances in Neural Information Processing Systems 8, D.S. Touretzky, M.C. Mozer, M.E. Haselmo (eds), MIT press, 1996.

[Sol98] Solbach, L.: *An Architecture for Robust Partial Tracking and Onset Localization in Single Channel Audio Signal Mixes*. Doktorarbeit, Distributed Systems Department, Technische Universität Hamburg-Harburg, 1998.

[sol03] *„Solo Explorer 1.0“*. http://www.recognisoft.com, 2003.

[SS97] Scheirer, E. und M. Slaney: *Construction and evaluation of a robust multifeatures speech/music discriminator*. IEEE Transactions on Acoustics, Speech and Signal Processing (ICASSP'97), Seiten 1331–1334, 1997.

[SS01] Schmidt, R. und H.G. Schaible: *Neuro- und Sinnesphysiologie*. Springer, 4. Auflage, 2001.

[SSZ97] Schiebler, T., W. Schmidt und K. Zilles: *Anatomie*. Springer, 7 Auflage, 1997.

[Ste99] Stemplinger, I. M.: *Beurteilung, Messung und Prognose der Globalen Lautheit von Geräuschimmissionen.* Herbert Utz Verlag, 1999.

[Ter79] Terhardt, E.: *Calculating virtual pitch.* Hearing Research, 1:155–182, 1979.

[TSS82] Terhardt, E., G. Stoll und M. Seewann: *Algorithm for extraction of pitch and pitch salience from complex tonal signals.* J. Acoust. Soc. Am., 71(3):679–688, 1982.

[VDC01] Verfaille, V., P. Duhamel und M. Charbit: *LIFT: Likelihood-Frequency-Time Analysis For Partial Tracking and Automatic Transcription of Music.* Proc. of the COST G-6 Conference on Digital Audio Effects (DAFx-01), Limerick, Irland, 2001.

[VK01] Virtanen, T. und A. Klapuri: *Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation.* Workshop on Applications of Signal Processing to Audio and Acoustics, 2001.

[VK02] Virtanen, T. und A. Klapuri: *Separation of Harmonic Sounds Using Linear Models for the Overtone Series.* ICASSP, 2002.

[VR81] Vos, J. und R. Rasch: *The perceptual onset of musical tones.* Perception & Psychophysics, 29(4):323–335, 1981.

[Wat95] Waterman, M.S.: *Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics.* Chapman & Hall/CRC, 1995.

[Wer25] Wertheimer, M.: *Über Gestalttheorie.* Philosophische Zeitschrift für Forschung und Aussprache, 1:39–60, 1925.

[WGR99] Walmsley, P., S. Godsill und P. Rayner: *Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters.* In Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 1999.

[wid03] *„Widi music recognition system 2.7“. Music Recognition Team.* http://www.widisoft.com, 2003.

[Zen94] Zenner, H.P.: *Hören.* Thieme Verlag Stuttgart, 1. Auflage, 1994.

[ZF01] Zwicker, E. und H. Fastl: *Psychoacoustics. Facts and Models.* Springer-Verlag, 2001.

[ZP90] Zwicker, E. und W. Peisl: *Cochlear preprocessing in analog models, in digital models and human inner ear.* Hear. Res., 44:209–216, 1990.

[ZZ87] Zwicker, E. und M. Zollner: *Elektroakustik.* Springer-Verlag, Berlin, 1987.