MATLAB®-Tool für die Parameterberechnung zur Additiven Klangsynthese

Diplomarbeit

von

Cornelia Falch

durchgeführt am

Institut für Elektronische Musik und Akustik

der Universität für Musik und Darstellende Kunst Graz

Betreuer:

Univ.Ass. Dipl.Ing. Alois Sontacchi

O.Univ.-Prof. Mag. Dipl.Ing. Dr. Robert Höldrich

Graz, September 2002

In erster Linie möchte ich meinen Eltern herzlich danken, die mir durch ihre Hilfe und Unterstützung dieses Studium ermöglichten. "Danke, dass ihr meine oft nicht sehr einfachen Entscheidungen akzeptiert und immer an mich geglaubt habt!"

Weiters danke ich besonders Herrn Alois Sontacchi und Herrn Robert Höldrich für die sehr gute Betreuung während der Diplomarbeit.

Dem gesamten Team des IEM danke ich für die angenehme Atmosphäre und ständige Hilfsbereitschaft!

Speziell möchte ich noch Markus, Peter und Michael für die interessanten Diskussionen im DiplomandInnenraum und auf der Brücke sowie dem netten Klima beim Arbeiten danken.

Zum Schluss bedanke ich mich noch bei meinem Bruder Meinhard, der alle meine Höhen und Tiefen der letzten Monate am intensivsten miterlebt hat. "Danke für deine Geduld!"

Zusammenfassung

Die Additive Klangsynthese ist nach wie vor das qualitativ beste Resyntheseverfahren. Die notwendige große Rechenleistung hat in früheren Jahren den Echtzeiteinsatz verhindert, heute stellt dies jedoch kein Problem mehr dar. Eine Herausforderung bleibt allerdings noch immer die Berechnung der Parameter, es sind dies die Frequenz-, Amplituden-, sowie Phasenverläufe der einzelnen Sinusgeneratoren.

Im Rahmen dieser Diplomarbeit werden unterschiedliche, bereits bekannte Analyseverfahren (z.B. Spectral Modeling Synthesis, SMS; MQ Algorithm; Reassignment) untersucht und miteinander verglichen. Eine geeignete Kombination daraus soll zur möglichst genauen Bestimmung der gesuchten Parameter führen. Da sich die Additive Klangsynthese ausschließlich auf die Addition der harmonischen und inharmonischen Teiltöne eines Signals beschränkt, muss das zu analysierende Signal zudem noch in einen deterministischen bzw. einen stochastischen Part zerlegt werden.

Das in MATLAB entwickelte Programm "SOUND ANALYSIS" bildet die Analyseeinheit zu einem bereits bestehenden Synthesealgorithmus, der ebenfalls in MATLAB generiert wurde. Somit war das Ergebnis bzw. das Ziel der Arbeit fix vorgegeben. Die Performance des Systems wurde mit Hilfe von unterschiedlichen Audiosignalen in mehreren Tests untersucht und daraus die wesentlichen Kriterien der für diese Kombination aus Analyse- und Synthesestufe geeigneten Eingangssignale abgeleitet.

Abstract

One of the most powerful tools for resynthesing sound signals represents the additive sound synthesis (ASS). Since the speed and computational power of computers increased dramatically in recent years, the use of this synthesis has become even more important. Thus, also analysis algorithms for estimating the parameters needed in ASS, that are the frequency, amplitude and phase, respectively, of a sound signal, were improved and extended.

The aim of this diploma thesis is to investigate existing analysis software (i.e. Spectral Modeling Synthesis, SMS; MQ Algorithm; Reassignment), and combine them in order to obtain best detection results for the three desired parameters. An additional decomposition of the input into a deterministic and a stochastic signal seems to be reasonable as the ASS only takes account of the harmonic and inharmonic part. Thus, a separate processing of both signal types is employed, which leads to audible improvements when resynthesing the original.

Within this thesis a program called "SOUND ANALYSIS" is especially delveloped and designed to act as an analysis tool for an already existing synthesis algorithm, both written in MATLAB programming language. Several tests performed with different audio signals showed the limits and characteristica of the entire system.

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit der Analyse von Klangsignalen sowie der Aufbereitung und Speicherung bestimmter Charakteristika dieser Signale, welche für eine nachfolgende Resynthese notwendig sind. Die Modifikation dieser Merkmale, welche vor der Rekonstruktion ausgeführt werden kann, stellt ein bedeutendes Kompositionswerkzeug in der Computermusik dar. Deshalb ist eine einfache und flexible Handhabung der gespeicherten Information erstrebenswert. In Abb. 1.1 ist die Grundstruktur eines beliebigen Analyse/Synthese-Systems mit einer zusätzlichen Einheit zur Variation der Daten dargestellt.

In der Signalverarbeitung werden Schallereignisse jeglicher Art, wie zum Beispiel Musik oder Sprache, durch unterschiedliche mathematische Modelle beschrieben, die, je nach Anwendung, spezielle Kriterien aufweisen. Zur digitalen Reproduktion gibt es eine Vielzahl unterschiedlicher Synthesemethoden, deren wichtigstes Kriterium das implementierte Signalmodell mit seinen Parametern darstellt. Im wesentlichen unterscheidet man heutzutage drei Modelltypen: instrumentale, abstrakte und spektrale Modelle. Der erste Typ versucht, ein Modell für die Klangerzeugung, z.B. ein Musikinstrument oder die menschliche Stimme, zu finden und diese mit entsprechenden Parametern zu beschreiben. Der zweite Typ versucht, musikalisch wichtige Parameter eines Audiosignals in abstrakte Formeln zu extrahieren. Ziel des letzten Typs ist es, die direkt auf die Basilarmembran auftreffenden Schallwellen zu modellieren. Dabei werden weder

die Einflüsse von Außen- und Mittelohr, noch Details des Hörvorganges selbst (z.B. die Maskierung eines Tons) berücksichtigt. [Serra, Smith], [Tolonen, et. al]

In der Literatur findet man verschiedene Realisationen der spektralen Modelle. Zwei bekannte Vertreter sind die Additive Klangsynthese (AKS) und der Phasenvokoder [Dolson]. In der vorliegenden Arbeit wird ein Analysesystem vorgeschlagen, das zur Bestimmung der für die AKS notwendigen Parameter dient. Die Grundstruktur leitet sich aus den beiden in Kapitel 3 genauer erläuterten Algorithmen ab.

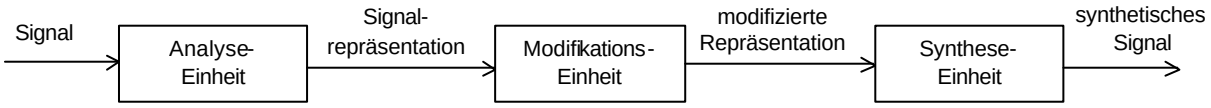


Abb. 1.1: Aufbau eines generellen Analyse/Synthese System

2 Klassifizierung und Darstellung von Signalen

Dieses Kapitel beschäftigt sich mit der Erklärung einiger grundlegender Begriffe, welche in der vorliegenden Arbeit häufig verwendet werden. Zu Beginn wird eine einfache Klassifizierung von physikalischen Signalen vorgenommen, die in den meisten Fällen zur Beschreibung von Musik und Sprache ausreicht. In der Signalverarbeitung spielt die Darstellung von Signalen eine wichtige Rolle, weshalb anhand dieser Einteilung einige Möglichkeiten der Signalrepräsentation vorgestellt werden.

2.1 Arten von Signalen

Im allgemeinsten Fall kann ein Signal als mathematische Funktion x beschrieben werden, die von der Zeit t abhängig ist:

$$x = f(t)$$

Unter der Annahme, dass diese Funktion eine periodische Sinusschwingung mit der Amplitude A und der Frequenz n ist, wird x zu:

$$x(t) = A\cos(2\pi nt)$$

wobei hier A und n als konstant und von der Zeit unabhängig angenommen werden. Generell gesehen trifft dies für die in der Natur vorkommenden Signale

nicht zu. Diese weisen eine Modulation der Amplitude und/oder der Frequenz auf und werden deshalb auch als nichtstationär bezeichnet:

$$x(t) = A(t)\cos\left[2\pi n(t)t\right]$$

Des weiteren bestehen natürliche Signale nicht nur aus einem Sinus, sondern aus der Superposition mehrerer periodischer Teilschwingungen. Daraus ergibt sich eine Unterteilung in Ein- und Mehrkomponentensignale, welche folgendermaßen beschrieben werden:

$$x(t) = \sum_{n=1}^{N} A_n(t)\cos\left[2\pi n_n(t)t\right]$$

Bisher wurden ausschließlich periodische, deterministische Signale betrachtet. Jedem Musik- und Sprachsignal ist zusätzlich noch ein stochastischer Anteil, zum Beispiel Rauschen, überlagert. Zur Beschreibung dafür sind komplexere Funktionen notwendig, auf die allerdings hier nicht näher eingegangen wird. Damit lässt sich ein beliebiges physikalisches Signal darstellen mit:

$$x(t) = \sum_{n=1}^{N} A_n(t)\cos\left[2\pi n_n(t)t\right] + r(t) \qquad (2.1)$$

2.2 Darstellung von Signalen

2.2.1 Eindimensionale Repräsentationen

Die Beschreibung, Speicherung und Übertragung von Signalen erfolgt fast ausschließlich als Funktion von der Zeit $x(t)$ und liefert somit als erste Art der Darstellung die Zeitrepräsentation. Es sei zunächst ein zeitinvariables Mehrkomponentensignal bestehend aus drei Teilschwingungen betrachtet:

$$x(t) = A_1\cos(2\pi n_1 t) + A_2\cos(2\pi n_2 t) + A_3\cos(2\pi n_3 t)$$

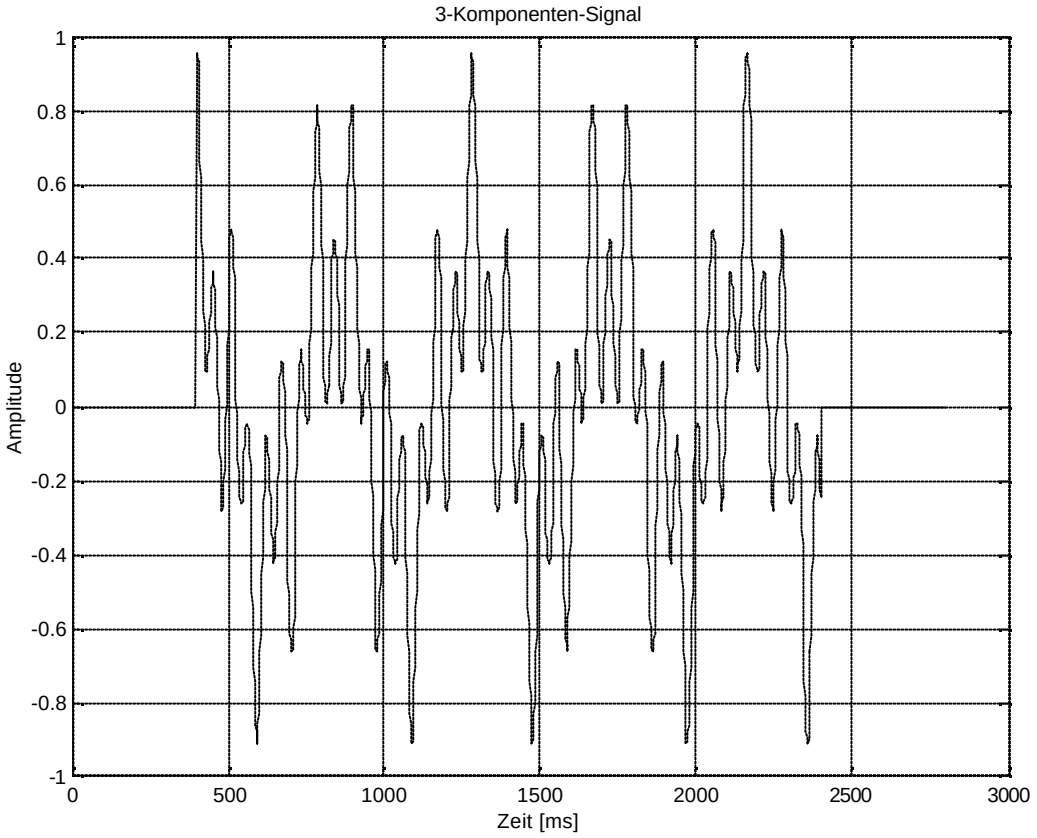


Abb. 2.1: Beispiel zur Zeitrepräsentation

Aus Abb. 2.1 ist zwar die zeitliche Struktur des Signals (Beginn, Ende, Dauer der Einzelfunktionen) unmittelbar ersichtlich, die Teilfrequenzen sind jedoch nicht ohne weiters bestimmbar. Deshalb wäre eine Darstellung der Funktion in Abhängigkeit von der Frequenz wünschenswert, was man mit Hilfe der Fourier Transformation erreicht und als Frequenzrepräsentation bezeichnet:

$$X(n) = \mathcal{F}\{x(t)\}$$

In diesem Fall (siehe Abb. 2.2) kann allerdings keine Aussage über die zeitliche Entwicklung des Signals getroffen werden, da die Fourier Transformation lediglich eine Aufspaltung des Signals in die einzelnen Frequenzkomponenten gibt, jedoch keine Auskunft über deren Dauer bzw. zeitliche Lokalisation liefert. Da bei beiden grafischen Interpretationen die gesuchte Funktion jeweils nur von einer Variable (entweder $x(t)$ oder $X(n)$) abhängt, spricht man von eindimensionalen Repräsentationen. Aus der Kombination beider Varianten erhält man die gesamte Signalinformation.

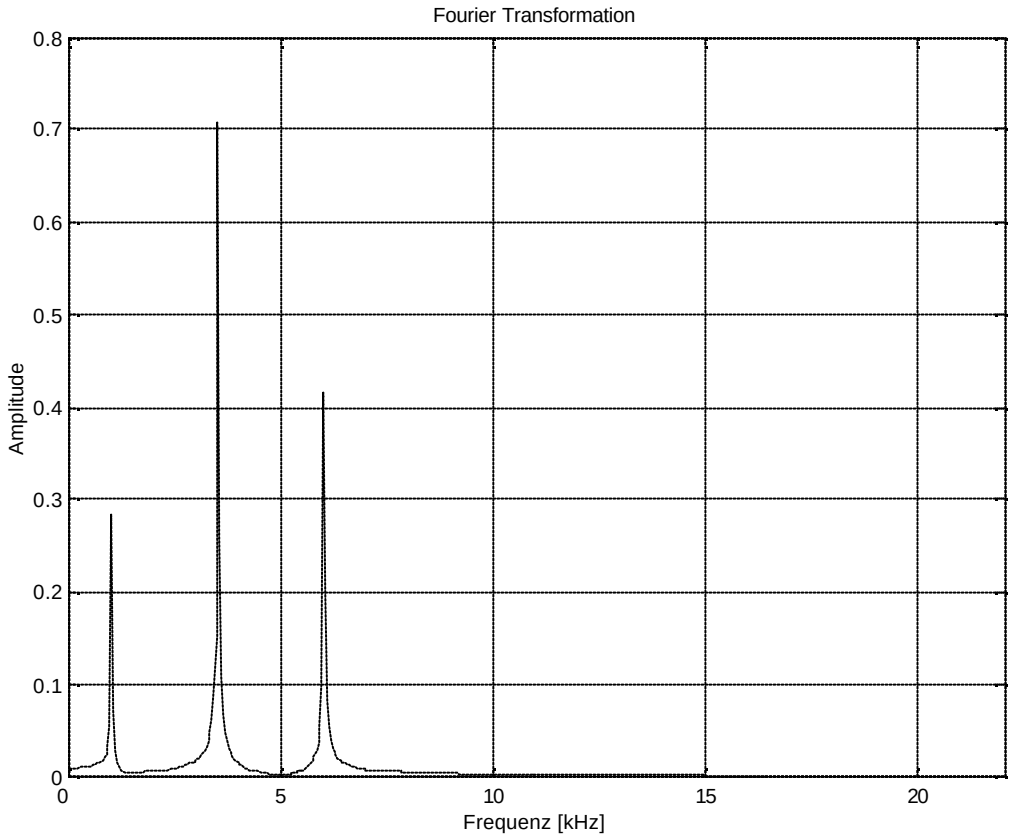


Abb. 2.2: Beispiel zur Frequenzrepräsentation

Um die Beschränkung von Zeit- und Frequenzrepräsentation auf zeitinvariante Signale zu verdeutlichen, wird im nächsten Beispiel ein Signal mit linearem Frequenzanstieg (ein sogenanntes Chirp-Signal) verwendet, wobei der Einfachheit halber die Amplitude wiederum als konstant angenommen wird:

$$x(t) = A\cos\left[2\pi n(t)t\right]$$

Abb. 2.3 liefert einen Ausschnitt von $50ms$ aus diesem zeitvarianten Signal und es ist unschwer zu erkennen, dass die Kombination der beiden Diagramme keine befriedigende Darstellung des Frequenzverlaufs bietet. Es wäre naheliegend, nach einer Alternative zu suchen, die exakt den Zusammenhang zwischen Frequenz und Zeit aufzeigt, $v = f(t)$. Eine solche Möglichkeit bietet die Einführung der Momentanfrequenz $f_i(t)$, welche als Ableitung der Phase eines Signals nach der Zeit definiert ist:

$$f_i(t) = \frac{1}{2\pi}\frac{\partial F\left\{x_a(t)\right\}}{\partial t}$$

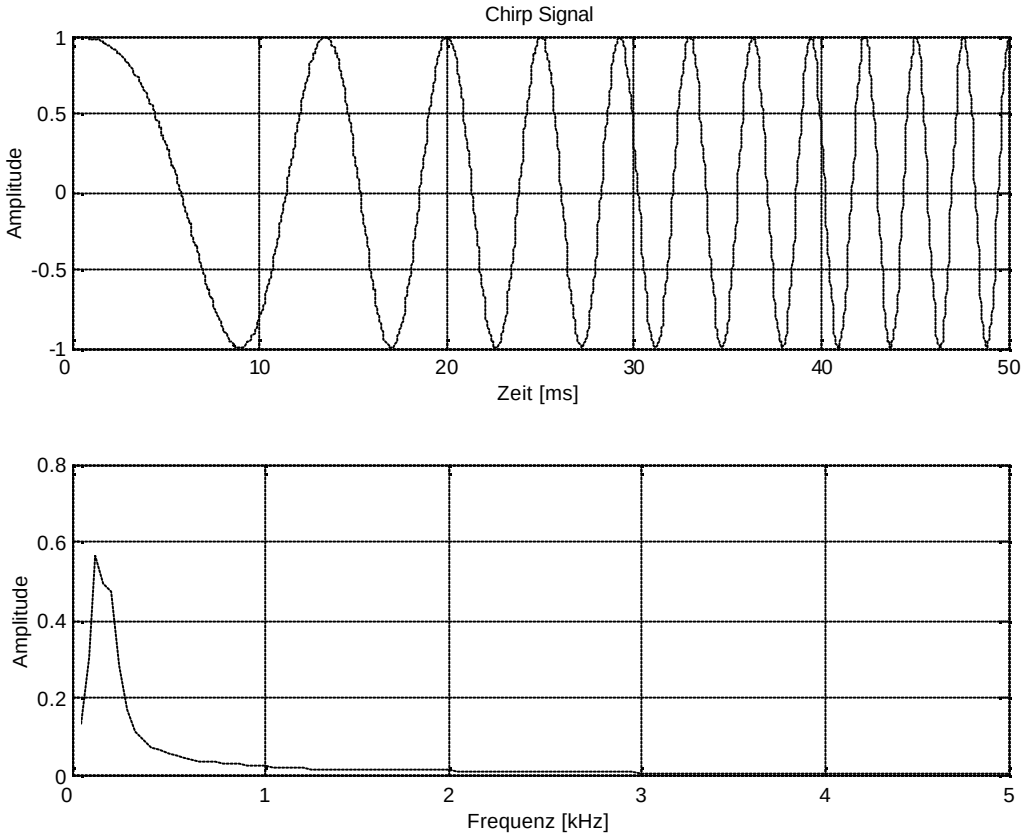


Abb. 2.3: Zeit- und Frequenzrepräsentation

Dabei wird allerdings eine komplexe Eingangsfunktion $x_a(t)$ vorausgesetzt, weshalb reellwertige Signale zuvor entsprechend aufbereitet werden müssen:

$$x_a(t) = x(t) + j\mathcal{H}\{x(t)\}$$

Die Funktion $x_a(t)$ wird als analytisches Signal und $\mathcal{H}\{x(t)\}$ als Hilbert Transformation von $x(t)$ bezeichnet. Die zeitliche Entwicklung der Frequenz des gesamten Chirp-Signals ist in Abb. 2.4 dargestellt.

Die Momentanfrequenz charakterisiert das Frequenzverhalten als Funktion der Zeit. In äquivalenter Weise wird das lokale Zeitverhalten als Funktion der Frequenz mit der Gruppenlaufzeit t_g beschrieben:

$$t_g(n) = -\frac{1}{2p}\frac{\partial F\{X_a(n)\}}{\partial f}$$

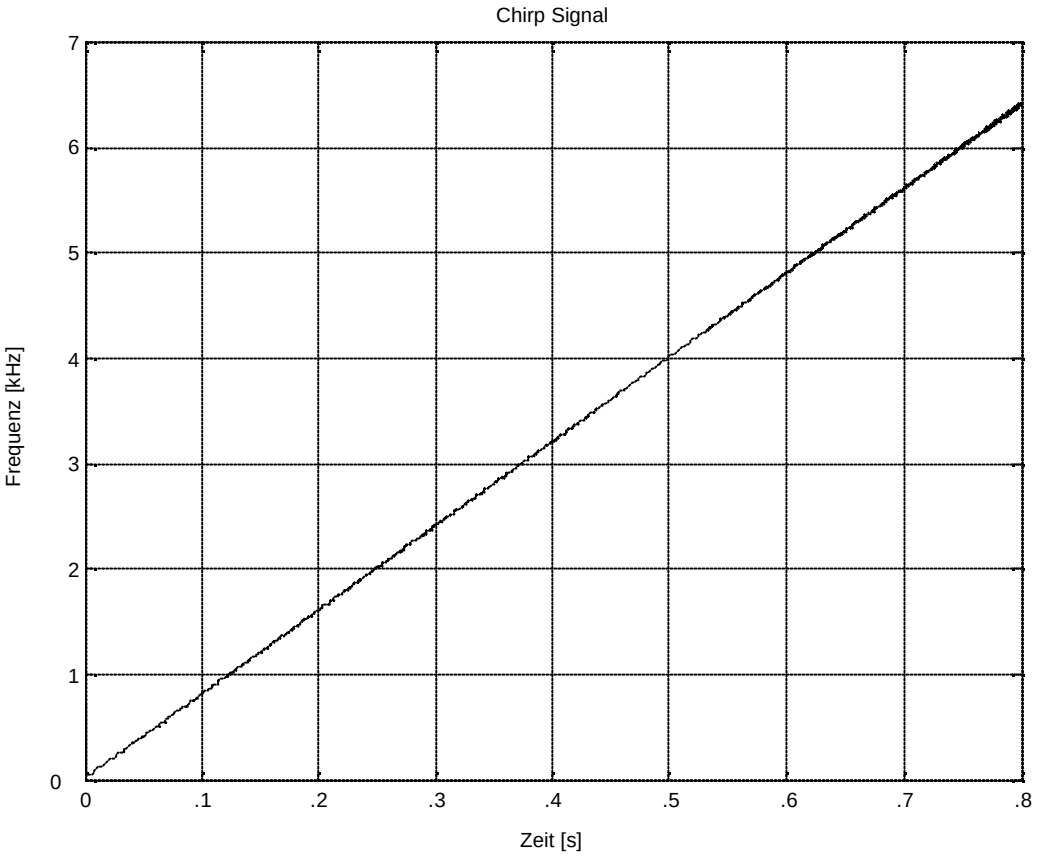


Abb. 2.4: Momentanfrequenz

Diese Größe gibt darüber Auskunft, zu welchem Zeitpunkt t eine bestimmte Frequenz n das erste Mal im Signal auftritt (vgl. Abb. 2.5).

Wiederum sind diese Arten der eindimensionalen Darstellung auf einen bestimmten Signaltyp begrenzt. Während die Verwendung der Momentanfrequenz Einkomponentensignale voraussetzt (zu jedem betrachteten Zeitpunkt darf ausschließlich eine Frequenzkomponente existieren), liefert die Gruppenlaufzeit nur sinnvolle Ergebnisse für Signale, deren gesuchte Frequenz lediglich zu einem Zeitpunkt beginnt (die Frequenzkomponente darf nicht mehrmals hintereinander auftreten) [Auger, et al].

Aus den vorangegangenen Erkenntnissen lässt sich schlussfolgern, dass eindimensionale Repräsentationen zur Darstellung von nichtstationären Multikomponentensignalen unbrauchbar sind. Da die meisten in der Natur vorkommenden Signale eben diese Eigenschaften aufweisen, ist die Einführung von zweidimensionalen Repräsentationen unentbehrlich.

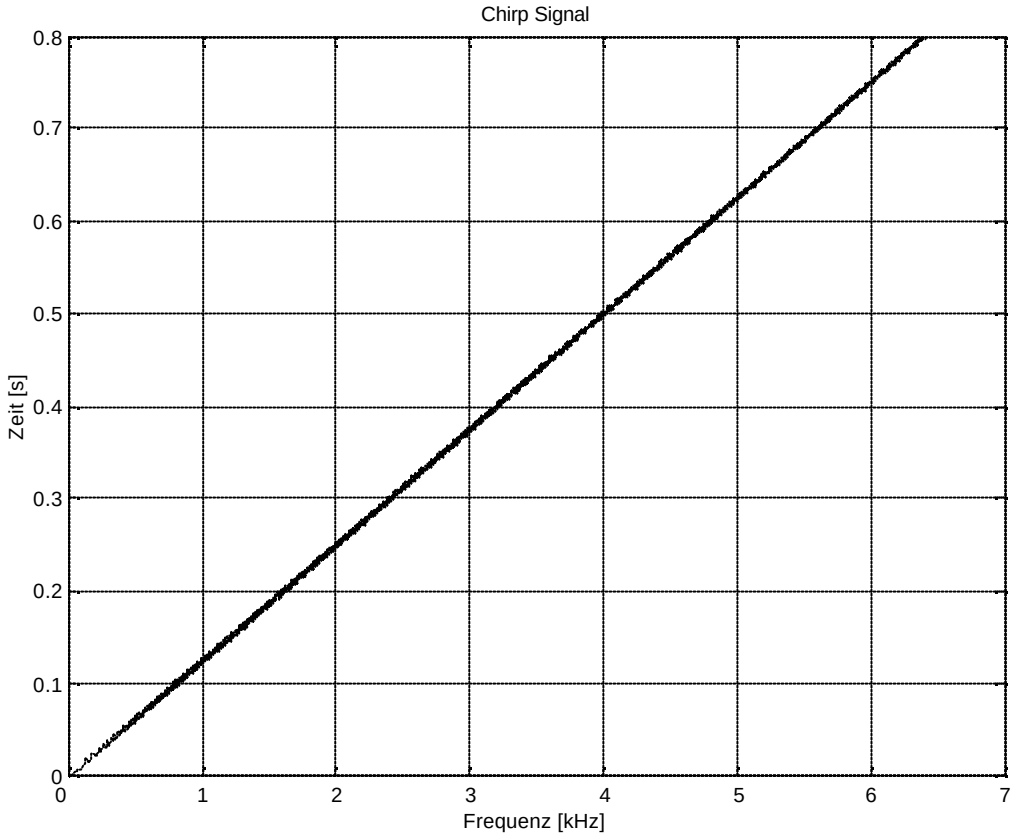


Abb. 2.5: Gruppenlaufzeit

2.2.2 Zweidimensionale Repräsentationen

Als erste Variante der gemeinsamen Darstellung der Abhängigkeit eines Signals von der Zeit und der Frequenz eignen sich lineare Zeit-Frequenz-Repräsentationen, welche mit Hilfe der Kurzzeit-Fouriertransformation (KZFT) gebildet werden. Eine ausführliche Erklärung dieser Transformation findet sich in Kapitel 4.2. Zunächst ist ausschließlich das Ergebnis interessant: Es besteht aus einer zeitlichen Sequenz komplexer Spektren und beinhaltet somit die Information über die Zeit- und Frequenzauflösung. Abb. 2.6 zeigt den Amplitudenverlauf eines Signals, das sich aus zwei linearen Frequenzchirps zusammensetzt, wobei einer um Dt verzögert ist.

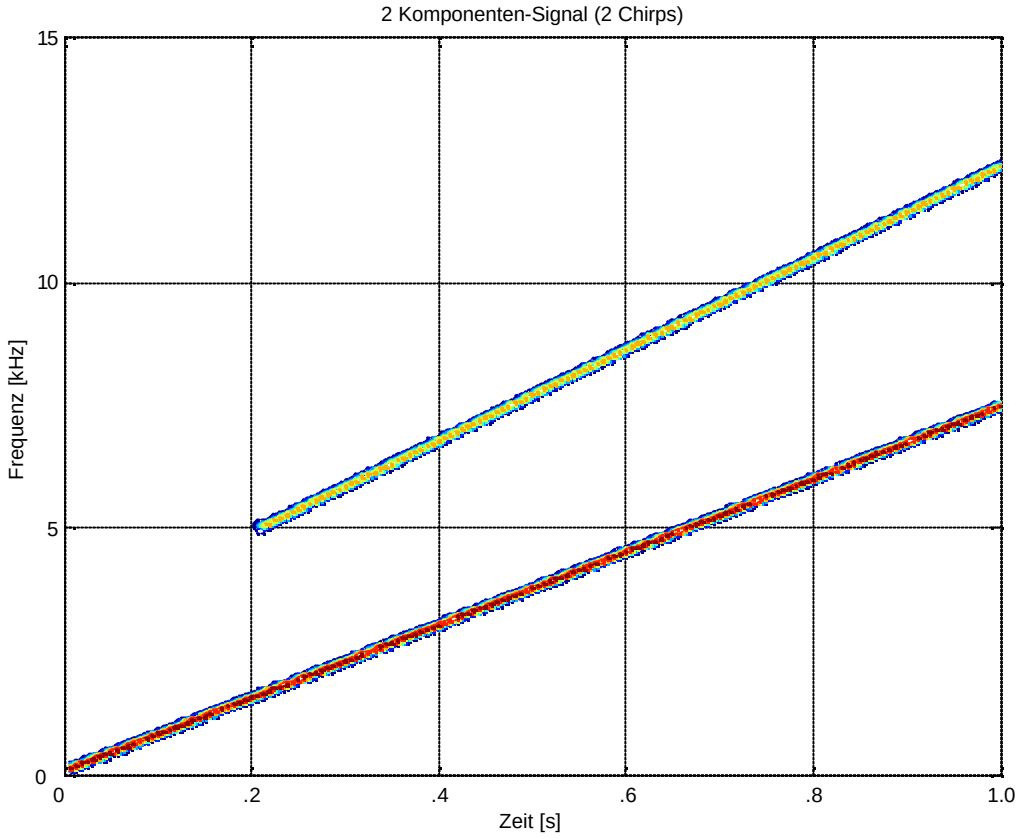


Abb. 2.6: Zeit-Frequenzrepräsentation

Zwei weitere Vertreter der linearen Zeit-Frequenz-Repräsentation sind die Wavelet Transformation (WT) und die Gabor Expansion (GE). Die Idee der WT ist die Projizierung des Eingangssignals $x(t)$ auf eine Gruppe von Funktionen, den sogenannten "Wavelets", die von einer gemeinsamen Elementarfunktion ("Mutterfunktion") abgeleitet werden. In der zeitkontinuierlichen Ebene sind sie definiert als:

$$T_x(t,a;Y) = \int_{-\infty}^{\infty} x(s) Y^*_{t,a}(s)\,ds$$

mit $Y_{t,a}(s) = |a|^{-\frac{1}{2}} Y\left(\frac{s-t}{a}\right)$. Die Variable a entspricht hierbei einem Skalierungsfaktor, wobei gilt: $|a|>1$ bedeutet eine Erweiterung des Wavelets Y, $|a|<1$ bedeutet eine Komprimierung von Y. Für $a=1$ stellt die WT eine echte Zeit-Frequenz-Repräsentation dar (für $a \neq 1$ spricht man von einer Zeitskalierungsrepräsentation). Der grundlegende Unterschied zur KZFT liegt darin, dass durch die Wahl des Skalierungsfaktors a die Bandbreite und Dauer der

Wavelets variiert werden können, ohne jedoch ihre Form zu ändern. Im Gegensatz zur KZFT, welche ein einziges Analysefenster fixer Länge verwendet, wird bei der WT für hohe Frequenzen ein kurzes Fenster und für tiefe Frequenzen ein langes Fenster eingesetzt. Dadurch wird die Auflösebeschränkung der KZFT teilweise aufgehoben: die Bandbreite B ist proportional der zu analysierenden Frequenz n bzw. gilt: $\frac{B}{n} = Q = konst.$, weshalb diese Transformation auch als "Constant-Q Analysis" bezeichnet wird.

Als Umkehrung der KZFT bzw. als Rekonstruktion (Synthese) eines Signals kann die GE verstanden werden. Die Syntheseformel entspricht einer linearen Superposition von zeit- und frequenzversetzten Elementarsignalen und lautet in diskreter Form:

$$x(t) = \sum_n \sum_m F_x[n, m; h] g_{n,m}(t)$$

wobei $g_{n,m}(t) = g(t - nt_0)e^{j2\pi mn_0 t}$ als Gabor Repräsentation bezeichnet wird. Die Elementarsignale $g_{n,m}(t)$ werden auch Gabor logons genannt. Die Koeffizienten $F_x[n, m; h]$ heißen Gabor Koeffizienten und enthalten Informationen über den Zeit- und Frequenzinhalt eines Signals zu einem bestimmten Zeit- und Frequenzpunkt (nt_0, mn_0).

Da keine der beiden Transformationen im vorgeschlagenen Algorithmus zur Anwendung kommt, wird auf eine detailliertere Beschreibung verzichtet. Ausführliche Informationen zu beiden Repräsentationen finden sich in [Auger, et al], [Höldrich], [Mallat], [Gabor].

Eine weitere Möglichkeit der zweidimensionalen Darstellung ergibt sich durch die Betrachtung der Energieverteilung eines Signals und führt zur Gruppe der quadratischen Zeit-Frequenz-Repräsentationen, deren einfachste Form, das Spektrogramm, das Betragsquadrat der KZFT ist:

$$Sp(x; t, n) = \left| \mathcal{KZFT}\ \{x(t)\} \right|^2$$

Es liefert ein der linearen Abbildung von vorhin äquivalentes Ergebnis, siehe Abb. 2.7.

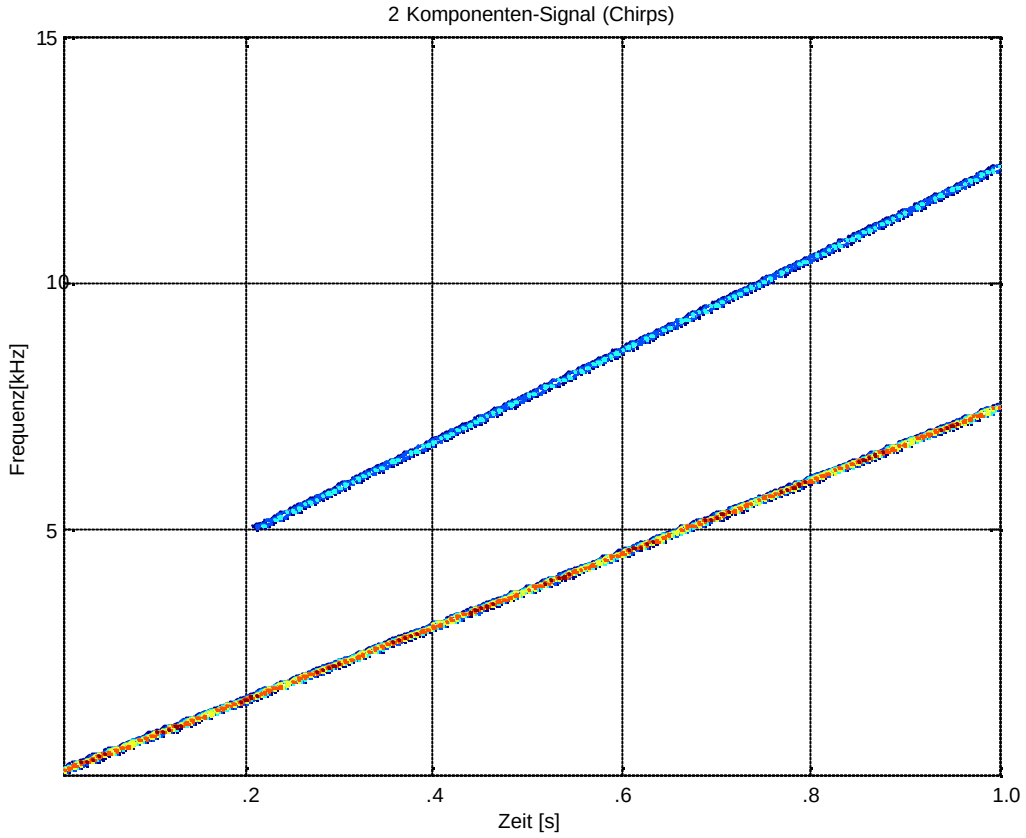


Abb. 2.7: quadratische Zeit-Frequenzrepräsentation, Energiebetrachtung

In beiden Abbildungen ist die Skalierung der Zeit- und Frequenzachse durch die Implementation der KZFT fix vorgegeben, auf die in Kapitel 4.2 ausführlich eingegangen wird. Infolgedessen kann jeder Analysewert exakt einem Rasterpunkt in der Zeit-Frequenz-Ebene zugeordnet werden. Betrachtet man diese Tatsache vom energetischen Standpunkt aus, so wird die gesamte Energie eines Zeit-Frequenz-Fensters seinem geometrischen Mittelpunkt zugeteilt. Für Energieanteile, die sich am Rande eines Fensters befinden, ergibt sich somit eine Fehllokalisation, bedingt durch die maximale Auflösung der Zeit-Frequenz-Darstellung. Eine schematische Darstellung dieses Problems zeigt Abb. 2.8(a): die tatsächliche Frequenzkomponente ist als Kreis, ihr zugewiesener Wert als "x" gekennzeichnet. In der Zeitebene ergibt sich somit ein maximaler Fehler von $e_t = \left|\frac{N}{2}\right|$, wenn, wie in Abb. 2.8(b) dargestellt, als Rasterbezugspunkt (mit "x" gekennzeichnet) die Mitte

des Fensters definiert wird[1]. Der tatsächliche Beginn des Signals ist durch einen Kreis markiert, das gefensterte Signal ist durchgezogen und die Fensterfunktion selbst strichliert gezeichnet. Die Abweichung des Analysewertes ("x") vom tatsächlichen Wert (Kreis) in der Frequenzebene hängt neben der Frequenzauflösung zusätzlich von der Fensterfunktion ab, in Abb. 2.8(c) ist sie für ein von-Hann-Fenster aufgezeigt. Zur genaueren Erläuterung des Einflusses der Fensterform sei auf die Kapitel 4.2 und 5.1 verwiesen, in denen unter anderem die gebräuchlichsten Fensterarten dargestellt sind (siehe Abb. 5.2).

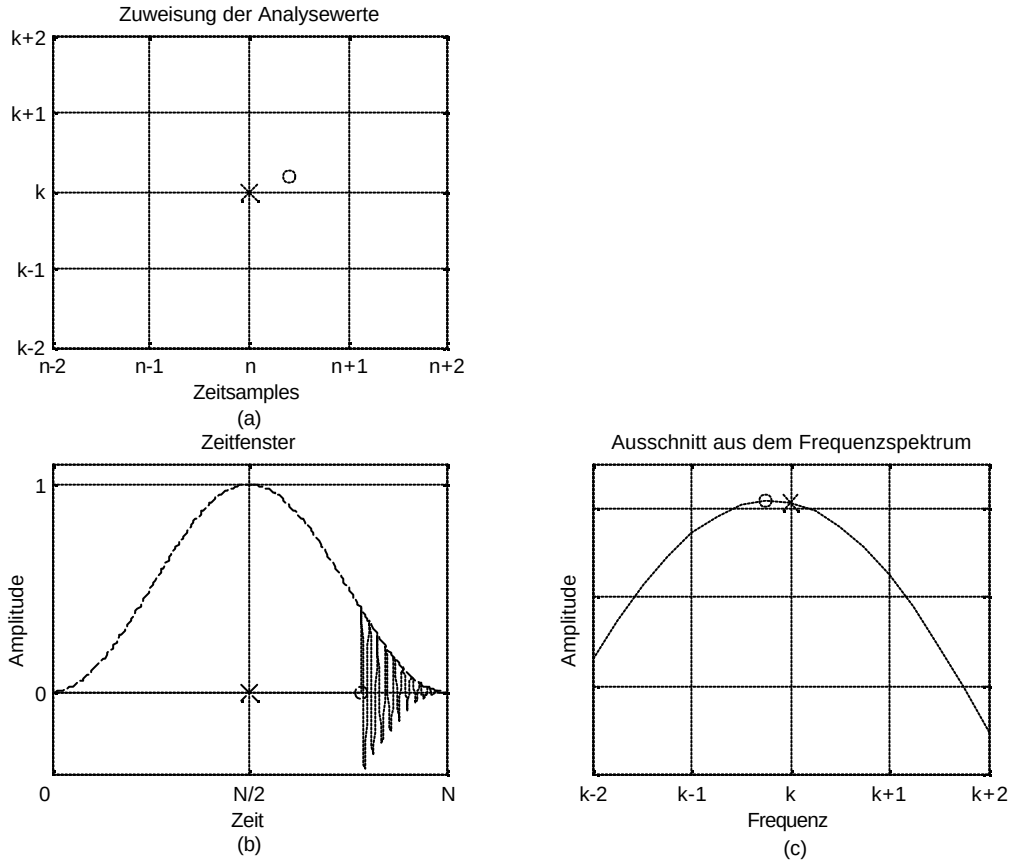


Abb. 2.8: Lokalisationsproblem

[Kodera, et al] bzw. [Auger, Flandrin, 1994] und [Auger, Flandrin, 1995] schlagen als Abhilfe vor, die Werte des Spektrogramms nicht dem Fenstermittelpunkt, sondern vielmehr dem Energieschwerpunkt zuzuweisen. Dies erfolgt durch die Miteinbeziehung der Phaseninformation des Spektrogramms. Die

[1] Wird als Bezugspunkt der Beginn des Fensters angenommen, wo wäre der maximale Fehler $e_t = |N|$.

Neuzuordnung der Punkte entlang der Zeitachse basiert auf dem Prinzip der Gruppenlaufzeit, allerdings muss dies für jede diskrete Frequenzkomponente des Signals getrennt vorgenommen werden. Analog dazu werden die Frequenzpunkte mittels Berechnung Momentanfrequenzen in jedem Zeitfenster reorganisiert.

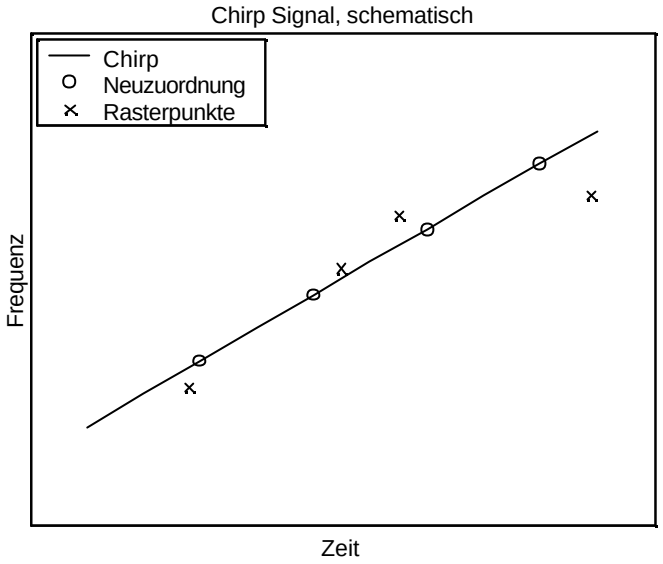


Abb. 2.9: Prinzip der Neuzuordnung

Abb. 2.9 veranschaulicht das Prinzip der Neuzuordnung des Spektrogramms anhand eines linearen Chirp-Signals. Die Kreuze kennzeichnen die Punkte des Spektrogramms, welche dem geometrischen Mittel des Fensters entsprechen, die Kreise bezeichnen den Schwerpunkt der Signalenergie. Auf die exakte Herleitung wird nicht näher eingegangen, sie wurde unter zu Hilfenahme der Gabor-Helstrom Transformation [Gabor], [Helstrom], [Montgomery, Reed] entwickelt und ist in [Kodera, et al] im Detail beschrieben.

Gegeben sei ein allgemeines nichtstationäres Mehrkomponentensignal $x(t)$, dessen Spektrogramm durch $Sp(x;t_0, f_0)$ definiert ist, wobei die Werte t_0 und f_0 Rasterpunkte der Zeit-Frequenz-Ebene sind. Die Neuzuordnung dieser Punkte geschieht mit:

$$t_0' = t_0 - \frac{1}{2p}\frac{\partial \arg\{Sp(x;t_0, f_0)\}}{\partial f_0}$$

$$f_0' = \frac{1}{2p}\frac{\partial \arg\{Sp(x;t_0, f_0)\}}{\partial t_0}$$

Eine aktuelle alternative Methode wird in [Plante, et al], [Hainsworth, Wolfe], [Hainsworth, et al] und [Fitz, et al] diskutiert und ersetzt die Verwendung der Phaseninformation durch die Einführung unterschiedlicher Fensterfunktionen, die zur Bildung des Spektrums benötigt werden[2]. Dabei werden die neuen Punkte folgendermaßen berechnet:

$$t_0'' = t_0 - \mathcal{R}\left\{\frac{X_t(x;t_0,f_0)X^*(x;t_0,f_0)}{\left|X(x;t_0,f_0)\right|^2}\right\}$$

$$f_0'' = f_0 + \mathcal{I}\left\{\frac{X_d(x;t_0,f_0)X^*(x;t_0,f_0)}{\left|X(x;t_0,f_0)^2\right|}\right\}$$

wobei $\mathcal{R}\{\cdot\}$ bzw. $\mathcal{I}\{\cdot\}$ den Real- bzw. Imaginärteil bezeichnen. $X(x;t_0,f_0)$ entspricht dem Spektrum der KZFT, gefenstert mit $w(t)$. $X_t(x;t_0,f_0)$ verwendet ein Fenster, welches mit t multipliziert ist, $w_t(t) = w(t)\cdot t$. $X_d(x;t_0,f_0)$ verwendet die Ableitung des Fensters $w(t)$ nach der Zeit, $w_d(t) = \frac{\partial w(t)}{\partial t}$. Somit verringert sich der Berechnungsaufwand des neuen Spektrums, da lediglich drei KZFT entsprechend der drei Fensterfunktionen durchgeführt werden müssen.

Weitere Darstellungsmöglichkeiten der Energieverteilung bieten die Wigner-Ville Verteilung, sowie die Cohen Klasse der bilinearen Zeit-Frequenz-Repräsentationen. Da keine der beiden bei der Entwicklung des in den folgenden Kapiteln beschriebenen Programms berücksichtigt wurden, wird auf eine eingehende Erklärung verzichtet und stattdessen auf einschlägige Literatur verwiesen, [Höldrich], [Auger, et al], [Mecklenbräuker, Hlawatsch], [Cohen]. Ausgehend von der Momentanleistung eines Signals $E_x = p_x(t) = \int_{-\infty}^{\infty} \left|x(t)\right|^2 dt$ einerseits sowie der spektralen Energiedichte $E_x = P_x = \int_{-\infty}^{\infty} \left|X(n)\right|^2 dn$ andererseits ist man bestrebt, eine gemeinsame Energieverteilung in Zeit und Frequenz zu finden. Die Lösung führt zu:

[2] Die Verwendung der Phaseninformation unterliegt einer Reihe von Restriktionen (Kodera, et al], [Auger, Flandrin, 1995]), weshalb der rechnerische Aufwand der Implementation nicht zu unterschätzen ist, [Plante, et al].

$$E_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r_x(t,n)\,dt\,dn$$

Weiters müssen noch folgende zwei Randbedingungen erfüllt werden:

$$\int_{-\infty}^{\infty} r_x(t,n)\,dt = \left|X(n)\right|^2$$

$$\int_{-\infty}^{\infty} r_x(t,n)\,dn = \left|x(t)\right|^2$$

Durch die Integration der Zeit-Frequenz Energieverteilung über eine Variable erhält man die Energiedichte der anderen Variable.

2.3 Zusammenfassung

Die Kategorisierung der in der Natur vorkommenden Signale liefert die Basis zur mathematischen Beschreibung und zur sinnvollen Darstellung derselben. Zunächst werden Schritt für Schritt die Unterschiede zwischen stationären und nichtstationären, deterministischen und stochastischen Signalen bzw. Ein- und Mehrkomponentensignale erläutert. Mit Hilfe dieser Eigenschaften kann letztendlich jedes beliebige physikalische Signal durch eine mathematische Formel ausgedrückt werden.

Im nächsten Abschnitt wird gezeigt, welche Art der Darstellung sich für ein bestimmtes Signal am besten eignet. Gleichzeitig ergibt sich daraus die Definition der ein- und zweidimensionalen Repräsentationen. Einige Vertreter werden (im Hinblick auf die Implementation im Programm "Sound Analysis") genauer besprochen. Der Vollständigkeit halber sind Alternativen mit Literaturhinweis angeführt.

3 Analyse/Synthese Algorithmen

Als Grundlage des im Rahmen dieser Diplomarbeit entwickelten Programms dient folgende Interpretation der Audiodaten: ein beliebiges Signal kann im Zeitbereich als Summe einzelner Sinusschwingungen dargestellt werden, wobei die Teilschwingungen nichtstationär und in Amplitude, Phase und/oder Frequenz verschieden sind:

$$s(t) = \sum_{p=1}^{P} A_p(t) \cos\left[w_p(t)t + j_p(t) \right] \tag{3.1}$$

Mit dieser Art der Zerlegung erhält man exakt jene Informationen, die für die anschließende Resynthese, die AKS, notwendig sind. Das Konzept der AKS beruht auf der Wiederzusammensetzung der Einzelkomponenten zu einem einheitlichen Klang, welcher ohne Implementation einer Modifikationsstufe möglichst genau dem Original entsprechen sollte. Abb. 3.1 zeigt das Blockschaltbild der Syntheseeinheit.

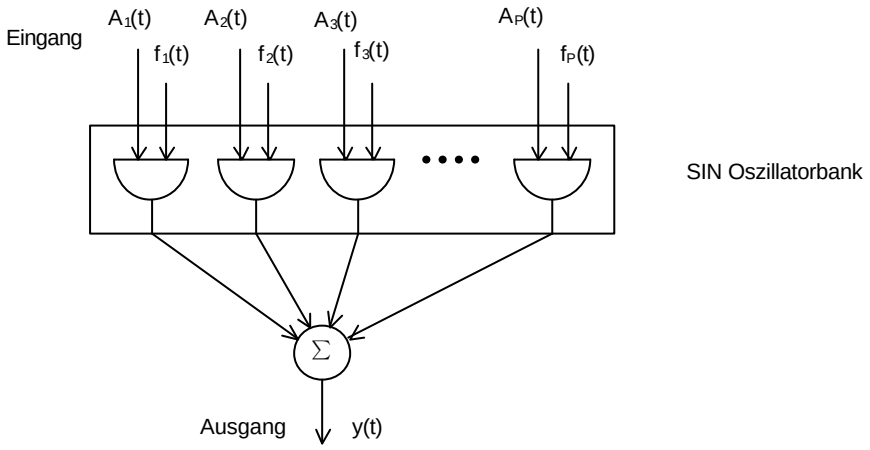


Abb. 3.1: Blockschaltbild der Additiven Klangsynthese

Im allgemeinen bestehen Audiosignale nicht nur aus genau detektierbaren Partialtönen, sondern auch aus überlagertem Rauschen und ähnlichen Artefakten. Die zeitliche Struktur dieses Anteils ist zu komplex, um mit einer sinnvollen Anzahl von Sinusoszillatoren akustisch zufriedenstellend nachgebildet werden zu können. Deshalb ist eine weitere Aufspaltung der Musik- oder Sprachdaten unerlässlich. Entsprechend der derzeit aktuellen Analyseverfahren ergibt sich somit eine getrennte Behandlung von deterministischen und stochastischen Komponenten.

$$s(t) = s_d(t) + s_s(t) \tag{3.2}$$

3.1 MQ Algorithmus

Eine der grundlegenden Techniken zur Analyse bzw. Resynthese von Sprachsignalen bietet der von McAulay und Quatieri in den achziger Jahren entwickelte MQ Algorithmus [McAulay, Quatieri]. Dabei liegen die zu analysierenden Daten in digitaler Form im Zeitbereich vor, welche mit Hilfe des Signalmodells aus Gleichung 3.1 interpretiert werden können. Jedes Audiosignal kann als Superposition von Sinusschwingungen mit den charakteristischen Parametern Amplitude, Frequenz und Phase betrachtet werden, wobei zu beachten ist, dass die einzelnen Sinuskomponenten nicht stationär, sondern zeitlich veränderlich sind. Diese zeitliche Struktur beinhaltet wesentliche

Signalinformationen und muss deshalb unbedingt in allen Verarbeitungsschritten so gut wie möglich konserviert werden.

Da die zu bearbeitenden Audiobeispiele ausschließlich in digitaler Form vorliegen, wird im weiteren Verlauf immer der zeitdiskrete Fall betrachtet. Eine Verallgemeinerung für die analoge, zeitkontinuierliche Ebene, in der Schallereignisse als zeitabhängige, mechanische Schwingungen definiert sind, ist jedoch stets zulässig. Mit den diskreten Abtastzeitpunkten $n = nT = t$ wird Gleichung 3.1 somit zu

$$s(n) = \sum_{p=1}^{P} A_p(n)\cos[w_p(n) + j_p(n)] \tag{3.3}$$

Die diskrete Frequenz $w(n)$ sowie die Phase $j(n)$ ergeben sich aus

$$w(n) = w(t)T$$

$$j(n) = j(t)T$$

3.1.1 Analyse

In der Analysestufe des MQ Algorithmus werden die Trajektorien der Amplitude, Frequenz und Phase ermittelt. Die dafür erforderlichen Schritte werden nachstehend kurz erläutert, die prinzipielle Struktur ist in Abb. 3.2 dargestellt.

- **Fensterung im Zeitbereich**

Das gesamte Signal $x(n)$ wird entlang der Zeitachse in Blöcke fixer Länge zerlegt, um eine bestimmte Zeitauflösung zu erhalten. Innerhalb eines solchen Blocks werden Amplitude, Frequenz und Phase als konstant angenommen. Diese Blöcke werden anschließend mit einer Fensterfunktion $w(n)$ multipliziert. Durch die Wahl der Fensterlänge kann die Genauigkeit der temporalen Rasterung beeinflusst werden: je kürzer die Datenblöcke, desto höher die Auflösung. Allerdings kann die Länge nicht beliebig klein gewählt werden, da diese auch gleichzeitig jener der anschließenden Fouriertransformation entspricht und hierbei aufgrund der im nächsten Schritt erläuterten Überlegung ein Konflikt entsteht.

Die Wahl der Fensterfunktion ihrerseits hat wiederum einen Einfluss auf die Frequenzauflösung der Signalkomponenten und stellt einen Kompromiss zwischen Breite der Hauptkeule bzw. Dämpfung der Nebenkeulen dar. Eine ausführliche

Diskussion bezüglich der in der Signalverarbeitung gebräuchlichsten Fenster bieten [Harris] und [Williams]. Allgemein gilt: eine schmale Hauptkeule verbessert einerseits die Genauigkeit der Frequenzbestimmung, andererseits steigt dadurch auch die Höhe der Nebenkeulen.

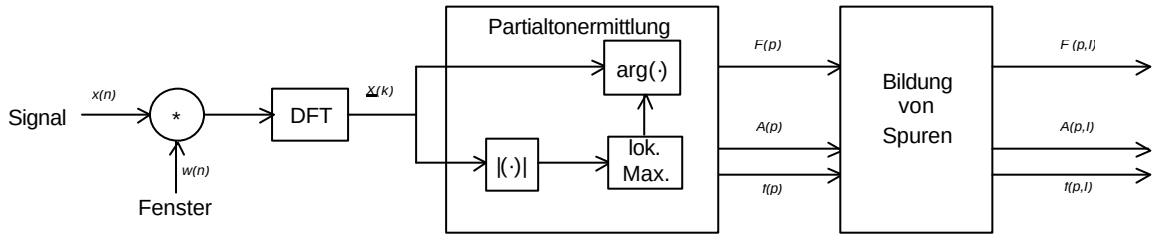


Abb. 3.2: Blockschaltbild des MQ Analysesystems

- ## Diskrete Fouriertransformation, DFT

Durch den Übergang von der Zeit- in die Frequenzebene können auf einfache Weise die komplexen, spektralen Komponenten $\underline{X}(k)$ eines Klanges extrahiert werden. Man bedient sich dabei der Fouriertransformation, wobei wegen seiner Recheneffizienz standardmäßig der sogenannte "fast Fourier transform algorithm", FFT, verwendet wird. Dieser kann allgemein sowohl auf kontinuierliche als auch auf diskrete Signale angewandt werden. Die diskrete Fouriertransformation, DFT, stellt die Implementation des FFT Algorithmus für zeit- und frequenzdiskrete Signale dar, Voraussetzung ist allerdings, dass die Länge der DFT, sprich die Anzahl der diskreten Abtastpunkte, ein vielfaches der Potenz zur Basis 2 beträgt, $N = 2^a$ mit $a \in \mathbb{N}$.

Da die Frequenz einer Schwingung dem Reziprokwert der Periodendauer entspricht, muss zur Detektion tiefer Frequenzen bzw. breiter Perioden die FFT-Länge entsprechend groß gewählt werden. Somit ergibt sich stets ein Kompromiss zwischen Zeit- und Frequenzauflösung. McAulay/Quatieri schlagen in ihrem Algorithmus vor, mindestens die 2 ½ fache Periodendauer der Grundfrequenz zu wählen.

Die Kombination dieser beiden Schritte, Unterteilung des Signals und FFT, wird in der Literatur als Kurzzeit Fouriertransformation (KZFT) definiert.

- **Abschätzung der Partialtöne**

Zunächst wird vom komplexen Spektrum $\underline{X}(k)$ der Betrag $|X(k)|$ berechnet. Alle ausgeprägten lokalen Maxima des Betragsspektrums werden als Teiltonkandidaten markiert und müssen keine weiteren Kriterien, wie zum Beispiel das Überschreiten einer minimalen Amplitudenschwelle, erfüllen. Die Werte für Amplitude $A(p)$, Phase $\Phi(p)$ und Frequenz $f(p)$ werden direkt aus dem komplexen FFT Spektrum ausgelesen, wobei p als Bezeichnung eines Partialtons dient.

- **Bildung von Spuren**

Um die zeitliche Entwicklung der einzelnen Komponenten besser verfolgen zu können, werden sogenannte Spuren gebildet. Das bedeutet, dass eine Verbindung zwischen den Kandidaten in benachbarten Zeitfenstern unter Zuhilfenahme bestimmter Kriterien hergestellt wird und diese durch einen gemeinsamen Spurindex I gekennzeichnet werden. Man erhält dadurch eine Abhängigkeit der Analyseparameter von p und I: $A(p,I)$, $\Phi(p,I)$ und $f(p,I)$. Die einfachste Art der Spurzuteilung, welche auch im MQ Algorithmus implementiert ist, erfolgt durch einen Frequenzvergleich. Kandidaten, deren Frequenz sich in aufeinanderfolgenden Zeitpunkten innerhalb der zulässigen Grenzen verändern, werden einem Partialton zugeordnet und miteinander verbunden.

Prinzipiell ergeben sich drei unterschiedliche Fälle während der Fortführung der Spuren:

- Eine bereits vorhandene Spur wird fortgesetzt, wenn im aktuellen Zeitframe ein Anwärter gefunden wird, der dem Frequenzkriterium genügt.
- Findet sich für eine bereits vorhandene Spur kein passender Treffer im aktuellen Frame, so wird diese Spur beendet.
- Alle übrigen, noch nicht einer Spur zugeteilten Komponenten, erzeugen eine neue Spur.

3.1.2 Synthese

Die Synthesestufe besteht aus einer Reihe von Sinusoszillatoren, deren Anzahl der maximal detektierten Teiltonanzahl P entspricht. Jede Sinusschwingung wird entsprechend den vorangegangenen Analysewerten amplitudenmoduliert sowie mit der zughörigen Phase versehen. Zur Aufbereitung der Amplitude erfolgt eine lineare Interpolation zwischen frameweise detektierten Werten $A'(p,I)$. Um einen maximal stetigen Phasenverlauf $\Phi'(p,n)$ zu erhalten, wird eine kubische Interpolation der Phasenwerte $\Phi'(p,I)$ und eine anschließende Phasenkorrektur durchgeführt. Die Frequenz muss ebenfalls zwischen den Analysewerten $f'(n,I)$ interpoliert werden. Das nachgebildete synthetische Signal $y(n)$ resultiert letztendlich aus der Summation aller Oszillatorausgänge $y(n,p)$ und sollte das Originalsignal ohne hörbare Unterschiede nachbilden. Darüber hinaus verspricht der MQ Algorithmus robustes Verhalten gegenüber dem Nutzsignal überlagertem Rauschen. Ursprünglich wurde er für Sprachsignale konzipiert, kann aber wegen seiner generellen Struktur auch auf Musik und diverse andere Signale angewandt werden.

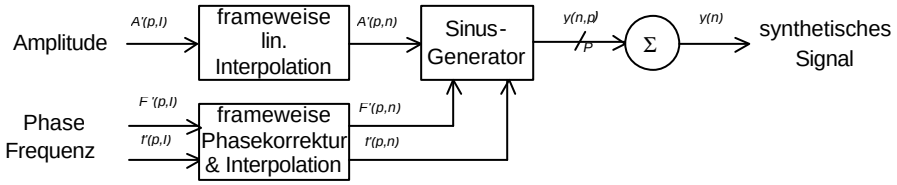


Abb. 3.3: Blockschaltbild des MQ Synthesesystems

3.2 SMS Algorithmus[1]

Im Gegensatz zu dem im MQ Algorithmus verwendeten Signalmodell wird hier eine getrennte Behandlung von deterministischem und stochastischem Teil vorgenommen (analog Gleichung 3.2). Deshalb wird das Eingangssignal $s(n)$ wie

[1] SMS steht für: spectral modeling synthesis

folgt interpretiert (es wird wiederum davon ausgegangen, dass die Daten in zeitdiskreter Form vorliegen):

$$s(n) = s_d(n) + s_s(n) = \sum_{p=1}^{P} \left\{ A_p(n) \cos\left[Q_p(n)\right] \right\} + r(n) \tag{3.4}$$

Das Modell der Summation von Sinusschwingungen ist zum Reproduzieren von Rauschen unbrauchbar, da einerseits aufgrund seiner breitbandigen Natur die Anzahl der Oszillatoren in der Synthesestufe extrem groß wäre. Dadurch steigt natürlich auch der Rechenaufwand. Andererseits ist die zeitliche Struktur eines Rauschsignals willkürlich und nicht vorhersehbar, somit ist die Bildung von Spuren praktisch nicht sinnvoll möglich. Allerdings sind bei einem stochastischen Signal weder die exakte Momentanamplitude noch die -phase von Bedeutung, weshalb zur Modellierung lediglich die Einhüllende der Amplitude verwendet wird. Die Nachbildung des breitbandigen Signals erfolgt somit mit Hilfe von gefiltertem weißen Rauschen.

3.2.1 Analyse

Die Abschätzung des deterministischen Anteils erfolgt auf ähnliche Weise wie im MQ Algorithmus, abgesehen von der getrennten Behandlung des rauschähnlichen Signalanteils. Wiederum wird das gesamte Zeitsignal mittels Fensterung in Blöcke unterteilt und von jedem die KZFT gebildet. Zur Verbesserung der Zeitauflösung erlaubt der SMS Algorithmus eine zeitliche Überlappung dieser Blöcke, die durch die sogenannte Hopsize definiert ist. Eine schematische Darstellung der Überlappung zeigt Abb. 3.4.

Ein weiteres Detail, welches in diesem Formalismus integriert ist, bietet die Möglichkeit, durch Anhängen von Nullen an den gefensterten Ausschnitt eine Art spektrale Interpolation zu erzielen, was eine exaktere Frequenzbestimmung zulässt. Die Auswirkung des sogenannten "zero-padding" ist in Abb. 3.5 zu sehen. Die strichpunktierte Linie kennzeichnet die Frequenzantwort eines von-Hann-Fensters, dessen FFT Länge genau der Länge der Funktion entspricht. Anhand der durchgezogenen Linie ist deutlich die Interpolation zwischen den Stützstellen k und der daraus resultierende verbesserte Kurvenverlauf zu erkennen. Da die FFT Länge das 8fache der originalen Signallänge beträgt, spricht man von 8fach zero-padding und erhält somit die 8fache Auflösung der Darstellung.

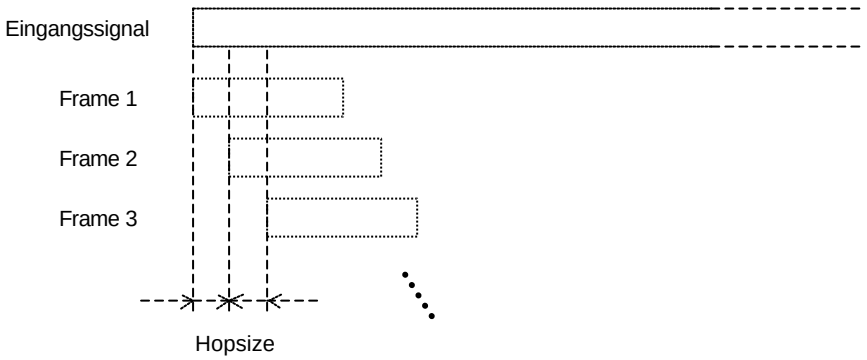


Abb. 3.4: Bildung zeitlich überlappender Frames

Vom Spektrum werden anschließend alle lokalen Maxima markiert und einem genauerem Auswahlverfahren unterzogen, da nicht alle ausgeprägten Spitzen gleich signifikant sind. Es werden sowohl für Amplitude als auch für die Frequenz Gültigkeitsbereiche definiert, innerhalb derer die entsprechenden Parameter der Sinuskomponenten liegen müssen.

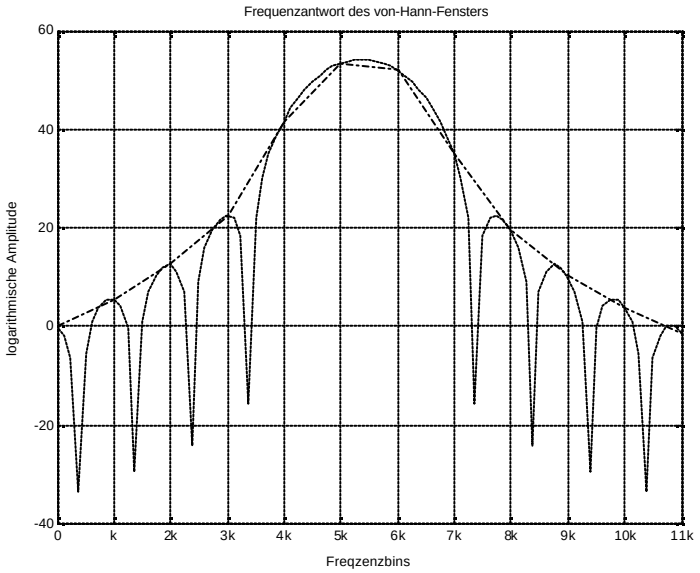


Abb. 3.5: Zero-padding

Die Generierung von Spuren zur zeitlichen Gruppierung der blockweise ermittelten Teiltöne ist in beiden Algorithmen ident. Serra/Smith schlagen zusätzlich eine Variation dieses Analyseteils für harmonische Klänge vor, in der primär eine Grundfrequenz bestimmt und mit den restlichen Spuren ein harmonisches Gefüge aufgebaut wird. Kann ein detektierter Kandidat aufgrund

seiner Tonhöhe bzw. Frequenz nicht in diese Struktur eingegliedert werden, so scheidet er aus und wird nicht weiter berücksichtigt.

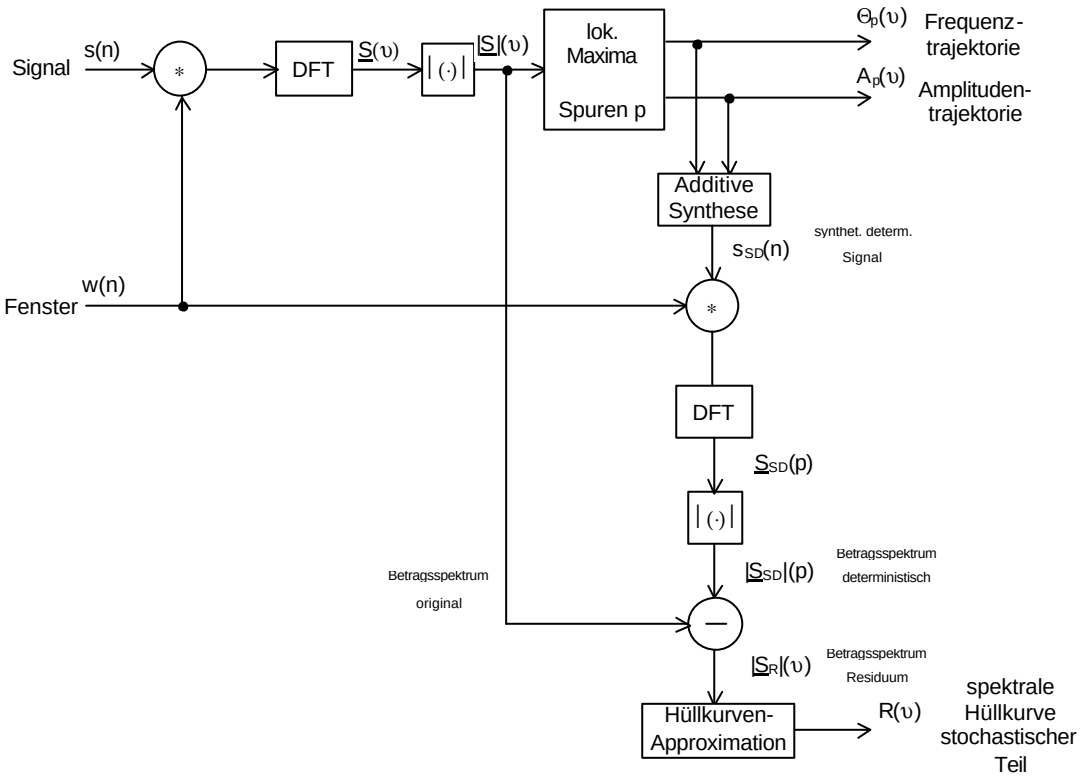


Abb. 3.6: Blockschaltbild des SMS Analysesystems

Nach der Abschätzung des deterministischen Signalanteiles erfolgt eine spektrale Subtraktion vom Original, woraus der stochastische Part resultiert. Die Notwendigkeit der Subtraktion im Frequenzbereich ergibt sich aus der Tatsache, dass in der Analysestufe die Bestimmung der Phase nicht berücksichtigt wurde. Diese wäre hingegen für eine Berechnung des Residuums im Zeitbereich unabkömmlich.

Das Restspektrum enthält Informationen über die generelle Frequenzcharakteristik bzw. die Hüllkurve der Amplitude, welche zusätzlich mittels einfacher, stückweise linearer Interpolation geglättet wird und nach Angabe von Serra/Smith für den vorliegenden Algorithmus ausreichend ist. Dabei wird jedes einzelne Spektrum in Q Abschnitte unterteilt, von jedem das lokale Maximum gesucht und zwischen diesen Punkten linear interpoliert.

Zusammenfassend ist die Wirkungsweise der Analyseeinheit in Abb. 3.6 grafisch dargestellt.

3.2.2 Synthese

Aus den Parametertrajektorien der Amplitude $A_p(n)$ und Frequenz $Q_p(n)$ wird mittels additiver Synthese wiederum ein Zeitsignal generiert und, analog Gleichung (3.3), die spektrale Hüllkurve $r(n)$ des stochastischen Signalanteils hinzuaddiert.

Im Unterschied zum MQ Algorithmus, der die Phase der einzelnen Sinuskomponenten getrennt analysiert und speichert, berechnet der vorliegende Formalismus die Momentanphase direkt aus der Momentanfrequenz durch Integration dieser. Detailliertere Informationen über den genauen Ablauf der Synthese des deterministischen Teils finden sich in [Serra, Smith]. Der prinzipielle Aufbau der Syntheseeinheit ist in Abb. 3.7 zu sehen, zusätzlich ist die Möglichkeit der Parametervariation durch die drei "Modifikationsblöcke" angedeutet.

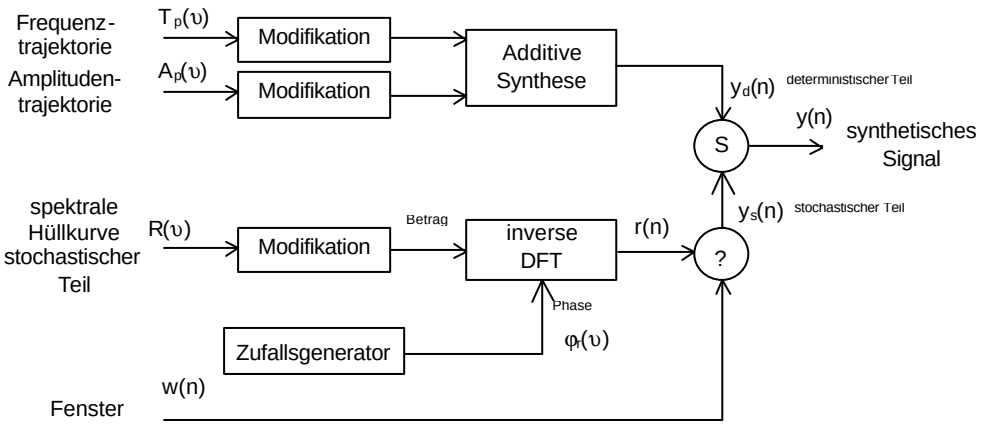


Abb. 3.7: Blockschaltbild des SMS Synthesesystems

Die Synthese des stochastischen Signals kann als Erzeugung von gefärbtem Rauschen verstanden werden, wobei seine Eigenschaften in der Repräsentation der spektralen Hüllkurven liegen. Intuitiv würde man nun einfach weißes Rauschen mit diesen Einhüllenden filtern, praktisch realisiert das SMS Verfahren das stochastische Signal jedoch durch die sogenannte "overlap-add" Technik. Von

jedem komplexen Spektrum wird der Betrag als Amplitudenabschätzung verwendet, wohingegen die Phase mittels Zufallsgenerator erzeugt wird, da sie keine notwendigen klanglichen Informationen enthält. Mittels inverser Fouriertransformation werden die entsprechenden Funktionen im Zeitbereich erzeugt, neuerlich gefenstert, wobei Synthese- und Analysefenster nicht ident sein müssen, und anschließend unter Berücksichtigung der Hopsize addiert.

4 Aufbau des Programms

Dieses Kapitel widmet sich der Gliederung der vorgeschlagenen Analysetechnik. Es beinhaltet eine Auflistung der implementierten Stufen, wie sie in Abb. 4.1 schematisch dargestellt sind, sowie eine kurze Erläuterung dieser. Weiters wird ein Überblick über die prinzipielle Funktionsweise der AKS gegeben.

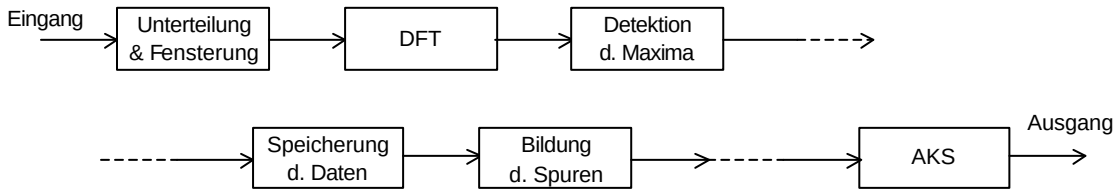


Abb. 4.1: Illustration der Analysestufen

4.1 Allgemeines

Das Programm "SOUND ANALYSIS" bestimmt von einem Zeitsignal zu äquidistanten Abtastzeitpunkten mögliche Kandidaten für seine Partialtöne. Als mögliche Kandidaten gelten in diesem Zusammenhang jene lokalen Maxima des Betragsspektrums, die aufgrund bestimmter Kriterien als tatsächliche Teiltöne des Originalsignals angenommen werden. Ein typisches Kriterium stellt zum Beispiel der Betrag der Amplitude des lokalen Maximums dar: Liegt dieser über einem gewissen Schwellwert, wird das Maximum als Teilton gekennzeichnet, andernfalls

ist es für die weiteren Berechnungen ohne Bedeutung, siehe Abb. 4.2. Die Definition der Kriterien sowie die damit verbundene Erkennung der Teiltöne findet in der Frequenzebene statt, weshalb als erster Schritt die Transformation vom Zeit- in den Frequenzbereich vorgenommen werden muss. Dies geschieht mit Hilfe der KZFT. Zur Beschreibung der Partialtöne dienen nicht nur die jeweiligen Amplituden, Frequenzen und Phasen, sondern zusätzlich auch deren zeitliche Entwicklung. Zu diesem Zweck werden sogenannte Spuren gebildet, die die momentanen Daten zu jedem Zeitpunkt unter festgelegten Bedingungen miteinander verbinden. Damit ist das Eingangssignal ausreichend genau charakterisiert und die Analysestufe abgeschlossen. Die gespeicherten Werte können in weiterer Folge der Additiven Klangsynthese entweder direkt oder über eine Modifikationsstufe zugeführt werden.

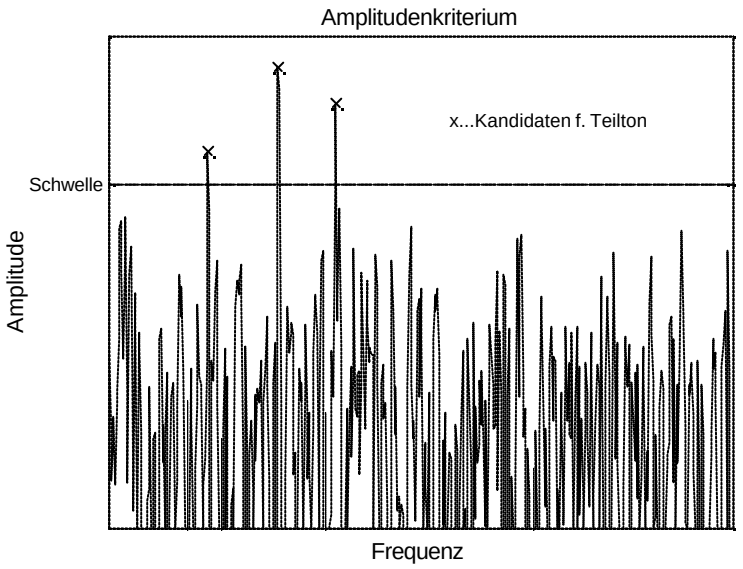


Abb. 4.2: Ermittlung der Teiltöne eines Signals

4.2 Berechnung der KZFT

- **Unterteilung des Eingangssignals**

  Das Eingangssignal $x_{in}(n)$ wird im Zeitbereich in einzelne Blöcke zerlegt, welche im weiteren als Frames bezeichnet werden:

$$\tilde{x}_m(n) \triangleq x_{in}(n - m \cdot hs)$$

wobei m die Framenummer angibt und $\tilde{x}_m$ somit das m-te Frame des Eingangs bezeichnet. n entspricht den Abtastpunkten des diskreten Signals[1] und hs bestimmt die Schrittweite zwischen den Frames, die Hopsize. Wird die Länge eines Frames mit N angegeben, so gilt für n:

$$n = \left[-\frac{N}{2}, \frac{N}{2} - 1 \right]$$

- **Fensterung der Signalframes**

Ebenfalls im Zeitbereich erfolgt die Multiplikation der Frames $\tilde{x}_m$ mit einer Fensterfunktion w:

$$x_m(n) = \tilde{x}_m(n) \cdot w(n)$$

- **Fouriertransformation**

Die Berechnung der Fouriertransformierten der gefensterten Daten x_m erfolgt unter der Verwendung der DFT:

$$X_m(k) = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} x_m(n) \cdot e^{-\frac{j2\pi nk}{N}}$$

mit $k = \left[-\frac{N}{2}, \frac{N}{2} - 1 \right]$ ... spektrale Bins

und $X_m(k)$ ... komplexes Spektrum des m-ten Frames.

Die Umrechnung von einem beliebigen diskreten Bin k_l auf die entsprechende analoge Frequenz f_l erfolgt mittels:

$$f_l = k_l \frac{Fs}{N} [Hz],$$

wobei Fs die Abtastfrequenz oder Samplingfrequenz in Hz darstellt.

Die Länge der DFT muss im allgemeinen nicht mit der Fensterlänge übereinstimmen, jedoch muss sie zur Verhinderung von Aliasing stets größer sein. In diesem Fall wird das gefensterte Signal durch das Anhängen von Nullen (zero

[1] In dieser Arbeit hängt der Begriff "Abtastpunkt" ausschließlich mit der Auflösung des Signals im Zeitbereich zusammen. Im Unterschied dazu werden die "spektralen Abtastpunkte" als Bins bezeichnet.

padding) auf die DFT Länge erweitert. Einerseits wird zero padding dann gebraucht, wenn die Länge des Fensters nicht 2^a mit $a \in \mathbb{N}$ entspricht (siehe Kapitel 3.1.1) und somit die Recheneffizienz des FFT Algorithmus nicht ausgenützt werden kann. Andererseits erreicht man durch Vergrößerung der DFT Länge um einen Faktor *zp* eine Verbesserung der spektralen Auflösung (siehe Abb. 3.5), wobei gilt:

$$N_{DFT} = zp \cdot N_{Fenster}$$

4.3 Detektion der Kandidaten

- **Betragsspektrum**

  Das logarithmische Betragsspektrum $X_{mag}(k)$ je Frame erhält man mit:

  $$X_{mag}(k) = 20\log\left|X(k)\right| \ [dB]$$

  Um die Schreibweise zu vereinfachen, wird auf den Frameindex *m* verzichtet.

- **Lokale Maxima**

  Ein lokales Maximum $\hat{X}_{mag}$ innerhalb eines Frames ist definiert durch:

  $$X_{mag}(k-1) \le \hat{X}_{mag}(k) \ge X_{mag}(k+1)$$

- **Validität der Maxima**

  Die Entscheidung, ob ein Maximum möglicher Kandidat für einen Partialton ist, wird durch zwei Bedingungen fixiert:

  Kriterium K1:

  Das Verhältnis von Maximum zu links- und rechtsseitigem Minimum, X_{mag-}, X_{mag+}, muss einen festgelegten Grenzwert L_1 erreichen:

  $$\hat{X}_{mag}(k) - \frac{1}{2}[X_{mag-}(k) + X_{mag+}(k)] \ge L_1 \qquad (4.1)$$

  Kriterium K2:

Es wird eine globale Schwelle L_2 definiert, die ebenfalls erreicht werden muss:

$$\hat{X}_{mag}(k) \geq L_2$$

Die Bins jener Kandidaten, welche beide Bedingungen erfüllen, werden gekennzeichnet:

$$p = k \ \forall \left\{ \hat{X}_{mag}(k) \geq (L_1 \wedge L_2) \right\}$$

4.4 Ermittlung von Amplitude, Frequenz und Phase

Unter Berücksichtigung eines Korrekturfaktors A_{cor}, welcher aufgrund der Fensterung des Signals erforderlich ist, kann die Amplitude jedes Kandidaten p_m in Frame m sofort angeschrieben werden:

$$X_{p,m} = \frac{1}{A_{cor}} 10^{\frac{\hat{X}_{mag,m}(p)}{20}}$$

Die Frequenz berechnet sich aus p_m mit:

$$f_{p,m} = p_m \frac{Fs}{N} \tag{4.2}$$

Somit ist die Frequenzauflösung mit $Df = \frac{Fs}{N}$ beschränkt, jedoch kann sie durch weitere Berechnungen verbessert werden. Zwei mögliche Algorithmen werden im Folgenden erklärt. Ähnliches gilt für die Genauigkeit der Amplitudenabschätzung, die an die korrigierte Frequenz angepasst wird und ebenfalls zu einem späteren Zeitpunkt genauer diskutiert wird.

Die Phase errechnet sich primär aus dem komplexen Spektrum jedes Frames:

$$F(k) = \arg\{X(k)\}$$

und wird in Analogie zur Amplitude emendiert.

4.5 Bildung von Spuren

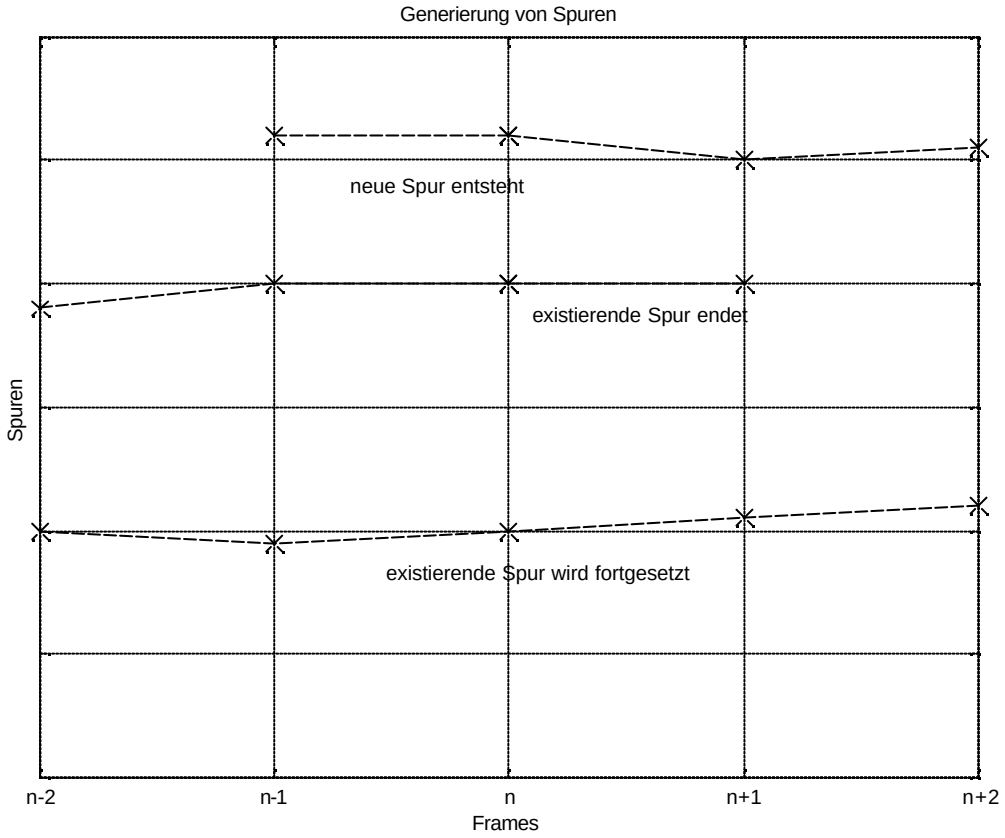


Abb. 4.3: Verknüpfung der Analysedaten

Das Zusammenschließen von Kandidaten, welche während ihrer zeitlichen Ausdehnung ähnliche Eigenschaften aufweisen, führt schließlich zur Bestimmung der Partialtöne des Signals. Dabei ist im Normalfall die Beobachtung der Frequenztrajektorien ausreichend, genauer gesagt werden alle Kandidaten von zwei aufeinanderfolgenden Frames miteinander verglichen. Befindet sich ein Kandidat des aktuellen Frames im Toleranzbereich einer bereits bestehenden Spur, wird er dieser zugeordnet und dadurch die Spur weitergeführt. Spuren, die im aktuellen Frame keinen "Treffer" finden, werden beendet, indem ihre Amplitude auf Null gesetzt wird. Äquivalent dazu können Spuren entstehen, wenn Kandidaten des aktuellen Frames übrigbleiben, das heißt, wenn für sie noch keine passende Spur existiert.

4.6 Additive Klangsynthese

Die Trajektorien der Amplitude, Frequenz und Phase jedes Partialtons dienen als Kontrollvariablen der Sinusoszillatoren der additiven Synthese. Für die allgemeine Anwendung der AKS kann auf die "Detektion der Partialtöne" verzichtet werden. In diesem Fall gelten alle lokalen Maxima des KZFT Spektrums als Teiltöne, deren Parameter keiner zusätzlichen Korrektur unterzogen werden, sondern direkt aus dem Spektrum extrahiert.

Zwischen den Amplituden- und Frequenzwerten von zwei aufeinanderfolgenden Frames wird linear interpoliert, um einen stetigen Verlauf ihrer Funktion zu erhalten. Mit $m^I = (m-1)\cdot hs$ und $m^{II} = m\cdot hs$ gilt:

$$A_k(l) = A_k(m^I) + \frac{A_k(m^{II}) - A_k(m^I)}{hs} l$$

$$f_k(l) = f_k(m^I) + \frac{f_k(m^{II}) - f_k(m^I)}{hs} l$$

$$l = 1,2,\ldots,hs-1$$

Hierbei kennzeichnet l die innerhalb eines Frames liegenden Abtastpunkte, denen die Amplitude $A_k(l)$ bzw. die Frequenz $f_k(l)$ zugewiesen wird.

Häufig wird während der Analyse auf eine exakte Berechnung der Momentanphase verzichtet, in der Synthese führt man diese dann über die Momentanfrequenz aus:

$$Q_k(n) = Q_k(n-1) + 2p\, f_k(n)T$$

In diesem Fall gilt n als allgemeiner Abtastpunkt, k dient als Index für einen beliebigen Frequenzbin im Spektrum, T entspricht der Periodendauer und ist gleich dem Kehrwert der Abtastfrequenz $T = \frac{1}{Fs}$. Zum Zeitpunkt $n = n_0$ gilt:

$$Q_k(n_0) = Q_{k,0}$$

Wird eine Abschätzung der Phase je Partialton und Frame in der Analyse miteinbezogen, gestaltet sich die Ermittlung der Momentanphase aufgrund folgender Tatsache als weitaus komplexer. Über die Beziehung

$f(n)=\frac{1}{2p}\left[F(n)-F(n-1)\right]$ sind Frequenz und Phase fix miteinander verbunden[2]. Somit wirken sich insgesamt vier Variablen auf die Momentanphase aus: $f(m^I)$, $f(m^{II})$, $F(m^I)$ und $F(m^{II})$. Eine lineare Interpolation ist in diesem Fall nicht mehr möglich, man benötigt eine Funktion mit mindestens drei Freiheitsgraden, wie sie ein Polynom dritter Ordnung besitzt:

$$y(x)=a_0+a_1x+a_2x^2+a_3x^3$$

Die Lösung dieser Gleichung in Bezug auf $Q_k(l)$ wird in [McAulay, Quatieri] im Detail erklärt und führt letztendlich zu folgendem Ergebnis:

$$Q(l)=F(m^I)+2p\ f(m^I)n+a_2n^2+a_3n^3$$

wobei die Parameter a_2 und a_3 mit Hilfe der Randbedingungen an den Framegrenzen berechnet werden können:

$$a_2(M)=\frac{3}{hs^2}\left[F(m'')-F(m')-2p\ f(m')hs+2p\,M\right]-\frac{2p}{hs}\left[f(m'')-f(m')\right]$$

$$a_3(M)=-\frac{2}{hs^3}\left[F(m'')-F(m')-2p\ f(m')hs+2p\,M\right]+\frac{2p}{hs^2}\left[f(m'')-f(m')\right]$$

Daraus ergibt sich ein Satz von Interpolationsfunktionen, die von der Variable M abhängig sind. Damit die Momentanphase einen maximal flachen Verlauf hat, wird M als dem Wert c nächstgelegene ganze Zahl gewählt, wobei

$$c=\frac{1}{2p}\left\{\left[F(m')+2p\ f(m')hs-F(m'')\right]+p\,hs\left[f(m'')-f(m')\right]\right\}$$

Die Summation der einzelnen Sinusschwingungen, welche in der Oszillatorbank erzeugt werden (vgl. Abb. 3.1), liefert folgendes zeitliche Signal:

$$y(n)=\sum_{p=1}^{P}A_p(n)e^{jQ_p(n)}\equiv x_{in}(n)$$

wobei A_p und Q_p gleich der Amplitude bzw. Phase des p-ten Partialtons sind und $p=1,2,...,P$. Folgt die Syntheseeinheit ohne Zwischenstufe der Analyseeinheit, so sollte $y(n)$ ohne wahrnehmbaren Unterschiede dem Original gleichzusetzen sein.

[2] Dies ergibt sich aus der allgemeinen Beziehung: $f(t)=\frac{\partial j(t)}{\partial t}$.

Wird eine Modifikation der Analysedaten (zeitliche Dehnung/Stauchung, Tonhöhenveränderung, Formantverschiebung) durchgeführt, so stehen der Synthesestufe die entsprechend abgeänderten Parameter $\tilde{A}_p$ und $\tilde{Q}_p$ als Eingangsgrößen zur Verfügung:

$$\tilde{y}(n) = \sum_{p=1}^{P} \tilde{A}_p(n) e^{j\tilde{Q}_p(n)} .$$

5 Beschreibung der einzelnen Analysestufen

In Anlehnung an das von J. O. Smith und X. Serra entwickelte Programm "PARSHL" [Smith, Serra] wird ein Konzept zur Analyse von Audiosignalen vorgeschlagen, welches durch seinen strukturierten Aufbau äußerst flexibel zu handhaben ist. Es wurde für keinen speziellen Signaltyp ausgelegt, sondern versucht, möglichst universell einsetzbar zu sein. Trotzdem sollten bestimmte Kriterien beachtet werden, um sinnvolle Analysedaten für die Weiterverarbeitung zu erhalten. Zum einen ist durch die vorgegebenen Einstellwerte der KZFT-Länge eine untere Grenzfrequenz bzw. das spektrale Auflösevermögen vorgegeben, wobei auf letzteres zusätzlich noch die gewählte Fensterfunktion einen Einfluss hat. In Abschnitt 5.7 sind die zur Verfügung stehenden Einstellmöglichkeiten des Programms aufgelistet. Andererseits ist die Detektion von Transienten ebenfalls durch die Länge der KZFT begrenzt. Weitere konkretere Eigenschaften für geeignete Eingangssignale sind im nächsten Kapitel angeführt, indem die Vor- und Nachteile des Algorithmus diskutiert und Verbesserungsvorschläge bzw. Alternativen angeboten werden. Dieses Kapitel widmet sich einer umfassenden Dokumentation der einzelnen Programmabschnitte.

5.1 Das Analysefenster

Bereits in diesem ersten Schritt werden wesentliche Entscheidungen getroffen, die sich auf die Performance des gesamten Analyseapparates auswirken. Die Wahl der Fensterfunktion, die zur Aufspaltung eines sehr langen Datenvektors in kleinere Blöcke dient, stellt in der Signalverarbeitung ein viel diskutiertes Thema dar, weshalb gewisse Grundkenntnisse im Umgang mit ihnen vorausgesetzt werden.

Das wichtigste Kriterium bei der Wahl des Fensters bildet der Kompromiss zwischen der zeitlichen bzw. spektralen Auflösung des zu modellierenden Signals, der durch die Dualität von Zeit und Frequenz gegeben ist. Eine exakte Untersuchung in beiden Ebenen ist deshalb unbedingt erstrebenswert, da das menschliche Gehör mit hoher Sensitivität auf jede einzelne reagiert. Dies lässt sich am einfachsten durch folgende Tatsache verdeutlichen: Der Wiedererkennungswert eines Instrumentes hängt zu einem beachtlichen Teil vom Anschlag bzw. Einschwingvorgang des Tons ab, welcher bekanntlich sehr kurz ist und eine Vielzahl verschiedener Frequenzen beinhaltet.

5.1.1 Beeinflussung der Zeit- und Frequenzauflösung durch die Fensterung

Betrachtet man die Auswahl des Fenstertyps zunächst von der spektralen Seite aus, treten vor allem zwei Charakteristika in den Vordergrund: Zum Einen die Breite der Hauptkeule, welche als Anzahl der Bins der maximalen Erhebung zwischen zwei Nulldurchgängen im Spektrum definiert ist, und zum Anderen die Höhe der größten Nebenkeulen, welche den Abstand zur Hauptkeule in Dezibel kennzeichnet.

Wünschenswert wäre eine möglichst schmale Hauptkeule (damit ergibt sich eine gute Auflösung) und eine geringe Höhe der Nebenkeulen (da sich diese negativ auf die Analyse auswirken). Diese Forderung ist anhand von Abb. 5.1 zu erkennen: ein Signal bestehend aus zwei Sinuskomponenten, deren Frequenzen eng beieinander liegen, wird einmal mit einem Hamming Fenster (durchgezogene Linie) und einem Rechteck Fenster (strichliert) gefenstert. Das Hamming Fenster

kann aufgrund seiner breiteren Hauptkeule die Teilschwingungen nicht trennen. Dies gelingt zwar dem Rechteck Fenster, jedoch sind dessen Nebenkeulen so hoch, dass eine Unterscheidung zwischen Hauptkeule und ersten Nebenkeulen sehr schwierig ist. Dadurch wird eine weitere Kompromisslösung notwendig, die in den zahlreichen Fensterrealisationen auf unterschiedliche Weise behandelt wird. Die am häufigsten verwendeten Fensterarten sind Rechteck, Dreieck, Hamming, von Hann und Kaiser, wobei das Rechteckfenster die schmalste Hauptkeule (2 Bins) von allen aufweist. Allerdings ist die Distanz von Hauptkeule zur höchsten Nebenkeule mit 13dB in vielen Fällen nicht ausreichend, weshalb auf andere Formen übergegangen wird. Mit einer 4 Bin breiten Hauptkeule und 43dB Dämpfung der Nebenkeulen stellt das Hammingfenster eine brauchbare Alternative dazu dar. Allgemein gilt: je schmäler die Hauptkeule, desto höher die Nebenkeulen und umgekehrt. Die genannten Fensterfunktionen sind in Abb. 5.2(a)-(e) dargestellt.

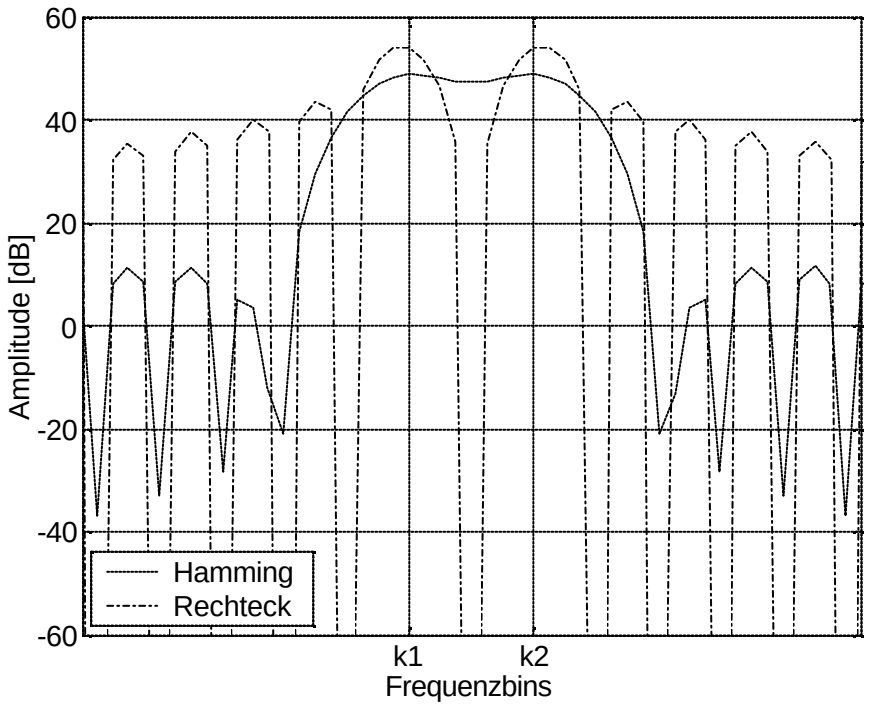


Abb. 5.1: Spektrum zweier eng beieinander liegender Sinuskomponenten

Wird der geforderte minimale Frequenzabstand zwischen zwei im Signal vorkommenden Teiltönen mit Df Hz vorgegeben, so gilt für die Hauptkeulenbreite B_{HK} in Hz die Bedingung:

$$B_{HK} \leq Df$$

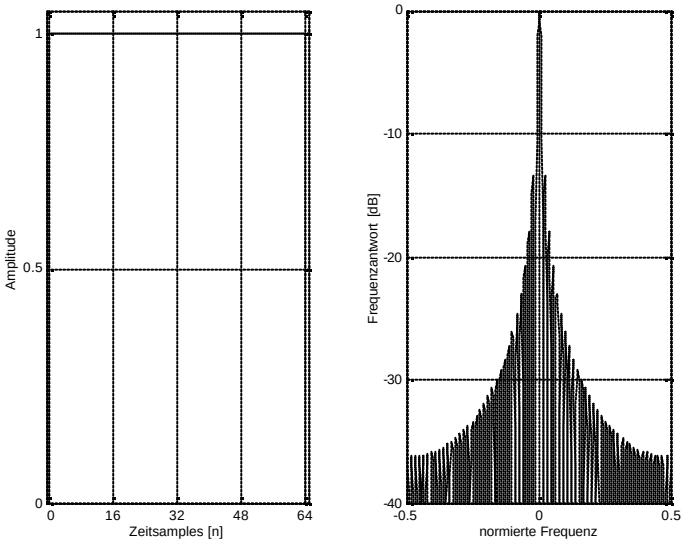
bzw. mit

$$K_{HK} = B_{HK} \frac{N}{Fs}$$

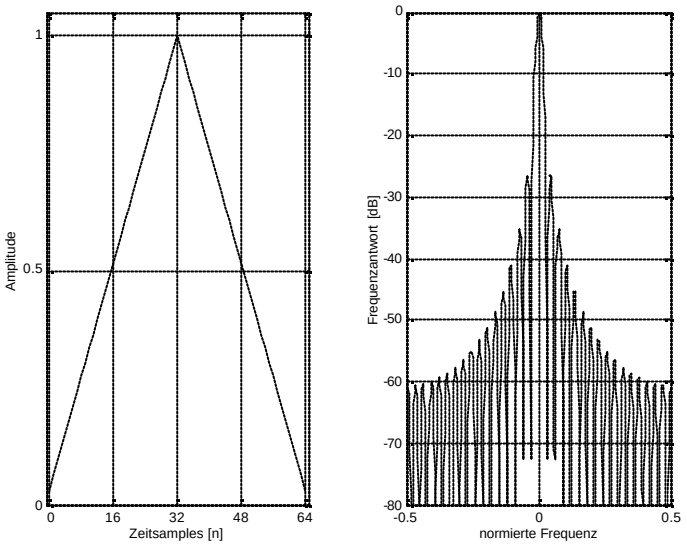
erhält man die Bandbreite in Bins, K_{HK}, wobei mit N und Fs zusätzlich die Länge des Fensters sowie die Abtastfrequenz einwirken. Nimmt man K_{HK} und Fs als fixe Größen an, so ist nur mehr N von Df abhängig:

$$N \geq K_{HK} \frac{Fs}{Df}$$

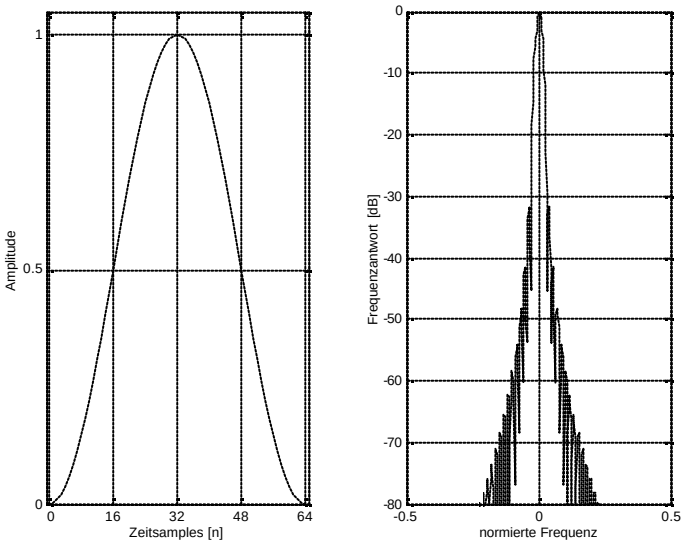
das heißt, je kleiner der minimale Frequenzabstand Df zweier Töne im Spektrum, desto länger muss das Analysefenster gewählt werden.



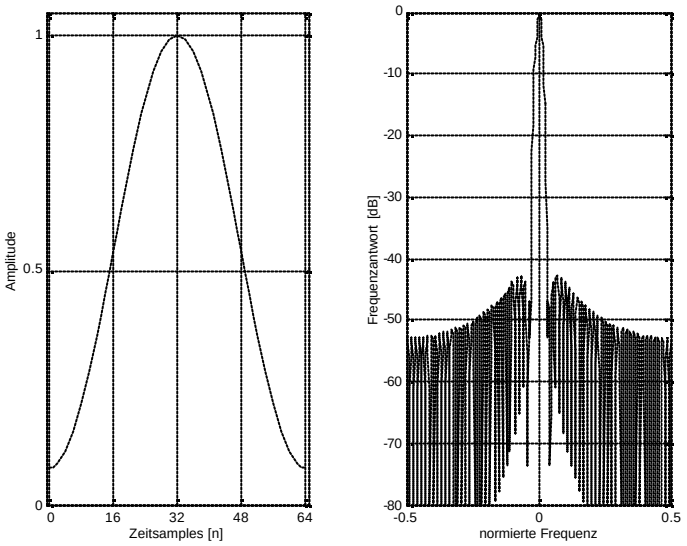
(a) Rechteck Fenster



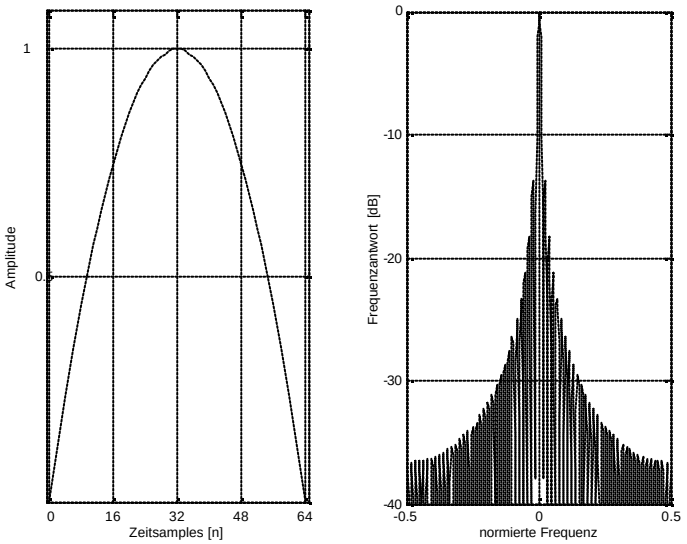
(b) Dreieck Fenster



(c) von Hann Fenster



(d) Hamming Fenster



(e) Kaiser Fenster

Abb. 5.2: Verschiedene Fenstertypen

Anders formuliert entspricht Df einer Periodendauer von $T_{Df} = \frac{1}{Df}$, womit bei der Verwendung eines Hammingfensters ($K_{HK} = 4$) von einem Signal mindestens das Vierfache von T_{Df} unterhalb des Fensters liegen müsste.

Wechselt man nun die Sichtweise und betrachtet das Signal hinsichtlich seiner zeitlichen Struktur, führen obige Überlegungen zu folgendem Resultat: Der kleinste aufzulösende Frequenzunterschied im Signal wird mit 40 Hz angenommen, womit sich bei einer Abtastfrequenz von 44.1kHz und einer Hammingfunktion eine Fensterlänge von $N = 4\frac{44100}{40} = 4410\,\text{Punkten}$ ergibt. Unterteilt man das gesamte Signal in Blöcke dieser Länge, erhält man eine Zeitauflösung von 100ms, was mehr als dem Zehnfachen des Auflösevermögens des menschlichen Gehörs entspricht und damit wenig zufriedenstellend ist. Zur Verbesserung können entweder die bereits bekannten Parameter variiert werden, oder man versucht, mit alternativen Ansätzen dieses Problem zu umgehen. Eine einfache Möglichkeit ergibt sich in der Überlappung der Zeitfenster und führt zur Definition der Hopsize. Sie gibt die Anzahl der Abtastwerte an, mit der sich aufeinanderfolgende Blöcke überschneiden und es ist sofort ersichtlich, dass die Erhöhung der Auflösung proportional mit der Hopsize steigt. Allerdings steigt damit ebenfalls die Rechenkomplexität, sodass sich im allgemeinen die Wahl der Hopsize nach der Art des Eingangssignals richtet. Ändert sich dieses nur langsam mit der Zeit, so kann die Hopsize entsprechend lange gewählt werden, umgekehrt muss sie bei raschen Signaländerungen verkleinert werden. Obwohl eine adaptive Realisation somit die beste Implementation darstellen würde, verwendet der vorliegende Algorithmus einen fix einstellbaren Wert[1].

5.1.2 Auswirkung der Fensterung auf die Amplitude

Am Beispiel eines analytischen Signals $x_a(n)$ mit konstanter Amplitude und Frequenz soll der Effekt des Fensters hinsichtlich der Amplitude erläutert werden:

$$x_a(n) = Ae^{j\frac{2\pi nk_x}{N}}$$

Die dazugehörige gefensterte Fouriertransformierte resultiert aus:

$$X_a(k) = \sum_{n=-\infty}^{\infty} x_a(n)w(n)e^{-j\frac{2\pi nk}{N}} =$$

[1] Dies gilt ebenso für die Länge des Analysefensters.

$$= A\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} w(n)e^{-j\frac{2\pi n(k-k_x)}{N}} =$$

$$= AW(k-k_x)$$

wobei $W(k-k_x)$ das um k_x Bins verschobene Spektrum des Fensters repräsentiert. Folglich verhält sich die Fouriertransformation eines gefensterten Signals wie ein Filter, dessen Ausgang die mit der Amplitude des Fensters gewichtete Signalamplitude ist.

$$W(k) = \sum_n w(n)$$

und hängt dementsprechend von der Funktion des Fensters ab. Für den Fall des Rechteckfensters entspricht die Summation exakt der Länge des Fensters, bei allen anderen liefert sie einen kleineren Wert. Die tatsächliche Amplitude bekommt man schließlich durch die Division:

$$A = \frac{X_a(k)}{W(k-k_x)}$$

5.1.3 Implementation in "SOUND ANALYSIS"

Das Programm bietet eine Auswahl von vier verschiedenen Fensterlängen (512, 1024, 2048 und 4096 Abtastpunkte), vier verschiedenen Werten für die Hopsize ($\frac{N}{2}$, $\frac{N}{4}$, $\frac{N}{8}$ und $\frac{N}{16}$) und drei Fenstertypen (Rechteck, Hamming, von Hann)[2].

Die bevorzugten Einstellungen, welche während der Entwicklungsphase hauptsächlich verwendet wurden, sind: von Hann Fenster der Länge 1024 mit einer Hopsize von $\frac{N}{8}$.

[2] Die Amplitudenfaktoren für Hamming und von Hann Fenster betragen $0.54N$ bzw. $0.5N$.

5.2 Abschätzung der Kandidaten für einen Partialton

Wie bereits im vorigen Kapitel erwähnt, verzichtet der vorgeschlagene Algorithmus auf die Möglichkeit, mit Hilfe von zero padding eine Interpolation des Frequenzspektrums, was einer Glättung der Funktion gleichkommt, zu erzielen. Dies erscheint im ersten Moment unlogisch, da durch das Anhängen von Nullen an das Zeitsignal die Lokalisation der Spitzen nicht nur rechnerisch, sondern auch grafisch erleichtert wird. Allerdings würde für ein zufriedenstellendes Maß an Genauigkeit der zero padding Faktor extrem hoch und somit der Rechenaufwand der FFT inakzeptabel groß werden. Deshalb wird die Verbesserung der Frequenzabschätzung in einer separaten Stufe ausgeführt, eine grafische Interpretation bzw. Darstellung des Spektrums ist zudem in dieser Anwendung nicht erforderlich.

Um die Recheneffizienz des FFT Algorithmus ausnützen zu können, muss die Länge der Transformationsblöcke ein Vielfaches der Potenz zur Basis 2 sein, $N = 2^a$, $a = \in \mathbb{N}$.

Der Berechnung der KZFT und der Ermittlung der Kandidaten für einen Partialton wurden die Kapitel 4.2 bis 4.4 gewidmet, deshalb soll hier nicht mehr näher darauf eingegangen werden. Stattdessen werden die Verbesserungsvorschläge für Amplituden- und Frequenzdetektion diskutiert.

5.2.1 Genauere Frequenzabschätzung

Das Frequenzauflösevermögen des Standardmodells ist für die meisten Anwendungen nicht ausreichend und wird deshalb durch entsprechende Modifikationen verfeinert. Im Zusammenhang mit der additiven Synthese sind eine Reihe von Verbesserungsalgorithmen bekannt, zwei davon kommen in "Sound Analysis" zur Anwendung, weitere sind in zahlreichen Publikationen veröffentlicht.

- **Verwendung der Momentanfrequenz**

Die Grundlage dieses Algorithmus bildet die zweidimensionale Darstellung der Daten mit Hilfe des in Kapitel 2.2.2 beschriebenen Spektrogramms, einer gleichzeitigen Interpretation des Signals entlang der Zeit- und Frequenzachse. Ein Spektrogramm erhält man aus der Aneinanderreihung von KZFT-Spektren und wird vor allem für die Analyse von nichtstationären Signalen verwendet. Innerhalb eines KZFT-Spektrums wird das Signal als quasistationär angenommen.

Wie ebenfalls in Kapitel 2.2.2 erläutert, kann durch die Berücksichtigung der Phaseninformation die Limitierung der maximalen Frequenzauflösung, welche aufgrund der vorangegangenen notwendigen Bearbeitung des Signals (Fensterung und Fourier Transformation) auftritt, aufgehoben werden. Somit können die Frequenzpunkte p_m aus Kapitel 4.4 bzw. deren äquivalente Frequenzwerte laut Gleichung (4.2) neu zugeordnet werden:

$$p'_m = \frac{Fs}{2p}\left[\arg(X_{p,m}) - \arg(X_{p,m-1})\right] \tag{5.1}$$

Im Unterschied zur allgemeinen Form in Gleichung (2.1) wird hier die diskrete Schreibweise verwendet, wobei X_m und X_{m-1} die komplexen KZFT des m-ten bzw. (m-1)-ten Frames bezeichnen. Die partielle Ableitung der Phase nach der Zeit wird durch die Subtraktion zweier aufeinanderfolgender Phasenwerte approximiert:

$$\frac{\partial \arg(X)}{\partial t} = \arg(X_m) - \arg(X_{m-1})$$

Die Phasenwerte je Frame werden mit $\arg(X) = \frac{\mathrm{Re}(X)}{\mathrm{Im}(X)} = \arctan(X)$ berechnet und liegen deshalb im Intervall $[-p, p]$ (Hauptwerte der Arkustangens-Funktion). Falls die Differenz der Phasen außerhalb dieses Wertebereichs liegt, muss eine entsprechende Korrektur vorgenommen werden.

Zur Reduktion der Fehler, die aufgrund von Rechenungenauigkeiten in der diskreten Ebene entstehen, wird folgende Überlegung implementiert. Die Kandidaten p_m wurden als lokale Amplitudenmaxima in der Frequenzebene definiert (vgl. Gleichung (4.1)). Das Frequenzspektrum eines Frames setzt sich aus lauter Frequenzantworten der Fensterfunktion, jeweils verschoben um die diskrete

Frequenz p_m, zusammen. Somit charakterisieren die Hauptkeulen des Analysefensters die maximalen Erhebungen im Spektrum und umfassen abhängig von der Fensterfunktion mehrere Bins, deren Phasenwerte annähernd ident sind, da sie zu einem analogen Frequenzanteil des ursprünglichen Signals gehören. Dank dieser Eigenschaft werden nicht nur die Werte p_m reorganisiert, sondern auch deren linke und rechte Nachbarbins $p_{l,m}$ und $p_{r,m}$.

$$p'_{l,m} = \frac{Fs}{2p}\left[\arg(X_{l,p,m}) - \arg(X_{l,p,m-1})\right] \tag{5.2}$$

$$p'_{r,m} = \frac{Fs}{2p}\left[\arg(X_{r,p,m}) - \arg(X_{r,p,m-1})\right] \tag{5.3}$$

Anschließend erfolgt eine lineare Mittelung über die drei Ergebnisse aus Gl. (5.1), (5.2) und (5.3):

$$p''_m = \frac{1}{3}\left(p'_m + p'_{l,m} + p'_{r,m}\right).$$

Dies stellt das endgültige Ergebnis der Verbesserung der Frequenzabschätzung unter Berücksichtigung der Momentanfrequenz dar.

- **Interpolation der Amplitude**

Eine andere Lösung der Frequenzkorrektur bietet die parabolische Interpolation der logarithmischen Amplitudenwerte [Serra]. Die allgemeine Gleichung einer Parabel lautet:

$$y(x) = a\,(x - g)^2 + b \tag{5.4}$$

und beinhaltet die drei unbekannten Parameter a, b und g, wobei diese folgende Bedeutungen haben: a gibt an, ob die Parabel konkav ($a > 0$) oder konvex ($a < 0$) ist bzw. gleichzeitig auch den Grad der Stauchung/Dehnung der Parabel; b kennzeichnet den Offset des Scheitelwertes zur x-Achse; g kennzeichnet den Abstand zum Brennpunkt in horizontaler Richtung. Zur Bestimmung der Unbekannten sind drei Wertepaare (x_1 / y_1), (x_2 / y_2) und (x_3 / y_3) notwendig.

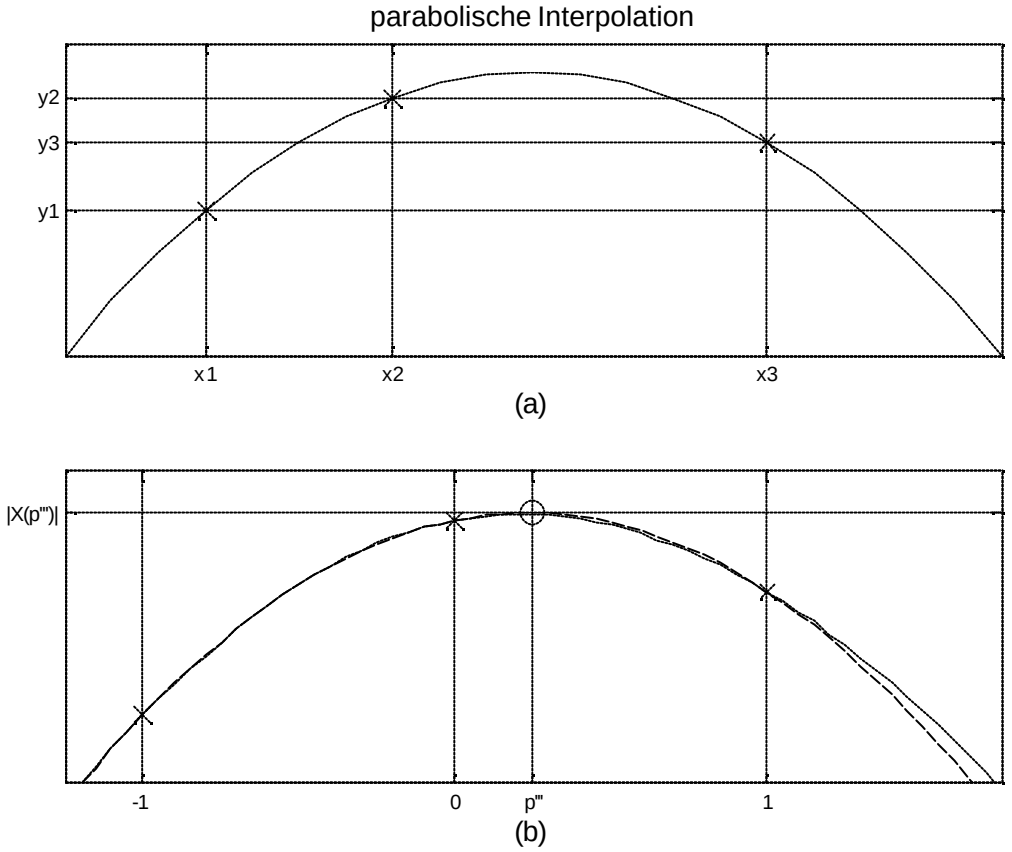


Abb. 5.3: (a) allgemeine Form
(b) Implementation zur Frequenzbestimmung

Diese erhält man aus der diskreten Frequenz des detektierten Maximums p, seinem linken und rechten Nachbarn p_l und p_r, sowie den dazugehörenden logarithmischen Amplituden $X_{mag,p}$, $X_{mag,pl}$ und $X_{mag,pr}$[3]. Als Vereinfachung kann (p_l, p, p_r) durch $(-1,0,1)$ ersetzt werden, wobei dadurch Scheitelwert und Brennpunkt nicht mehr auf der y-Achse liegen, sondern genau um g verschoben sind.

Mit Hilfe der Verschiebung g erhält man somit direkt die verbesserte Lokalisation der Frequenzkomponente:

$$p'''_m = p_m + g_m$$

Das Auflösen von Gleichung (5.4) nach g ergibt:

[3] Der Frameindex m wurde zur übersichtlicheren Schreibweise ausgelassen.

$$g = \frac{1}{2}\frac{X_{mag,pl} - X_{mag,pr}}{X_{mag,pl} - 2X_{mag,p} + X_{mag,pr}}$$

Für die weiterführenden Berechnungen werden die neuen Frequenzwerte einheitlich mit p'_m bezeichnet, das heißt, es ist unwesentlich, welcher der Formalismen zur Verbesserung geführt hat.

5.2.2 Genauere Amplitudenabschätzung

Die Korrektur der Amplitudenwerte kann ebenfalls auf mehrere verschiedene Arten erfolgen. Im Gegensatz zur Frequenz ist das Ohr jedoch Amplitudenverfälschungen gegenüber unempfindlicher [Zwicker, Feldtkeller]. Das Programm "Sound Analysis" impliziert folgende drei Varianten:

- **Explizite Berechnung der Fourier Transformation**

Von jedem neuen Frequenzwert p'_m wird die Amplitude durch die Verwendung der Fourier Transformation in expliziter Schreibweise gebildet, da die MATLAB Funktion "FFT" auf die Bearbeitung diskreter Bins beschränkt und zur Berechnung von Zwischenwerten ungeeignet ist.

$$X'(p') = \left| \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} x(n)w(n)e^{-j\frac{2p\,p'n}{N}} \right|$$

Gleichzeitig könnte in analoger Weise die Phase aktualisiert werden. Praktisch wird dies in "Sound Analysis" jedoch durch die lineare Mittelung der Phasenwerte der drei Hauptbins (p_l, p, p_r) realisiert.

- **Kubische und Spline Interpolation**

Beide Interpolationsarten bedienen sich der MATLAB Funktion "INTERP1"[4].

[4] Mehr Information darüber findet sich in der MATLAB Online-Hilfe: http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml.

- **Parabolische Interpolation**

Unter Verwendung der in Punkt 5.2.1 beschriebenen Formel der parabolischen Interpolation kann ebenfalls eine Verbesserung der Amplitudenabschätzung erfolgen. Die drei Unbekannten a, b und g werden für den allgemeinen Fall mit den Wertepaaren (x_1 / y_1), (x_2 / y_2) und (x_3 / y_3) ermittelt:

$$g = \frac{1}{2}\frac{(y_3 - y_1)(x_2 - x_1)(x_2 + x_1) - (y_2 - y_1)(x_3 - x_1)(x_3 + x_1)}{(y_3 - y_1)(x_2 - x_1) - (y_2 - y_1)(x_3 - x_1)}$$

$$a = \frac{y_2 - y_1}{(x_2 - g)^2 - (x_1 - g)^2}$$

$$b = y_2 - a(x_2 - g)^2$$

Daraus ergibt sich für die neue Amplitude $X''(p')$:

$$X''(p') = a(p' - g)^2 + b$$

5.2.3 Zusammenfassung

Die unterschiedlichen Verbesserungsvorschläge liefern sowohl bei der Frequenz- als auch bei der Amplitudenabschätzung ähnliche Werte. Damit hat ihre Auswahl keinen gravierenden Einfluss auf die Performance des gesamten Analyseapparates, sondern wurden vielmehr zu Testzwecken im Programm inkludiert.

Eine interessante Alternative wird allerdings in [Rodet] und detaillierter in [Depalle, Tromp] und [Depalle, Hélie] entwickelt und als parametrische Modellierung der KZFT bezeichnet. Es wird ein iterativer Algorithmus mit dem Ziel der Minimierung des mittleren quadratischen Fehlers von Originalsignal und einem entsprechenden Modell vorgeschlagen.

$$|E|^2 = |S_{orig} - S_{modell}|^2 \rightarrow min.$$

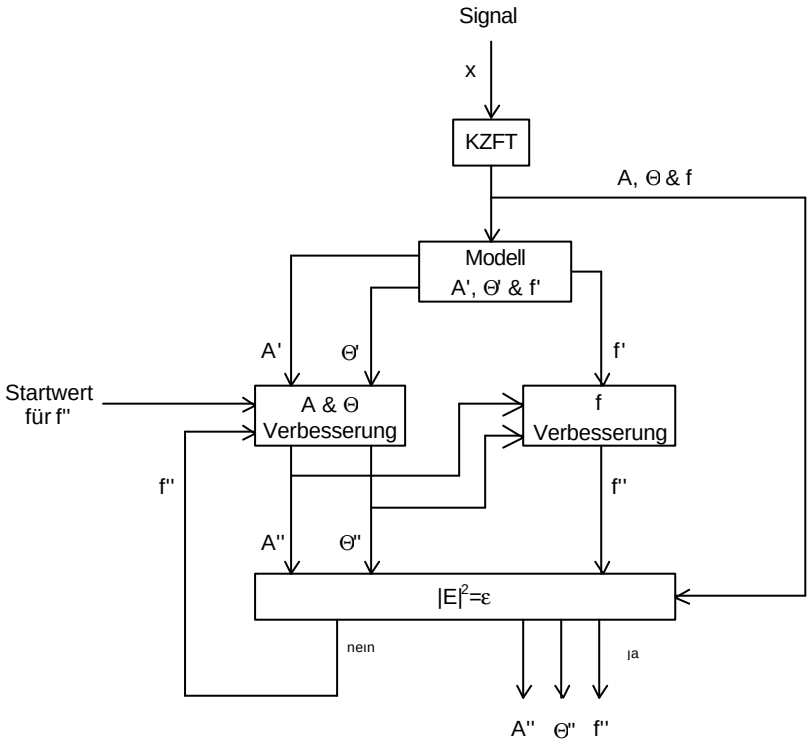


Abb. 5.4: Prinzip der parametrischen Modellierung

$$S_{modell}(F) = \sum_{k=1}^{L} \left\{ \frac{a_k}{2} e^{j\Phi_k} W(F - f_k) + \frac{a_k}{2} e^{-j\Phi_k} W(F + f_k) \right\}$$

wobei mit a_k, Φ_k und f_k die Teiltöne des Signals bezeichnet sind und $W(\cdot)$ die fouriertransformierte Fensterfunktion bedeutet. Weiters steht F hier für die diskrete Frequenzvariable.

Der Berechnungsfehler $|E|^2$ inkludiert nicht nur den nicht-deterministischen Anteil des Signals, sondern ebenso etwaige Teiltöne, welche vom Modell nicht detektiert wurden. Das Modell besteht wiederum aus der Summe von Sinusschwingungen. Die Annäherung der Bedingung erfolgt in zwei Schritten. Zunächst wird die genaue Frequenz als bekannt vorausgesetzt (Startfrequenz) und damit eine Verbesserung der Amplitude und Phase berechnet. Anschließend wird mit diesen Werten die Frequenz korrigiert, welche als neue Eingangsvariable für den Amplitude-Phase-Algorithmus dient. Die Anzahl der Iterationen hängt letztendlich von der gewünschten Genauigkeit der Approximation ab. Die Vorgehensweise bei der Berechnung ist in Abb. 5.4 schematisch dargestellt.

Da die Komplexität der verwendeten Implementationen beträchtlich geringer ist und trotzdem ein zufriedenstellendes Ergebnis erreicht wird, wurde der iterative Ansatz nicht weiter verfolgt.

5.3 Zeitliche Struktur der Analysedaten

Die zeitliche Struktur der Analysedaten ist grundsätzlich durch die Wahl der DFT Länge und der Hopsize vorgegeben. Wie in Kapitel 2.2.2 gezeigt, kann diese durch die Betrachtung des Energieschwerpunktes eines Signalanteils präzisiert werden. Für die Frequenzachse ist dies bereits in Kapitel 5.2.1 geschehen, deshalb liegt es nahe, selbiges auch für die Zeitachse zu realisieren.

5.3.1 Verwendung der Gruppenlaufzeit

Im Gegensatz zur Frequenzkorrektur ist hier nicht der exakte Beginn jedes einzelnen Partialtons erforderlich, man implementiert die Zeitkorrektur nach folgender Idee: Wird eine Frequenzkomponente in der ersten Hälfte des Analysefensters detektiert, so wird sie diesem Fenster zugeordnet. Befindet sie sich jedoch in der zweiten Hälfte, so wird ihre Amplitude der entsprechenden Komponente im nächsten Fenster aufsummiert und im aktuellen Fenster auf Null gesetzt.

$$X_{p,m+1} = X_{p,m+1} + X_{p,m}$$

und

$$X_{p,m} = 0$$

Die Bestimmung des Startzeitpunktes erfolgt mit Hilfe der Gruppenlaufzeit je Teiltonkandidat und Frame.

Die Schwierigkeit der Zuweisung der Amplitude dem nachfolgenden Frame liegt darin, dass zuerst ein gültiger Kandidat dafür definiert werden muss. Infolgedessen kann die Reorganisation nicht an dieser Stelle im Programm durchgeführt werden, sondern benötigt vorher das Gruppieren der einzelnen Frequenzkomponenten, das im nächsten Abschnitt dargestellt ist. Die Information

über eine eventuelle zeitliche Verschiebung wird trotzdem hier gewonnen und für die spätere Auswertung gespeichert.

5.3.2 Bildung von Spuren

Bis zu diesem Zeitpunkt wurden die einzelnen Frames getrennt voneinander auf das Vorhandensein signifikanter Frequenzanteile untersucht. Um eine Verbindung zwischen diesen losen, ungeordneten Werten herzustellen, werden sie unter Berücksichtigung ihrer Frequenz zu Spuren zusammengeführt. Obwohl die Verarbeitung der Amplituden- und/oder Phaseninformation ebenfalls dazu verwendet werden kann, ist die Beschränkung auf das Frequenzkriterium für den vorgeschlagenen Algorithmus ausreichend genau.

Ziel dieser Strukturierung ist die Generierung einer kontinuierlichen zeitlichen Entwicklung der Analysewerte. In weiterer Folge wird diese Eigenschaft, die das Originalsignal ebenfalls aufweist, auf das synthetisierte Signal übertragen[5]. Folglich wird der wahrnehmbare Unterschied zwischen dem Originalklang und der künstlichen Nachbildung verringert, wodurch letzteres ein natürlicheres Hörempfinden vermittelt.

Zur Beschreibung der Wirkungsweise des Prozesses wird ein beliebiges Frame m mit seinen detektierten Kandidaten $(r_1, r_2, ..., r_R)$ sowie das vorangegangene Frame $m-1$ mit den Kandidaten $(p_1, p_2, ..., p_P)$ betrachtet. Im allgemeinen stimmt die Anzahl der Werte p_i nicht mit jener der Werte r_j überein, $P \neq R$. Von jedem p_i wird zunächst ein zulässiger Toleranzbereich Δf_{pi} definiert, wobei dieser vom Frequenzwert von p_i abhängig ist:

$$\Delta f_{pi} = A f_{pi}$$

Der Parameter A ist grundsätzlich frei wählbar, wird jedoch in "Sound Analysis" auf Werte zwischen 1 und 2 beschränkt. Dies bedeutet, dass die maximale Abweichung dem Frequenzintervall einer Oktav (Frequenzverhältnis = 2:1) entspricht. Im nächsten Schritt wird für jeden einzelnen Wert r_j geprüft, ob sich

[5] Dies gilt generell für den deterministischen Signalanteil. Sehr rasche Änderungen im Signal werden hier nicht detektiert und bearbeitet, sondern im stochastischen Restgefüge berücksichtigt.

dieser innerhalb eines Toleranzbereiches Δf_{pi} befindet, wobei vier verschiedene Fälle auftreten können:

Fall 1: r_j liegt genau innerhalb eines Δf_{pi},

Fall 2: r_j liegt innerhalb mehrerer Δf_{pi},

Fall 3: mehrere r_j liegen innerhalb eines Δf_{pi} oder

Fall 4: r_j liegt außerhalb aller Δf_{pi}.

Dementsprechend erfolgt die Zuteilung zu einem p_i bzw. die Entstehung und Fortführung einer Spur. Zur Veranschaulichung dienen Abb. 5.5(a)-(d). Das aktuelle Frame wird jeweils mit m bezeichnet, dessen Vorgänger mit $m-1$. Die Toleranzgrenzen der Frequenzabweichung von einem Frame zum nächsten werden *fu* für das untere bzw. *fo* für das obere Limit genannt.

In Fall 1 wird r_j einfach mit jenem p_i verbunden, dessen einziger Treffer es ist und somit eine bereits bestehende Spur fortsetzt. Die Kennzeichnung einer Spur erfolgt durch einen Spurindex, mit dem der neu hinzukommende Wert versehen wird. Wenn ein r_j für mehrere Spuren geeignet wäre (Fall 2), so wird nach dem Minimum von $\left| f_{pi} - f_{rj} \right|$ gesucht und der passende Index notiert. Gleiches gilt für die umgekehrte Situation, bei dem auf eine Spur mehrere Werte r_j zutreffen würden (Fall 3). In Abb. 5.5(b) und (c) ist die gültige Verbindung mit einer durchgezogenen Linie gekennzeichnet, die entsprechenden Punkte sind mit Kreuzen markiert. Jene Treffen, die aufgrund des Minimum-Kriteriums ausscheiden, sind strichliert gezeichnet bzw. mit Kreisen versehen. Es kann auch vorkommen, dass in Frame $m+1$ eine Frequenzkomponente enthalten ist, die in den bisherigen Frames noch nicht präsent war (Fall 4). Somit wird eine neue Spur gebildet, welche in Frame m mit der Amplitude

$$X_{p\,i,m-1} = 0$$

und der Phase

$$j_{\,pi,m-1} = j_{\,rj,m} - 2p\ f_{rj,m}$$

beginnt. Dies ist in Abb. 5.5 (d) durch die strichpunktierte Verbindung und dem "Karo-Punkt" dargestellt.

Nachdem mit dieser Methode alle Kandidaten des aktuellen Frames verarbeitet worden sind, muss noch überprüft werden, ob eine bereits existierende Spur keinen Treffer gefunden hat. In diesem Fall wird sie in Frame m beendet mit

$$X_{r\,j,m} = 0$$

und

$$j_{rj,m+1} = j_{p\,i,m} + 2p\,f_{pi,m}$$

Bei beiden Phasenberechnungen muss darauf geachtet werden, das Ergebnis mittels Modulo-Funktion in den Wertebereich $[0,2p]$ zu bringen, um einen stetigen Phasenverlauf zu gewährleisten.

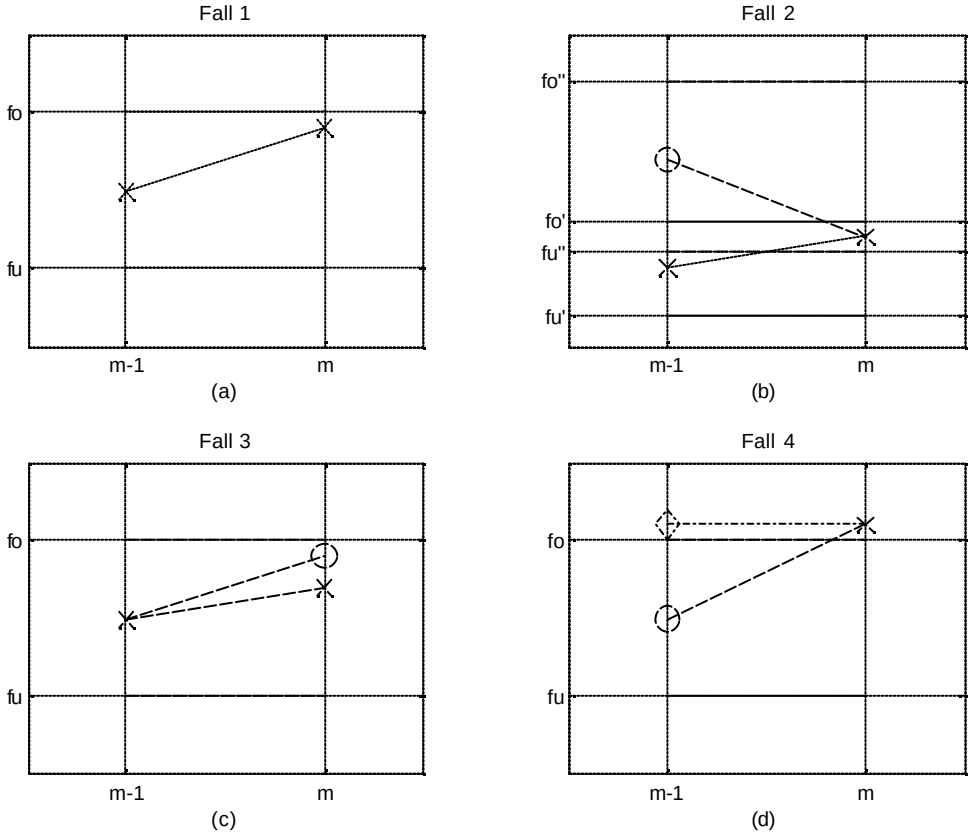


Abb. 5.5: Möglichkeiten der Zuordnung
durchgezogen: gültige Verbindungen
strichliert: ungültige Verbindungen

5.3.3 Zusammenfassung

Eine vereinfachte Darstellung der drei möglichen Zustände einer Spur – Entstehung, Fortsetzung, Beendigung – ist in Abb. 4.3 grafisch dargestellt. Der tatsächlich implementierte Algorithmus ist in Abb. 5.7 als Flussdiagramm abgebildet. Die Amplitude einer neuen Spur startet immer mit dem Wert Null im vorigen Frame, die Phase wird vom aktuellen Wert rückgerechnet und der Frequenzwert wird beibehalten. Für das Auslaufen einer Spur gelten äquivalente Bestimmungen.

Im Anschluss kann schließlich die zeitliche Umstrukturierung durchgeführt werden, wobei sich diese lediglich auf den Beginn von neuen Spuren bezieht. Bereits bestehende Spuren sind davon ausgeschlossen.

5.4 Weitere Kriterien für Spuren

Das vorangegangene Kapitel beschreibt die Ordnung und Gruppierung der losen Analysedaten aufgrund eines einfachen Frequenzkriteriums. In den meisten praktischen Anwendungen reicht diese Forderung allein für ein akzeptables Syntheseresultat aus (siehe MQ Algorithmus [McAulay, Quatieri] bzw. PARSHL [Smith, Serra]). Trotzdem bringt folgende Zusatzbedingung einen weiteren Vorteil mit sich. Durch das Definieren einer minimalen Lebensdauer bzw. einer maximalen Ruhedauer jeder Spur wird die Kontinuität des Gesamtspektrums zudem verfeinert.

- Minimale Lebensdauer: Sie gibt an, in wie vielen aufeinanderfolgenden Frames eine Spur präsent sein muss, um für die Synthese weiterverwendet zu werden. Aus Abb. 5.6 ist ersichtlich, dass zu kurze Spuren nicht weiterverwendet werden – im gezeigten Fall sind dies Spur 2 und Spur 4.
- Maximale Ruhezeit *RZ*: Sie bestimmt die Anzahl der aufeinanderfolgenden Frames, in denen die Spur nicht vorhanden sein

muss. Während dieser Zeit wird die Spur durch Interpolation der Amplitude, Phase und Frequenz fortgesetzt. Die Idee, welche der Einführung dieser Größe zugrunde liegt, hängt mit der Ermittlung der Teiltonkandidaten zusammen. Dabei wird das Kriterium der minimalen globalen Amplitudenschwelle S verwendet (siehe Kapitel 4.3), welches als variabler vom Benutzer einstellbarer Parameter im Programm integriert ist. Bei ungünstiger Wahl dieses Grenzwertes kann es vorkommen, dass die (nichtstationäre) Amplitude eines Teiltons in aufeinanderfolgenden Frames alternierend oberhalb bzw. unterhalb dieser Schwelle liegt, vgl. Abb. 5.8a. Die Folge wäre eine fehlerhafte, unvollständige Detektion dieses Teiltons (siehe Abb. 5.8b), die sich negativ auf die Effizienz der Analysestufe auswirken würde. In den ersten beiden Fällen ($d1$ und $d2$) ist die Anzahl der Frames, in denen die Amplitude des Teiltons unterhalb der Schwelle S liegt, kleiner als die Ruhezeit, also erfolgt eine Interpolation und Weiterführung der Spur. Die Anzahl der "Ruheframes" $d3$ ist allerdings größer als RZ, darum wird Spur x beendet und eine neue Spur y an der entsprechenden Stelle erzeugt, Abb. 5.8(c).

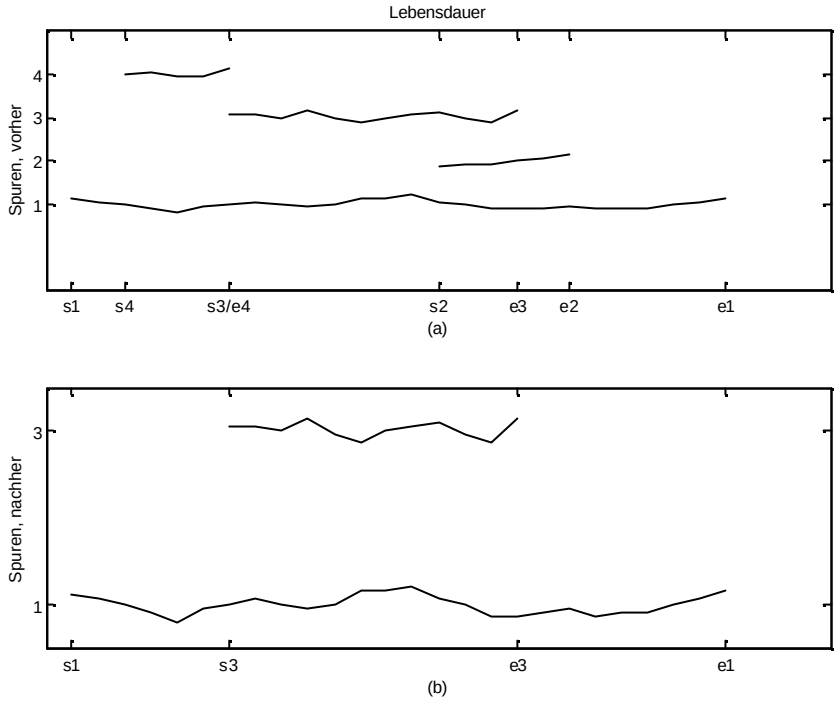


Abb. 5.6: Spurverlauf mit Berücksichtigung der Lebensdauer

Die Implementation beider Parameter in "Sound Analysis" wurde als frei wählbare Werte realisiert.

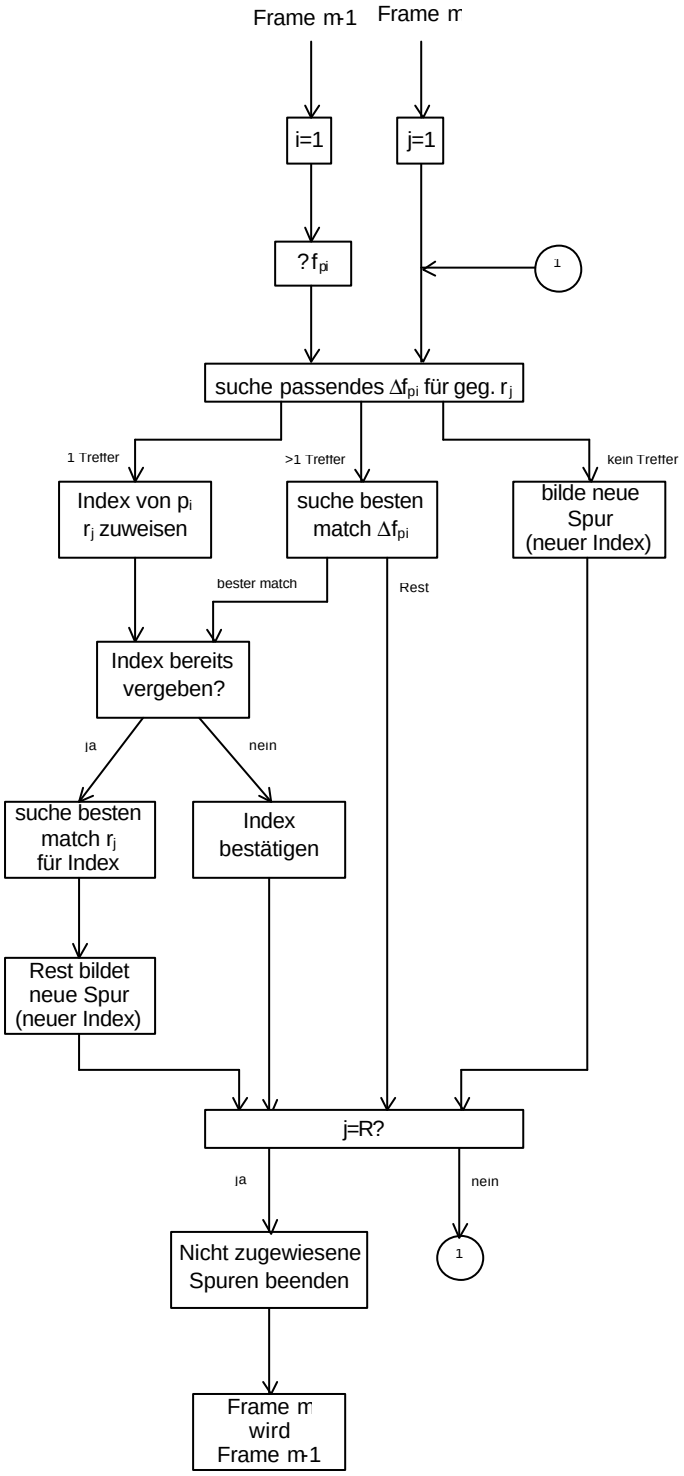


Abb. 5.7: Flussdiagramm des Programmteils "Bildung von Spuren"

All jene Spuren, die keiner der beiden obigen Definitionen genügen, werden als ungültig gekennzeichnet und scheiden für die Speicherung und Weitergabe an das Syntheseprogramm aus. Somit reduziert sich außerdem die Anzahl der Sinusoszillatoren in der Additiven Klangsynthese, das wiederum verringert den erforderlichen Rechenaufwand.

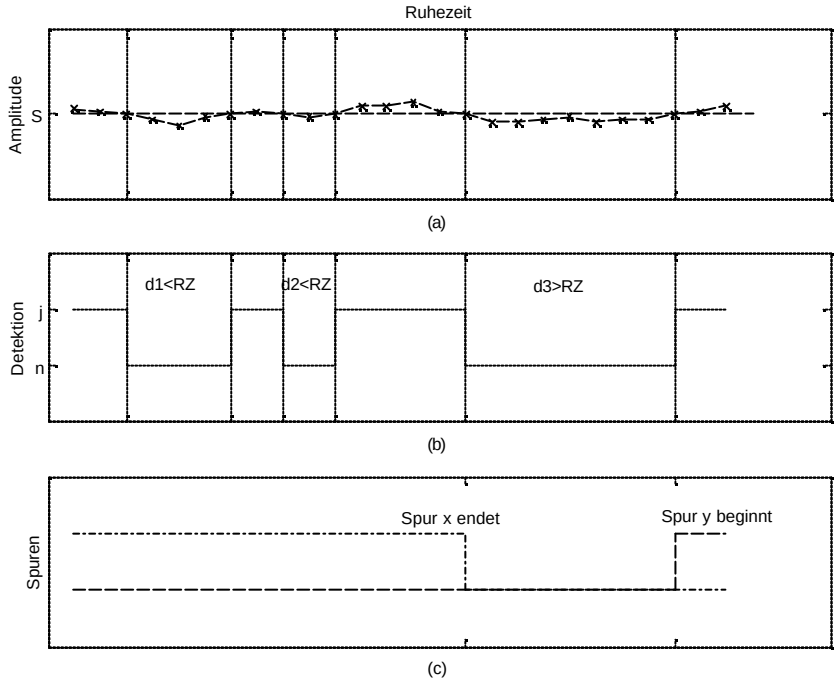


Abb. 5.8: Spurverlauf mit Berücksichtigung der Ruhezeit

5.5 Der stochastische Teil

Sämtliche Verarbeitungsschritte, welche bisher in diesem Abschnitt besprochen wurden, dienen der Detektion der deterministischen Komponenten des Originalsignals. Unter Berücksichtigung einiger Einschränkungen für die zu analysierenden Signale wie auch für die Flexibilität der Analyse selbst, liefert dieser Formalismus ein zufriedenstellendes Ergebnis der Signalbeschreibung. Das Eingangssignal sollte quasistationär sein mit einem hohen Signal-Rausch-Abstand. Stark verrauschte Signale und Transiente sind für eine Synthese mittels AKS

ungeeignet. Da aber in jedem natürlichen Audiosignal Transiente vorkommen, werden diese als Klangsamples gespeichert und nicht der Analyse zugeführt. In der Synthese werden sie unverändert in die Analysedaten integriert, wobei bei der Zusammensetzung auf einen kontinuierlichen Signalverlauf geachtet werden muss, um hörbare Ungenauigkeiten zu vermeiden. Ein Nachteil ergibt sich dann, wenn zwischen Analyse und Synthese zusätzlich eine Modifikation der Parameter erwünscht wird, da die Änderung der Klangsamples problematisch und sehr unflexibel ist – es steht kein Modell für die Transienten zur Verfügung, dessen Parameter variiert werden können [Verma, Meng]. Aus diesem Grund werden im vorgeschlagenen Programm in einer separaten Stufe jene Daten analysiert, die nicht durch das deterministische Modell beschrieben worden sind.

5.5.1 Subtraktion im Zeitbereich

Die Gewinnung des Restsignals kann auf unterschiedliche Art und Weise erfolgen, die einfachste ergibt sich bei Betrachtung des ursprünglich aufgestellten Modells aus Gleichung 3.2. Man erkennt sofort, dass eine Subtraktion im Zeitbereich das gewünschte Resultat liefert:

$$s_s(n) = s(n) - s_d(n)$$

Dabei bezeichnet $s(n)$ das Original, $s_d(n)$ den resynthetisierten deterministischen und $s_s(n)$ den stochastischen Anteil. Voraussetzung für die Gültigkeit dieser Formel ist allerdings die samplegenaue Übereinstimmung der beiden Funktionen $s(n)$ und $s_d(n)$, andernfalls führt die Subtraktion zu einem fehlerhaften Ergebnis.

Bei der Entwicklung des Programms "Sound Analysis" ergab sich exakt dieses Problem – die Synthesestufe war bereits vorhanden, vollständig unabhängig von der Analysestufe und somit auch nicht 100%ig kompatibel mit ihr. Auf die Implementation einer eigenen Syntheseeinheit mittels AKS wurde aus zeitlichen Gründen verzichtet. Neben der AKS bietet die sogenannte Overlap-Add-Synthese eine alternative Möglichkeit der Zusammensetzung der Analysedaten [Smith, Serra]. Dabei wird von jedem KZFT Spektrum die inverse Fourier Transformation gebildet:

$$x_m(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} X_m(k) e^{j\frac{2\pi nk}{N}}$$

Anschließend werden die einzelnen gefensterten Zeitframes $x_m(n)$ entsprechend der Hopsize hs überlappt und addiert:

$$x(n) = \sum_{m=0}^{M-1} x_m\,(n - m \cdot hs)$$

Die Rekonstruktion eines Signals mittels Overlap-Add-Algorithmus ist zwar recheneffizienter als die Verwendung der AKS, ihr Nachteil liegt darin, dass das gesamte KZFT Spektrum zur Verfügung stehen muss, es darf also keine Datenreduktion erfolgen. Das vorgeschlagene Analysekonzept verletzt allerdings diese Vorschrift, indem es eine Reihe von Restriktionen inkludiert, welche die gültigen Spuren bzw. Partialtöne einhalten müssen. Demzufolge ist die zeitliche Subtraktion nicht zielführend und es muss nach einer weiteren Maßnahme zur Ermittlung der verbleibenden Daten gesucht werden.

5.5.2 Subtraktion im Frequenzbereich

Die spektrale Subtraktion gestaltet sich etwas schwieriger als jene im Zeitbereich. Dies lässt sich damit erklären, dass durch die Fensterung und Transformation in die Frequenzebene die Ermittlung der deterministischen Frequenzbins lediglich eine Abschätzung bedeutet. Außerdem bestimmt nicht nur der markierte maximale Bin, welcher in der vorliegenden Arbeit mit p oder p_m bezeichnet wird, eine im Signal vorkommende Frequenz, sondern es tragen noch diverse Nebenbins dazu bei, deren Anzahl von der nichtstationären Eigenschaft des Signals abhängig ist. Abb. 5.9 verdeutlicht die Aufweitung der Hauptkeule aufgrund zeitlichen Veränderung der Frequenz. Andererseits ist das überlagerte Rauschen ein Signal breitbandiger Natur und beeinflusst infolgedessen ebenso die Subtraktion.

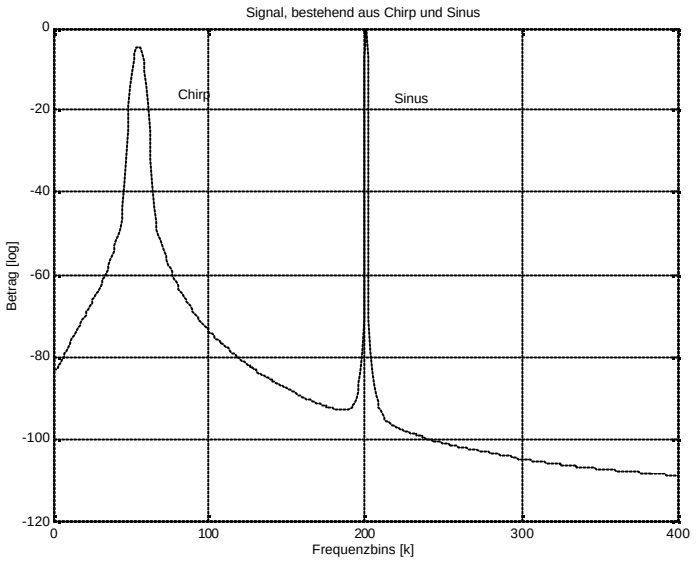


Abb. 5.9: Aufweitung der Hauptkeule durch nichtstationäre Frequenz

Trotz der problematischen Voraussetzungen stellt diese Methode eine brauchbare Lösung zur Gewinnung der stochastischen Daten dar und die wesentlichen Charakteristika können hinreichend genau modelliert werden. Im Gegensatz zur Beschreibung des deterministischen Anteils genügt hier eine grobe Abschätzung der Amplitudeneinhüllenden des Spektrums, eine detaillierte Frequenzanalyse sowie die Ermittlung der Phase sind nicht notwendig. Die Subtraktion beinhaltet folgende Schritte:

Ermittlung aller Bins und deren Amplituden zwischen links- und rechtsseitigem Minimum eines Kandidaten $\rightarrow k_d, X_d(k_d)$,

Subtraktion dieser vom Gesamtspektrum $X_{mag}(k)$. Die Subtraktion im linearen Amplitudenbereich liefert für die gekennzeichneten Bins den Wert 0. Es wird jedoch in dieser Stelle mit logarithmischen Amplituden gerechnet. Somit müssten die ermittelten Bins theoretisch auf $-\infty dB$ gesetzt werden, praktisch sind $-200dB$ ausreichend $\rightarrow X_s(k)$,

Lineare Interpolation der Einhüllenden an diesen Stellen $\rightarrow X'_s(k)$,

Rückrechnung auf eine lineare Amplitude $\rightarrow X_{slin}(k)$,

Generierung einer willkürlichen Phase mittels Zufallsgenerator $\rightarrow f_s(k)$,

Berechnung der inversen Fourier Transformation $\rightarrow x_s(n)$,

Multiplikation mit einem Sinusfenster $\rightarrow x_{ws}(n)$[6],

Überlappung und Addition der einzelnen gefensterten Frames $\rightarrow s_s(n)$.

Die Fensterung der Zeitframes ist erforderlich, damit beim Zusammenfügen der Einzelframes wiederum ein kontinuierlicher Verlauf der Funktion entsteht.

Das somit gewonnene Signal $s_s(n)$ ist vom deterministischen Anteil völlig unabhängig, kann jedoch vor der Overlap-Add-Synthese (letzter Schritt) ebenfalls modifiziert werden.

5.6 Ergebnis

Wie schon erwähnt, stand bei der Entwicklung des Algorithmus ein eigenständiges Programm zur Additiven Klansynthese zur Verfügung, weshalb die Ausgabeparameter zweierlei unterschiedliche Gestalt aufweisen: Zum einen liegen die Analysedaten des deterministischen Signalanteils in vektorieller Form vor, sie sind bereits exakt für dieses Programm aufbereitet und können in einem eigenen File mit der Extension ".add" abgespeichert werden. Für den stochastischen Anteil ist zusätzlich die Synthese vorhanden, es gibt allerdings keine Möglichkeit, zuvor eine Modifikation der Daten vorzunehmen. Die zeitliche Funktion wird in ein ".wav" File umgewandelt und gespeichert. Nach Anwendung der AKS auf das ".add" File erhält man ebenfalls ein ".wav" File, welches, zu vorigem hinzuaddiert, eine Rekonstruktion des Originals ergibt.

Um die prinzipielle Wirkungsweise des Analyseapparates direkt testen zu können, wurde eine vereinfachte Version der AKS, welche ebenso bereitstand und zum Einbau besser geeignet war, in das Programm integriert. Damit kann ad hoc ein Vergleich zwischen Original und Nachbildung sowohl akustisch als auch grafisch angestellt werden.

[6] Genaugenommen müssten alle Schritte bis hierher noch mit dem Fensterindex m versehen werden.

5.7 Bedienungsoberfläche von SOUND ANALYSIS

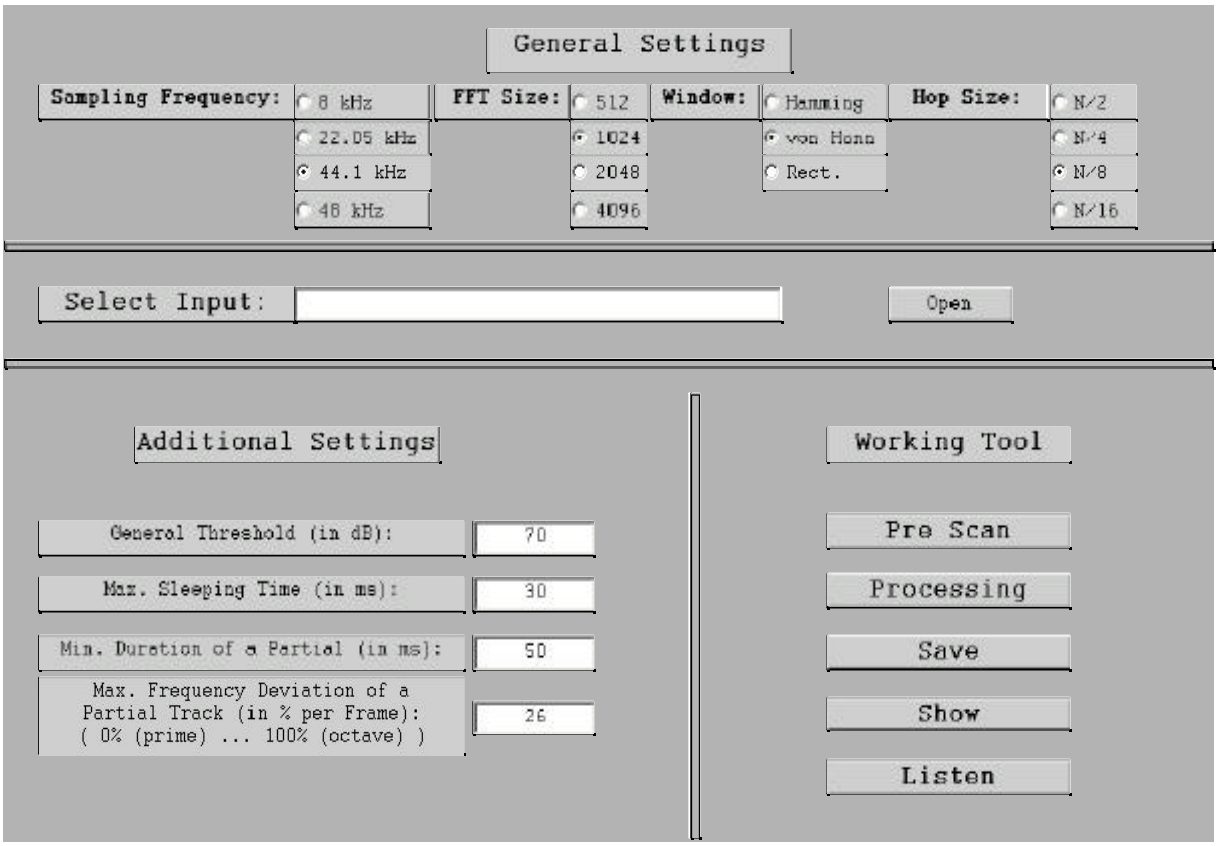


Abb. 5.10: grafische Oberfläche

Die Bedienungsoberfläche des Programms ist in Abb. 5.10 dargestellt und gliedert sich allgemein in vier Bereiche: generelle Einstellungen, Wahl des Eingangssignals, erweiterte Einstellungen und dem Arbeitsbereich.

- **Generelle Einstellungen (General Settings):**

Die für den Analysealgorithmus wesentlichen Parameter Samplingfrequenz, Länge der KZFT, Fensterfunktion und Hopsize können hier ausgewählt werden. Da das zu analysierende Signal im Normalfall bereits in digitaler Form vorliegt, ist die Wahl der Samplingfrequenz mit jener des Eingangssignals gleichzusetzen. Andernfalls bringt das Programm eine Fehlermeldung und liefert in weiter Folge falsche Analysewerte. Die restlichen drei Parameter sollten von den Eigenschaften des Signals abhängig verwendet werden, wobei sich speziell die Länge der KZFT und die Hopsize auf den Rechenaufwand auswirken. Je geringer der Anteil der Transienten in einem Signal ist bzw. je weniger Wert auf die zeitliche Struktur

gelegt wird, desto größer können "FFT size" und "hop size" gewählt werden und desto schneller sind die Berechnungen.

- **Wahl des Eingangssignals (Select Input):**

Das zu analysierende Eingangssignal muss in digitaler Form als Wave-Datei vorhanden sein. Es kann entweder direkt eingegeben (vollständigen Pfad angeben) oder mit Hilfe des "Open"-Buttons aus der Verzeichnisstruktur ausgewählt werden.

- **Erweiterte Einstellungen (Additional Settings):**

Diese frei wählbaren Parameter haben einen gravierenden Einfluss auf die Performance des Programms. In zahlreichen Testversuchen wurden die vorgegebenen Startwerte ermittelt, welche sich bei einer Vielzahl von Signalen zu einem zufriedenstellenden Ergebnis geführt haben. Es empfiehlt sich jedoch in der Praxis, diese zu variieren und auf jedes einzelne Signal individuell einzustellen. Generell sind ausreichende Vorkenntnisse über den Signaltyp bei der Auswahl von großem Vorteil.

Der erste Parameter – "General Threshold" – entspricht exakt der in Kapitel 4.3 definierten globalen Schwelle L_2 (Kriterium K2). Die weiteren zwei Variablen – "Max. Sleeping Time" und "Min. Duration of a Partial" – dienen der Verbesserung des Algorithmus' und sind in Kapitel 5.4 erklärt. Der letzte Einstellwert – "Max. Frequency Deviation of a Partial Track" – stellt das Frequenzkriterium bei der Bildung der Spuren dar, siehe Kapitel 5.3.2. Die Angabe in Prozent je Frame gibt die maximal zulässige Abweichung von einem Abtastzeitpunkt zum nächsten an.

- **Arbeitsbereich (Working Tool):**

Die letzten fünf Bedienungselemente sind als sogenannte "push buttons" ausgeführt dienen der Bearbeitung des Eingangssignals. Mit Hilfe der "Pre Scan" Funktion kann der zeitliche Verlauf der im Signal vorkommenden Frequenzanteile in einem eigenen Fenster dargestellt werden. Zu bestimmten Zeitpunkten wird eine FFT der Länge der vorher eingestellten "FFT Size" durchgeführt und die Frequenzantwort in dB aufgezeichnet. Die Zeitpunkte werden in der Grafik mit "Time Instant" bezeichnet und sind in Sekunden angegeben. Dadurch kann die Einstellung der "General Threshold" erleichtert werden. Die Funktion "Processing"

führt die eigentliche Berechnung der Parametertrajektorien für Amplitude, Frequenz und Phase aus. Das Programm ermöglicht die Abspeicherung der Analysedaten einerseits als Wave-Datei, andererseits als Datei mit der Extension .add. Die Struktur dieser Files ist exakt auf das bereits bestehende Programm der AKS abgestimmt. Des weiteren können der zeitliche Verlauf von Original- und synthetischem Signal dargestellt werden, sowie beide hintereinander angehört werden. Für die beiden letzten Optionen und die Speicherung als Wave-Datei wurde direkt in SOUND ANALYSIS eine vereinfachte Version der AKS implementiert, welche zur schnellen Auswertung der Analysewerte dient. Es wird darauf hingewiesen, dass die hiermit gewonnen Resultate lediglich als Unterstützung im Umgang mit den erweiterten Einstellungen gedacht sind und von den Ergebnissen, welche mittels Synthetisierung der .add Files erzielt werden, abweichen können.

6 Testbeispiele

In diesem Kapitel wird anhand von Testbeispielen die Performance des vorgeschlagenen Analyseprogramms diskutiert. Die Auswahl der Audiosignale erfolgte wahllos und ohne bestimmte Restriktionen. Alle besitzen dasselbe Aufnahmeformat: Monospur mit 16bit Auflösung und 44.1kHz Samplingfrequenz.

Als erstes Beispiel wird der Ton einer Oboe analysiert, welcher ein relativ einfaches harmonisches Spektrum besitzt. Die wählbaren Parameter waren bis auf die "General Threshold" auf die voreingestellten Werte gesetzt. Da dieser Schwellwert von der Lautstärke des Signals bzw. dem Signal-Rausch-Abstand abhängig ist, wurde er mit 80dB festgelegt.

Die Analyseergebnisse sind – jeweils zusammen mit dem Original – einmal als Zeitfunktion, Abb. 6.1 bzw. als Frequenzspektrum, Abb. 6.2, dargestellt. Auch die Kombination zweier Töne liefert ein zufriedenstellendes Resultat; der Beginn des zweiten Tones ist in Abb. 6.3 sehr gut zu erkennen. Bei der folgenden Variante wird die Oboe von Streichen begleitet. Dadurch entsteht ein dichteres Frequenzspektrum, siehe Abb. 6.4. Um Auflösung etwas zu verfeinern, wurde hier die Länge der KZFT-Blöcke auf 2048 erhöht.

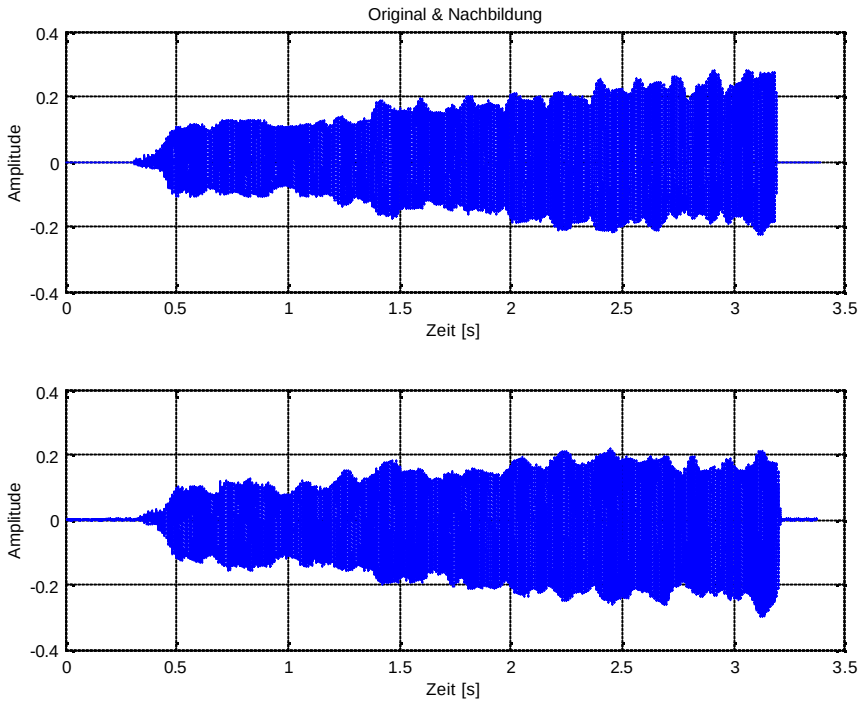


Abb. 6.1: zeitlicher Verlauf des Tones

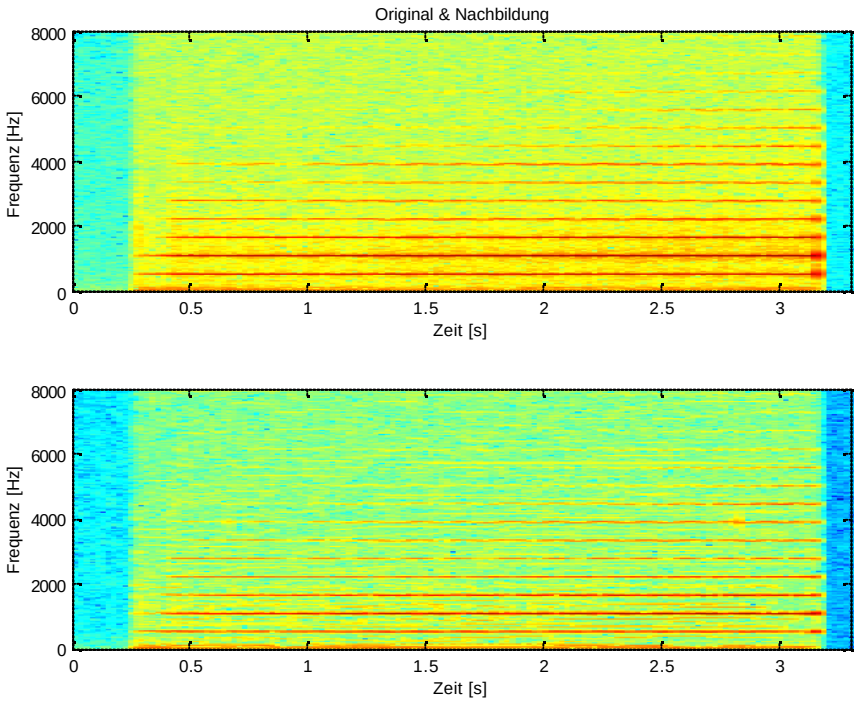


Abb. 6.2: Spektrum einer Oboe

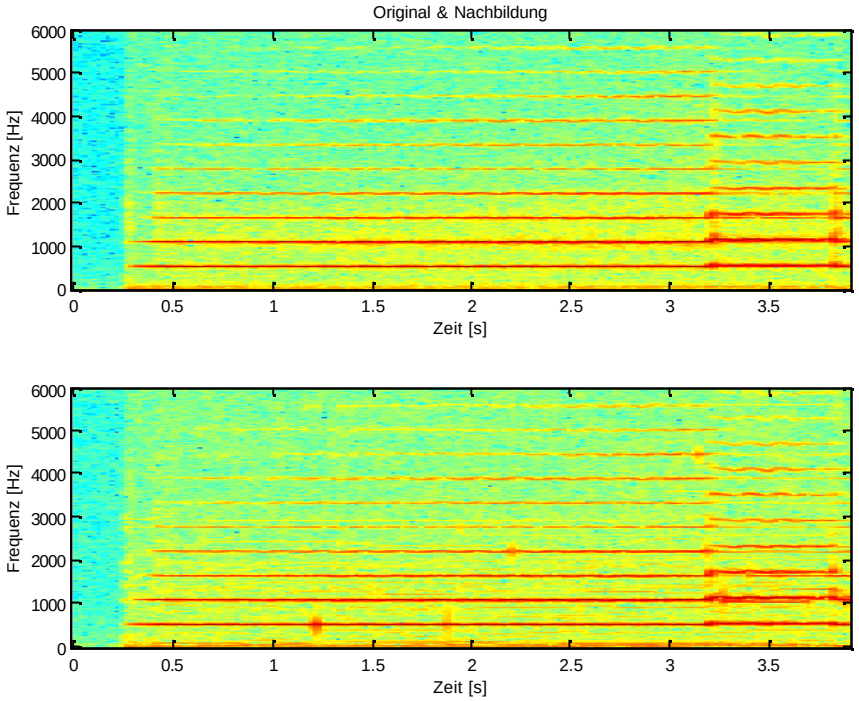


Abb. 6.3: zwei aufeinanderfolgende Töne einer Oboe

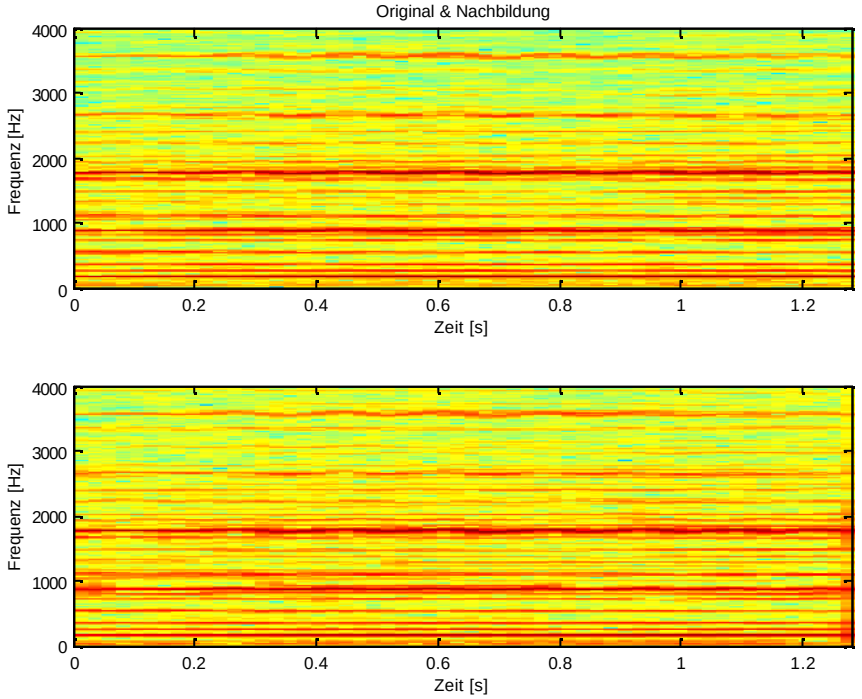


Abb. 6.4: Oboe und Streicher

Im nächsten Beispiel wird ein Klavierton untersucht, der mit einem tieffrequenten Störgeräusch überlagert ist. Zunächst wurde die Analyse wieder mit folgenden Einstellungen durchgeführt:

- Sampling Frequency: 44.1kHz
- FFT Size: 1024
- Window: von Hann
- Hop Size: N/8

In Abb. 6.5 sind die zeitlichen Verläufe von Original und Nachbildung zu sehen, die jedoch wenig Information über die Genauigkeit der Analyse liefern. Mehr Aufschluss über das Resultat wird durch den Vergleich der beiden Spektrogramme gewonnen, die in Abb. 6.6 dargestellt sind. Das maximale Auflösevermögen im Bereich unter 200Hz ist deutlich zu erkennen ist. Die exakte Auflösung berechnet sich aus obigen Werten sowie der 4 Bin breiten Hauptkeule des von-Hann-Fensters mit:

$$\Delta f = \frac{Fs}{N} b_{HK} = \frac{44100}{1024} 4 = 172 Hz$$

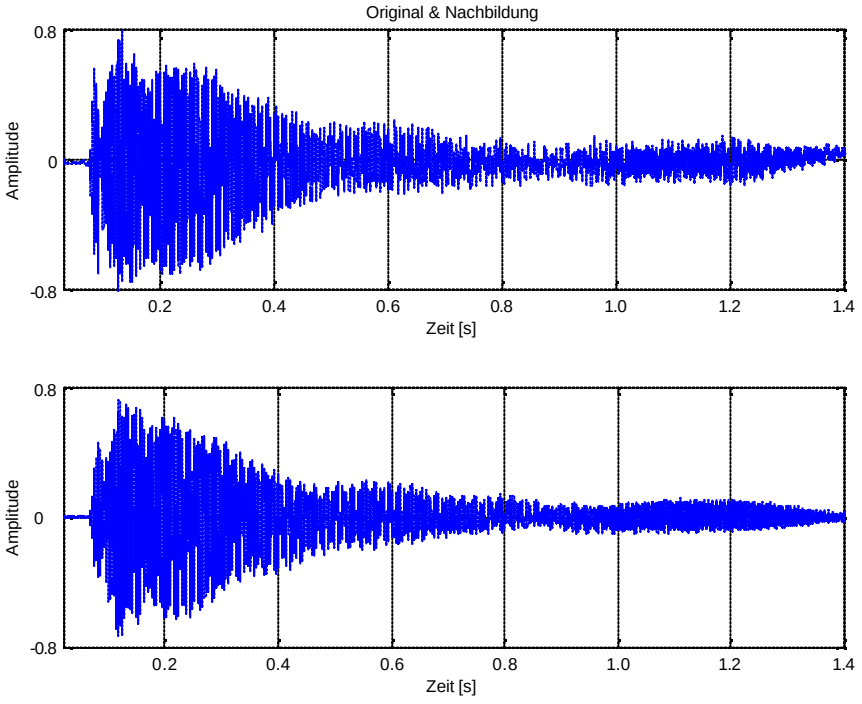


Abb. 6.5: zeitliche Darstellung des Klaviertones

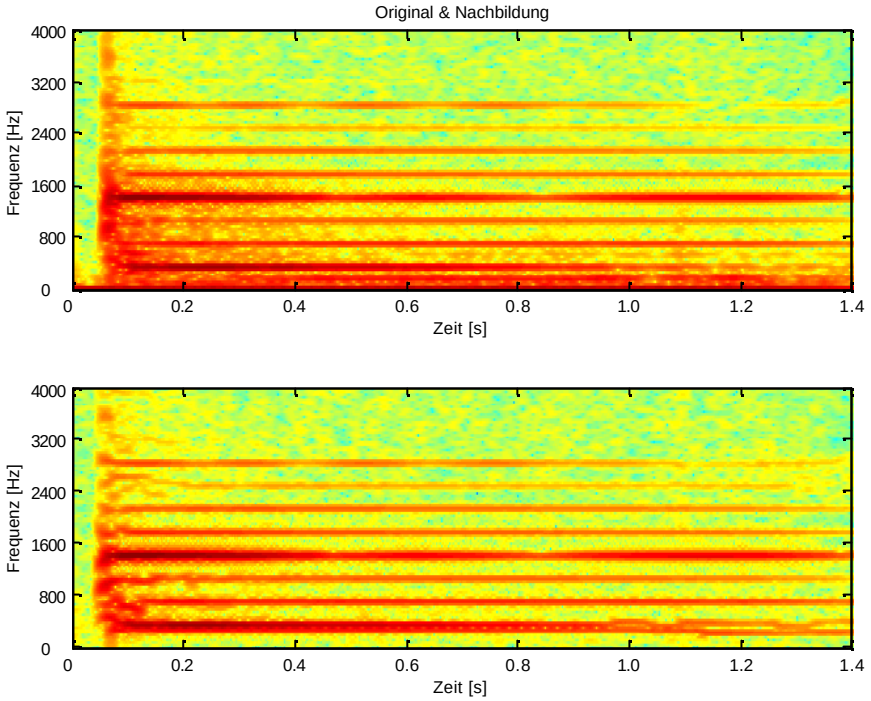


Abb. 6.6: Frequenzspektrum, N=1024

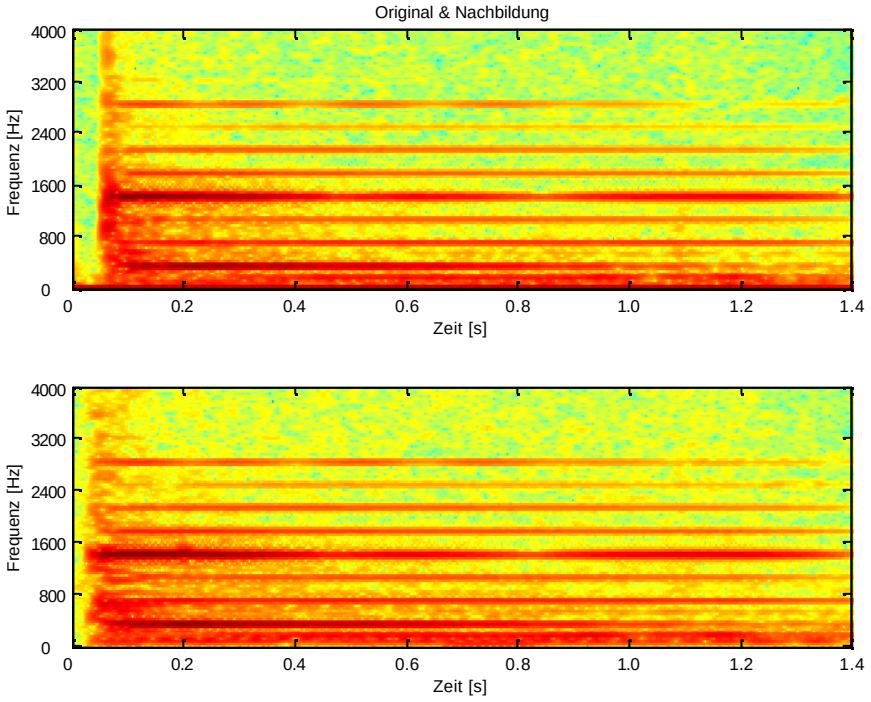


Abb. 6.7: Frequenzspektrum, N=4096

Durch die Vergrößerung von N von 1024 auf 4096 Punkte wird eine Verbesserung der Analyse im tieffrequenten Bereich erzielt, welche nicht nur grafisch, siehe Abb. 6.7, sondern auch akustisch ein deutlich zufriedenstellenderes Ergebnis liefert. Ein Nachteil, der dabei entsteht und ebenfalls in Abb. 6.7 zu erkennen ist, ist die Verschlechterung der zeitlichen Struktur des Signals. Der Beginn des Tones ist deutlich verschwommen bzw. erscheint "vorgezogen".

Schwierigkeiten ergeben sich bei der Verarbeitung von Signalen mit Spektren, deren Frequenzanteile teilweise eng beieinander liegen und welche zudem noch aufgrund von Einschwingvorgängen eine akkurate zeitliche Auflösung erfordern. Dies ist z.B. bei zwei aufeinanderfolgenden Klaviertönen der Fall, welche wiederum von einem tieffrequenten Störgeräusch überlagert sind. Abb. 6.8 zeigt, dass weder mit 2048 noch mit 4096 Punkten der FFT ein zufriedenstellendes Ergebnis erzielt werden konnte. Auch die Variation der Hop Size ergab keine wesentliche Verbesserung (vgl. die beiden unteren Grafiken).

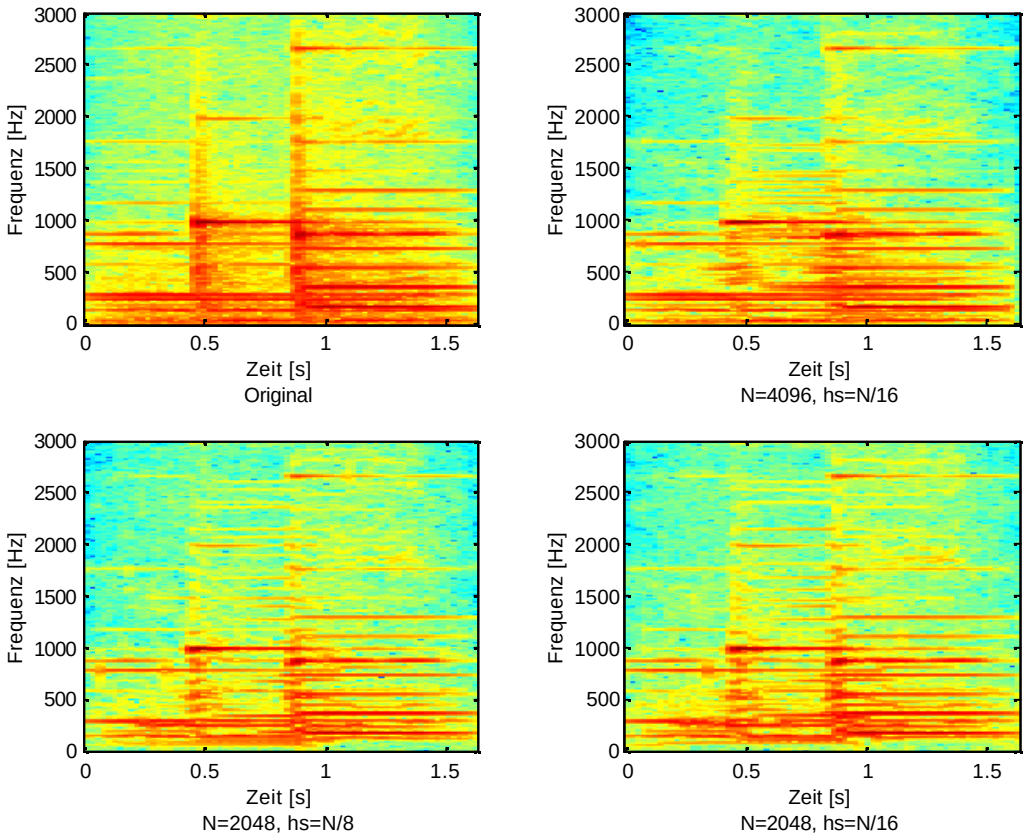


Abb. 6.8: zwei aufeinanderfolgende Klaviertöne

Resümee: Während der Testphase hat sich gezeigt, dass mit Hilfe der Funktion "Pre Scan" relativ leicht eine grobe Einstellung der "General Threshold" gefunden werden kann. Manchmal führt der Versuch der Verfeinerung jedoch zu folgender nachteiligen Eigenschaft des Programms: Diese Schwelle, die das erste Auswahlkriterium für einen gültigen/ungültigen Teilton des Signals darstellt, ist eine fixe, von der Zeit unabhängige, Größe. Sie kann auch als statisch bezeichnet werden. Da die Amplituden der einzelnen Teiltöne eines Musik- oder Sprachsignals zeitlich variabel sind, wäre ein weiterer Verbesserungsvorschlag, diese Schwelle dynamisch zu realisieren.

Die restlichen drei Parameter der "Additional Settings" zeigen sich wider Erwarten als eher unempfindlich und ihre Beeinflussung ist oft nicht leicht erkennbar. Generell kann die Einstellung aller Analyseparameter durch gute Vorkenntnisse über die Frequenzstruktur sowie die zeitliche Entwicklung des Signals erleichtert werden.

Literaturverzeichnis

[Auger, et al]
Auger F., Flandrin P., Gonçalvès P., Lemoine O.: 'Time-Frequency Toolbox for the use with MATLAB, Tutorial'; Centre National de la Recherche Scientifique, 1995/96

[Auger, Flandrin, 1994]
Auger F., Flandrin P.: 'La réallocation: une méthode générale d'amélioration de la lisibilité des représentations temps-fréquence bilinéaires'; Centre National de la Recherche Scientifique, Report, 1994

[Auger, Flandrin, 1995]
Auger F., Flandrin P.: 'Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method'; IEEE Trans. on Signal Processing, Vol. 43, No. 5, Mai 1995

[Cohen]
Cohen L.: 'Time-Frequency Distributions: A Review'; IEEE Proceedings, Vol. 77, No. 7, Juli 1989

[Depalle, Hélie]
Depalle Ph., Hélie T.: 'Extraction of Spectral Peak Parameters using a Short-Time Fourier Transform Modeling and no Sidelobe Windows'; IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Oktober 1997

[Depalle, Tromp]
Depalle Ph., Tromp L.: 'An Improved Additive Analysis Method Using Parametric Modelling of the Short-Time Fourier Transform'; ICMC Proceedings, 1996

[Dolson]
Dolson M.: 'The Phase Vocoder: A Tutorial'; CMJ, Vol. 10, No. 4, Winter 1986

[Fitz, et al]
Fitz K., Haken L., Christensen P.: 'Transient Preservation under Transformation in an Additive Sound Model'; ICMC Proceedings, 2000

[Gabor]
Gabor D.: 'Theory of Communications'; J. IEE, Vol. 93, Part III, No. 26, November 1946

[Hainsworth, et al]
Hainsworth S.W., Macleod M.D., Wolfe P.J.: 'Analysis of Reassigned Spectrograms for Musical Transcription'; IEEE Workshop of Signal Processing to Audio and Acoustics, Mohonk, NJ. October 21-24th 2001

[Hainsworth, Wolfe]
Hainsworth S.W., Wolfe P.J.: 'Time-Frequency Reassignment for Music Analysis'; Proceedings ICMC, Havana, Cuba., September 17-22nd, 2001

[Harris]
Harris F.J.: 'On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform'; IEEE Proceedings, Vol. 66, No. 1, Januar 1978

[Helstrom]
Helstrom C.W.: 'An expansion of a signal in Gaussian elementary signals'; IEEE Trans. on Information Theory, Vol. IT 12, 1966

[Höldrich]
Höldrich R.: 'Zur Analyse und Resynthese von Klangsignalen unter Verwendung von Zeit-Frequenz-Repräsentationen mit verbesserter Lokalisation der Signalenergie'; Dissertation, Technische Universität Graz, 1994

[Kodera, et al]
Kodera K., Gendrin R., deVilledary C.: 'Analysis of Time-Varying Signals with Small BT Values'; IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP 26, No. 1, Februar 1978

[Mallat]
Mallat S.: 'A Wavelet Tour of Signal Processing'; Academic Press, ISBN 012466606X, 1999

[McAulay, Quatieri]
McAulay R.J., Quatieri T.F.: 'Speech Analysis/Synthesis Based on a Sinsoidal Representation'; IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP 34, No. 4, August 1986

[Mecklenbräuker, Hlawatsch]
Mecklenbräuker W.F.G., Hlawatsch F.: 'The Wigner Distribution – Theory and Applications in Signal Processing'; Amsterdam, The Netherlands, Elsevier, 1997

[Montgomery, Reed]
Montgomery L.K., Reed I.S.: 'A generalization of the Gabor-Helstrom transform'; IEEE Trans. on Information Theory, Vol. IT 13, 1967

[Plante, et al]
Plante F., Meyer G., Ainsworth W.A.: 'Improvement of Speech Spectrogram Accuracy by the Method of Reassignment'; IEEE Trans. on Speech and Audio Processing, Vol. 6, No. 3, Mai 1998

[Rodet]
Rodet X.: 'Musical Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models'; IEEE Time-Frequency and Time-Scale Workshop 1997

[Serra]
Serra X.: 'A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochstic Decomposition'; Dissertation, Center for Computer Research in Music and Acoustics, Stanford University, Report Stan-M-58, August 1989

[Serra, Smith]
Serra X., Smith J.O.: 'Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition'; CMJ, Vol. 14, No. 4, Winter 1990

[Smith, Serra]
Smith J.O., Serra X.: 'PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation'; Center for Computer Research in Music and Acoustics, Stanford University, Report Stan-M-43, August 1987

[Tolonen, et al]
Tolonen T., Välimäki V., Karjalainen M.: 'Evaluation of Modern Sound Synthesis Methods'; Helsinki University of Technology, ISBN 951-22-4012-2, Report 48, März 1998

[Verma, Meng]
Verma T.S., Meng T.H.Y.: 'Extending Spectral Modeling Synthesis with Transient Modeling Synthesis'; CMJ, Vol. 24, No. 2, Summer 2000

[Williams]
Williams Ch.S.: 'Designing Digital Filters'; Prentice/Hall International, Inc., 1986

[Zwicker, Feldtkeller]
Zwicker E., Feldtkeller R.: 'Das Ohr als Nachrichtenempfänger'; S. Hirzel Verlag, Stuttgart, 1967