Phasengenaue Sinustonanalyse mittels der Desam - Toolbox

Seminararbeit aus Algorithmen in Akustik und Computermusik 2

Peter Innerhofer

Betreuung: Dr. Franz Zotter, DI Matthias Frank

Graz, 3. Mai 2011

kunst uni graz

institut für elektronische musik und akustik

Zusammenfassung

Diese Arbeit soll einen gegliederten Überblick an fourierbasierter Audioanalyse bieten. Hauptaugenmerk liegt in Verbesserungsvorschläge zur klassischen Sinustonanalyse mittels der Short-Time-Fouriertransformation (STFT). Die Reassignment-Methode [LM07] stellt eine solche Optimierung dar. Die Desam-Toolbox [LBD+10] implementiert diese Reassignment-Methode und bietet darüber hinaus weitere umfassende Signalverarbeitungswerkzeuge. Unter Verwendung der Desam-Toolbox entstand eine Audioanalyse- und Syntheseapplikation, in der eine Reihe an Testsignalen analysiert, synthetisiert und verglichen wurde.

Inhaltsverzeichnis

1 Einleitung

Audioanalyse und Synthese ist ein äußerst umfangreiches Thema. Diese Seminararbeit kann nur einen kleinen Einblick ermöglichen. Hauptbestandteil ist die Verwendung der DESAM-Toolbox [LBD+10] für Matlab und Octave. Daneben werden einige Grundlagen der Audioanalyse theoretisch behandelt.

Die Desam-Toolbox bietet eine gut gegliederte state-of-the-art Sammlung von Signalverarbeitungswerkzeugen. Beispielhafte Anwendungsgebiete können dabei Musik-Information-Retrieval (MIR) Aufgaben werden, aber auch automatische Transkriptionen oder allgemeine Audioanalyse und Synthese. Die Toolbox ist gegliedert in Sinuston und spektrenbasierte Modelle, das Sinustonmodell wiederum in Langzeit- und Kurzzeit-Analyseverfahren. Diese Arbeit befasst sich ausschließlich mit dem fourierbasierten Sinustonmodell. Die Abbildung 1 bietet einen Überblick.

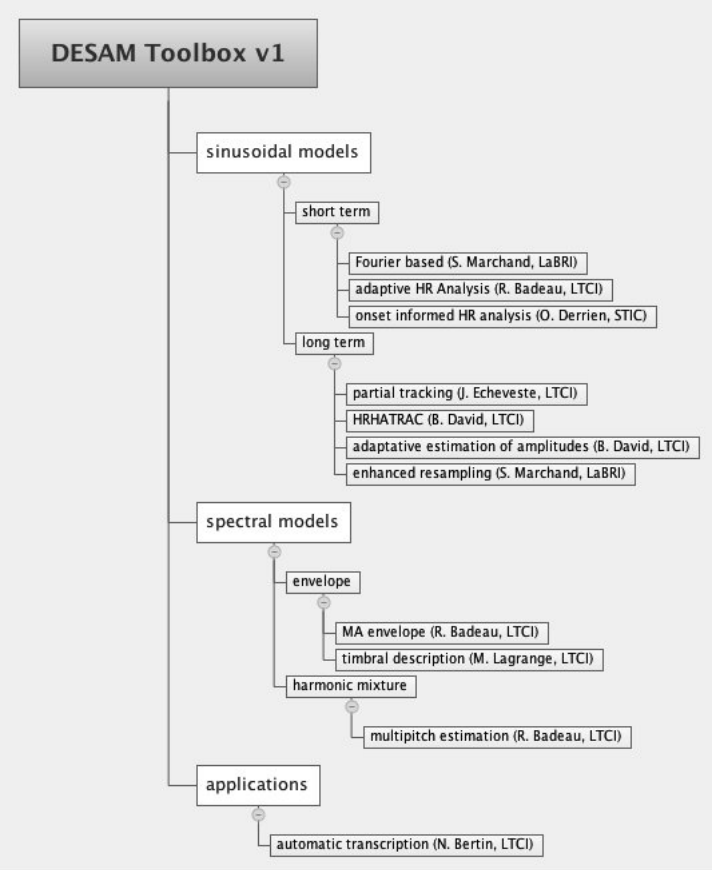


Abbildung 1: Organisation der Desam Toolbox. Quelle [LBD+10]

2 Grundlagen der Audioanalyse

Sinustonanalysen in praktischen Anwendungen benötigen eine genaue Bestimmung der Frequenzen und dessen zeitliches auftreten, da alle weiteren Verfahrensschritte darauf aufbauen. Schnelle Verfahren können über Kurzzeit Analysemethoden wie in der *Short-Time-Fouriertransformation* (STFT) implementiert werden. Um die dadurch entstandenen Ungenauigkeiten besser zu bewältigen, werden verschiedene Ansätze gewählt.

Eine Klasse an FFT-basierten Näherungen (engl. *estimators*) betrachtet das Energiespektrum rund um eine Frequenzkomponente und passt eine analytische Funktion danach an (z.B: polynominal); so werden exaktere Frequenzwerte errechnet [LM07]. Dazu gehört etwa die Applikation Parshl [SS86] [MQ86].

Eine zweite Klasse benutzt zusätzlich zum Frequenzspektrum die Phaseninformation um die Frequenzwerte zu schätzen. Dazu gehört die Reassignment Methode und das phasendifferenzbasierte Modell, wie sie in *Phasevocoder* eingesetzt werden [LM07].

Ein allgemeines Audiosignal s wird hier als eine von Summe von Sinusteiltöne mit den Parametern

$$s(t) = \sum_{p=1}^{P} a_p(t) exp(i\varphi_p(t)) \tag{1}$$

P, die Anzahl der Teiltöne, der Amplitude $a_p(t) = a_p exp(\mu_p t)$ und eine Funktion der Phase $\varphi_p(t) = \varphi_p + \omega_p t + \frac{1}{2}\psi_p t^2$ angenommen.

Als phasenbasiertes Sinustonmodell stellt die Desam-Toolbox die Reassignment-Methode zur Verfügung. Diese Methode, und welche Unterschiede zur reinen STFT bestehen, werden nachfolgend näher beschrieben.

2.1 Reassignment-Methode

Die Reassignment Methode versucht einen Nachteil der "Short-Time-Fouriertransformation" (STFT) zu verbessern. Der Nachteil besteht aus den kurzen Signalausschnitten, die in der STFT behandelt werden. Ein kleines Zeitfenster $h(t)$ führt zu einer großen Auflösung in der Zeit; hingegen führt ein kleines Zeitfenster zu einer schlechten Auflösung in der Frequenz ω. In der Literatur wird dies als Zeit-Frequenz *Trade-Off* bezeichnet [LM07]. Die STFT wird wie folgt berechnet

$$S_w(\omega, t) = \int_{-\infty}^{+\infty} s(\tau) w(\tau - t) e^{-j2\pi\omega(\tau - t)} d\tau, \tag{2}$$

wobei $w(t)$ die Fensterfunktion und $s(t)$ das Signal repräsentiert.

Eine andere Methode ist, den kurzen Signalausschnitt durch *zero-padding* zu verlängern, und dadurch eine höhere Auflösung in der Frequenzdomäne zu erreichen. Zero-padding erhöht aber nicht die Information im Signal und wie bereits besprochen, wollen schnelle Anwendungen kurze Signalausschnitt behandeln.

Ein weiteres Verfahren stellt die Wavelet-Transformation dar. Diese Transformation wählt unterschiedliche Auflösungen für unterschiedliche Frequenzen, adaptiert also die Länge des Zeitfenster für verschiedene Frequenzbänder [Mey93].

Die Reassignment-Methode basiert in Theorie der Fensterung (Windowing). Anstatt das Zeitfenster beliebig in der Länge zu wählen, wird das Fenster bzw. die Fensterfunktion adaptiv, abhängig vom Signal, verändert. Ideal wäre es, das selbe Signal zeitverehrt als Fensterfunktion zu verwenden [FACM03]. Ein so entstandenes Signal nennt man Wigner-Ville Verteilung (Wigner-Ville Distribution, WVD). Nachteile der Wigner-Ville-Distribution sind ihre teilweise negativen Werte, sowie das Übersprechen von interagierenden Komponenten, charakterisiert von oszillierenden Spektralanteilen - siehe Abbildung 2. Der daraus resultierende Ansatz fährt eine Summierung der WVD über alle Zeitwerte bzw. Frequenzwert des Spektogramms durch. Diese, durch Summierung resultierende Werte, werden als neue Energiemassepunkte den spektralen Werten zugeordnet, sprich reassigned. Dies gilt sowohl in der Zeit- als auch in der Frequenzdomäne. Man unterscheidet deshalb zwischen Zeit- und Frequenz-Reassignment. In welcher Form die Theorie der Fensterung mittels Reassignment in einen Algorithmus umgewandelt wird, hängt stark von dessen Implementierung ab, wie in [FACM03] beschreibt.

Die Reassignment-Methode nach Kodera, De Villedary und Gendrin [KGV78] berechnet neue Frequenz und Zeitwerte anhand der Gruppenlaufzeit und der Momentanfrequenz, respektive der Ableitung der Phase nach der Frequenz und der Zeit [LM07]

$$\hat{t}(t,\omega) = \frac{1}{2\pi}\frac{\partial}{\partial\omega}\varphi(t,\omega), \tag{3}$$

$$\hat{\omega}(t,\omega) = \frac{1}{2\pi}\frac{\partial}{\partial t}\varphi(t,\omega). \tag{4}$$

In der Gleichung 1 erhält man die Phase durch $Im\big(ln(a_p(t)e^{i\varphi_p(t)})\big) = \varphi_p(t)$. Folglich ergibt sich Frequenz-Reassignemts über [LM07, s. 8].

$$\hat{\omega} = \frac{1}{2\pi}\frac{\partial}{\partial t}\varphi(t,\omega) = \frac{1}{2\pi}Im\left(\frac{\partial}{\partial t}ln\big(S_w(\omega,t)\big)\right) \tag{5}$$

$$= \frac{1}{2\pi}Im\left(\frac{\frac{\partial}{\partial t}\big(\int_{-\infty}^{+\infty} s(\tau)w(\tau-t)e^{-j2\pi\omega(\tau-t)}d\tau\big)}{S_w(\omega,t)}\right) \tag{6}$$

$$= \frac{1}{2\pi}Im\left(\frac{j2\pi\omega S_w(\omega,t) - S_{w'}(\omega,t)}{S_w(\omega,t)}\right) \tag{7}$$

$$= \omega - \frac{1}{2\pi}Im\left(\frac{S_{w'}(\omega,t)}{S_w(\omega,t)}\right). \tag{8}$$

wo $S_{w'}$ das Signal mit abgeleiteter Fensterfunktion und S_w das gefensterte Signal bezeichnet. Beim Zeitreassignment wird Gleichungen 4 zu

$$\hat{t}(t,\omega) = t - Re\left(\frac{S_w^t(t,\omega)}{S_w}\right) \tag{9}$$

wo $S_w^t(t,\omega)$ die Fouriertransformierte des Signals multipliziert mit der Zeit bezeichnet. Ein solches, neu zugeordnete Spektrums ist in Abbildung2 dargestellt.

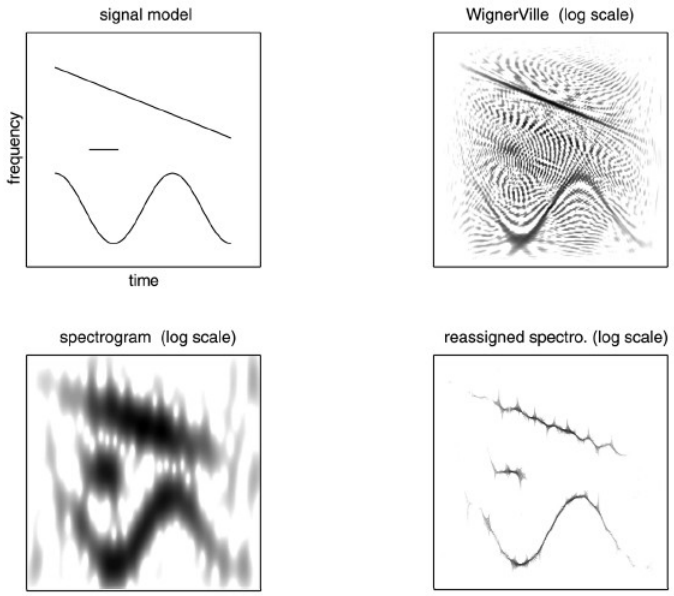


Abbildung 2: Dreikomponentensignal eingebettet in Rauschen. Ein ideales Zeit-Frequenz-Modell wird hier mit drei Bildern verglichen: Erstellt mit der Wigner-Ville Verteilung (oben rechts) einem schlecht aufgelösten Spektogramm (unten links) und dem Zeit-Frequenzbild der Reassignment Methode (unten rechts). Quelle [LM07]

3 Verwendung der Desam-Toolbox

Vorweg wird das phasenbasierte Sinustonmodell der Toolbox näher beschrieben. Anschließend folgt eine Beschreibung der entstandenen Applikation in Verwendung der Toolbox. Am Ende dieses Kapitels werden die Resultate der Analyse und Resynthese mittels der Desam-Toolbox dargestellt.

3.1 Die Reassignment-Methode in der Desam-Toolbox

Die Reassignment Methode im der Desam-Toolbox implementiert folgendes Modell:

$$s(t) = \sum_{p=1}^{P} a_p(t) exp(i\varphi_p(t)) \tag{10}$$

P bezeichnet die Anzahl der Frequenzscheitel, $a_p(t) = a_p exp(\mu_p t)$ eine modulierte Amplitude und $\varphi_p(t) = \varphi_p + \omega_p t + \frac{1}{2}\psi_p t^2$ eine Phasenmodulation. Die Rückgabewerte der Reassignmentfunktion bezeichnen die Amplitude a, die Amplitudenmodulation μ,

die Phase φ, die Frequenz ω und die Frequenzmodulation ψ eines spektralen Frequenzscheitelpunktes. Jedes Analysefenster wird nach den zero-phase-Pinzip [Smi07] vorweg gefiltert. Damit liegt keine falsche Phaseninformation vor. Dieses Modell implementiert eine Funktion der Phase zweiten Grades, und berücksichtigt somit lineare Frequenzänderungen innerhalb eines Fensters (Frequenzmodulation) sowie Phasenmodulation. Anders als in der Arbeit von McAulay [MQ86] oder dem Parshl-Programm bietet dieses Modell keine Möglichkeit der nichtlinearen Änderung der Frequenz. Hierfür müsste eine Funktion der Frequenz dritten Grades errechnet werden. Eine solche Funktion kann nur dann errechnet werden, wenn zwei hintereinander liegende Fenster betrachtet werden und jeweils die Frequenz als auch die Phase der beiden Fenster zur Berechnung der 4 Parameter herangezogen werden, wie bei McAulay [MQ86, S. 750] beschrieben. Als Basis für diese Berechnungen könnte aber die Reassignment-Methode eingesetzt werden. In der Praxis erweist sich aber schon die Bestimmung des Parameter ψ als nicht besonders Stabil und so stellt sich die Frage ob eine Funktion der Phase 3ter Ordnung notwendig ist.
Die Audiosynthese erhällt in der Desam-Toolbox wenig Aufmerksamkeit. Zur Verfügung steht eine Funktion zur Erzeugung des Signal aus den zuvor berechneten Parametern. Diese Signalausschnitte könnten danach, z.B: mittels der Overlapp-Add-Methode, zusammengeführt werden.

3.2 Die Applikation

Zur Übersicht die Verfahrensschritte der entstandenen Applikation:

1. Trennung des Eingangsignal in Signalblöcke
2. Analyse mittels der Reassignment Methode
3. Synthese eines Teilsignals anhand des Modell
4. Fensterung und Addition der Teilsignale

ad. 1: Das Signal wird in Teilsignalen zerlegt. Die Wahl der Hop-Size R ist gebunden an die Wahl der Fensterlänge/FFT-Bins. Die Hop-Size bezeichnet die Anzahl der diskrete Signalwerte zum nächsten Teilsignal. Da in der Analyse ein Hann Fenster benutzt wird, entspricht die ideale Hop-Size R = N/2/2 [SS86].

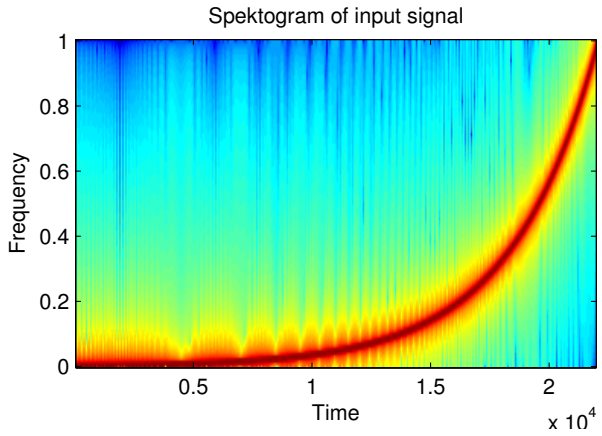
ad. 2: Bei der Reassignment-Methode kann die Anzahl der partiellen Frequenzscheitel angegeben werden, für welche obiges Modell angepasst wird. In diesem Schritt fehlt manches an Logik (Frame-to-Frame Peak Matching / Peak Continuation) wie sie in ausgereiften Peak-Tracking Applikationen wie in Parshl [SS86, s.13] Verwendung findet.

ad. 3 u. 4: Synthetisiert wird das Signal mittels der Overlapp-Add-Methode. Hierbei wird das Teilsignal aus den Analyseparametern erzeugt, danach mit dem Fenster multipliziert und mit Rücksicht der Hop-Size addiert. Wie zuvor bereits besprochen könnte die Parameter aus der Analyse dazu benutzt eine vollständige analytische Repräsentation eines Audiosignals zu errechnen und anhand dieser Repräsentation das Signal zu synthetisieren. Dies hätte aber den Umfang dieser Seminararbeit gesprengt.

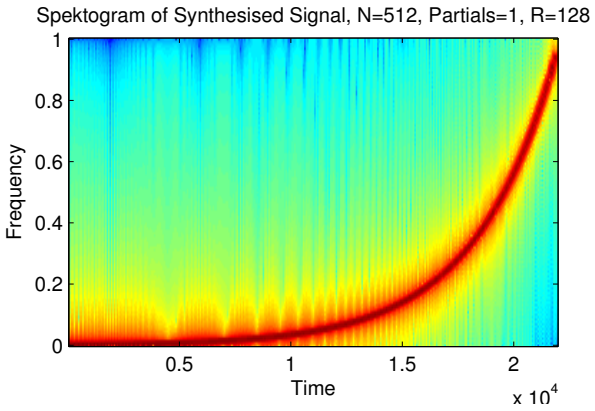
3.3 Resultate

Im Folgenden werden einige Testsignale analysiert, synthetisiert und dargestellt. Ein wichtiges Analysesignal ist der sogenannte Sweep. Der Sweep beschreibt eine quasi-exponentielle Frequenzänderung (kumulative Summe) bei gleich bleibender Amplitude. Weitere Testsignale waren ein menschliches Pfeifen, das Instrument Erhu (klassisches chinesiches Instrument), eine steirische Knopfharmonika und eine menschliche Stimmprobe, jeweils mit unterschiedlichen Fensterlängen und unterschiedlicher Anzahl der Teiltönen.

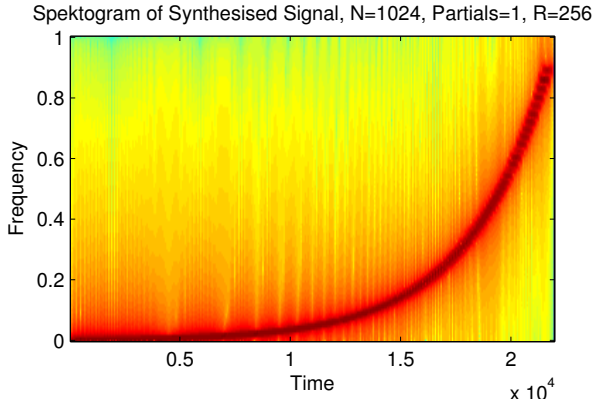
Die Präsentation der folgenden Diagrammen lässt keine genauen Vergleiche oder Messungen zu, jedoch können grundsätzliche Unterschiede diskutiert werden. Wichtiges Vergleichsmaterial bieten auch die synthetisierten Audiodateien mit ihren Originalen.



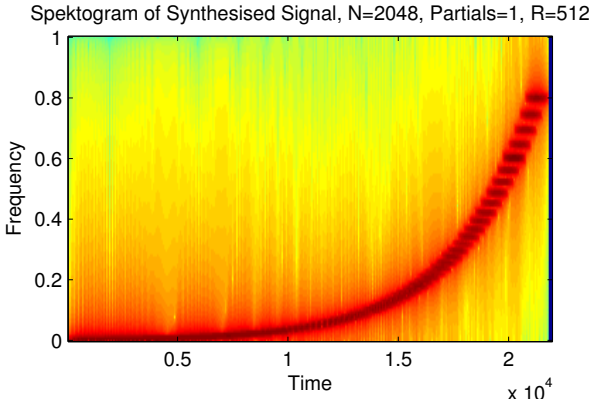
(a) input sweep



(b) synthesised sweep, N=512

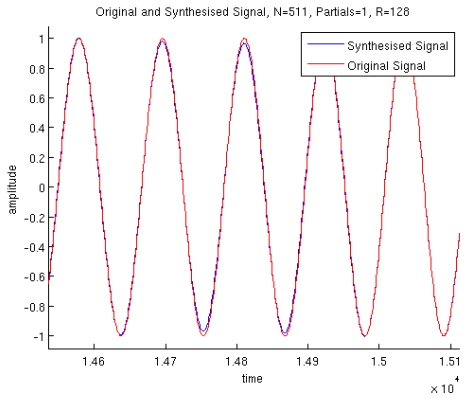


(c) synthesised sweep, N=1024



(d) synthesised sweep, N=2024

Abbildung 3: Vergleich unterschiedlicher Fensterlängen (N=512/1024/2048). Hier ist ersichtlich dass kurze Zeitfenster schnellen Frequenzänderungen leichter folgen können.



(a) Sweep Zeitdomäne (b) Sweep Zeitdomäne

Abbildung 4: Synthetisierter Sweep N=511/1023. Es kann ein qualitativer Unterschied zwischen den Audiodaten gehört werden. Kürzere Zeitfenster (N=511) können schnellen Frequenzänderungen besser folgen.

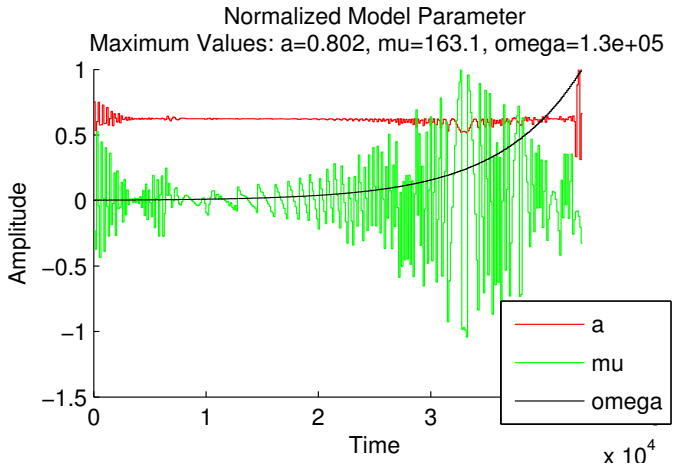
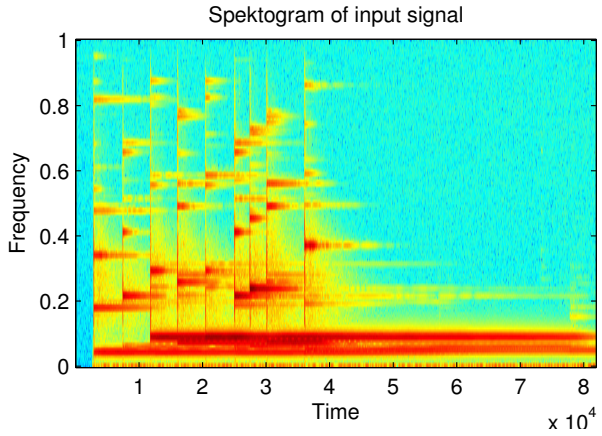
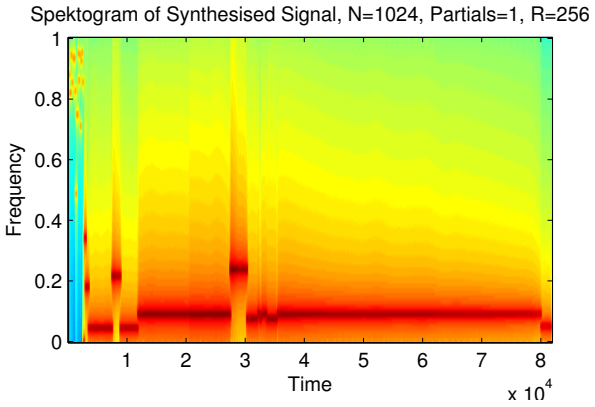


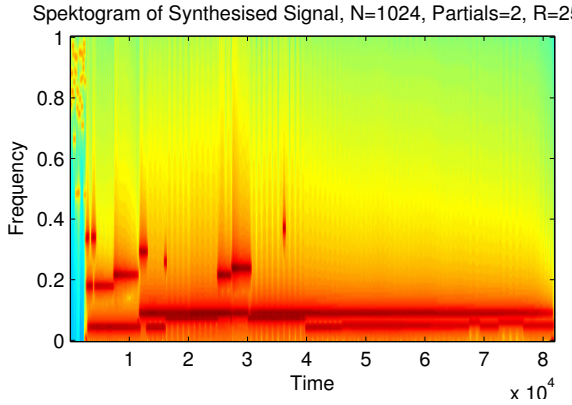
Abbildung 5: Normalisierte Modellparameter. Die Amplitude bleibt stabil, die Frequenz ω steigt korrekt an, μ schwingt teils stark, genauso sowie ψ hier der Übersicht weggelassen



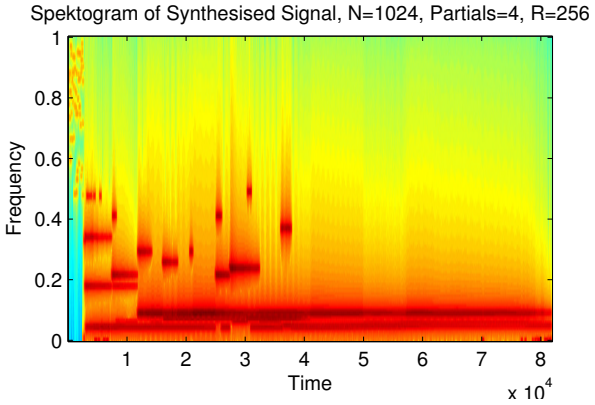
(a) Glockenspiel Input Spectrum



(b) Glockenspiel ein Teilton

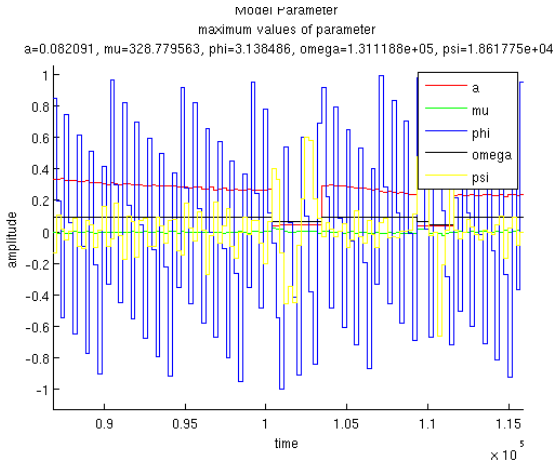


(c) Glockenspiel zwei Teiltöne

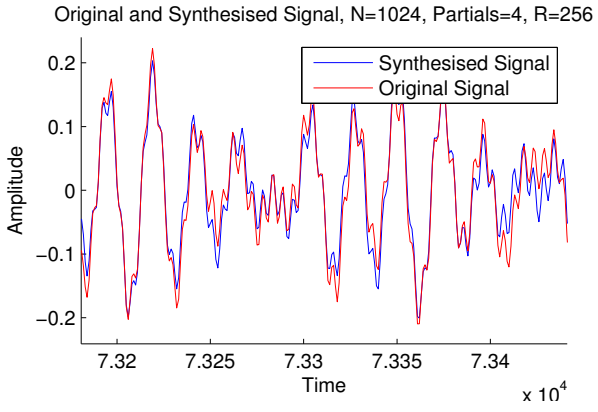


(d) Glockenspiel vier Teiltöne

Abbildung 6: Glockenspiel Synthese 1,2 und 4 Partials

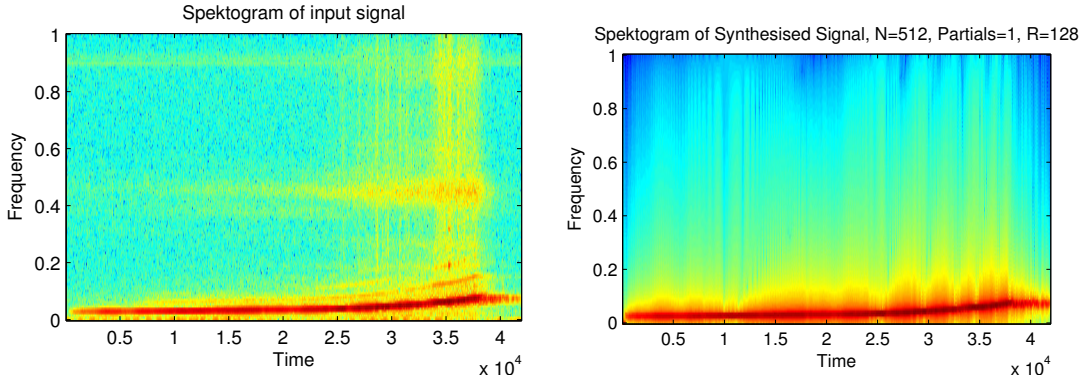


(a) Glockenspiel Synthese, Zeitdomäne



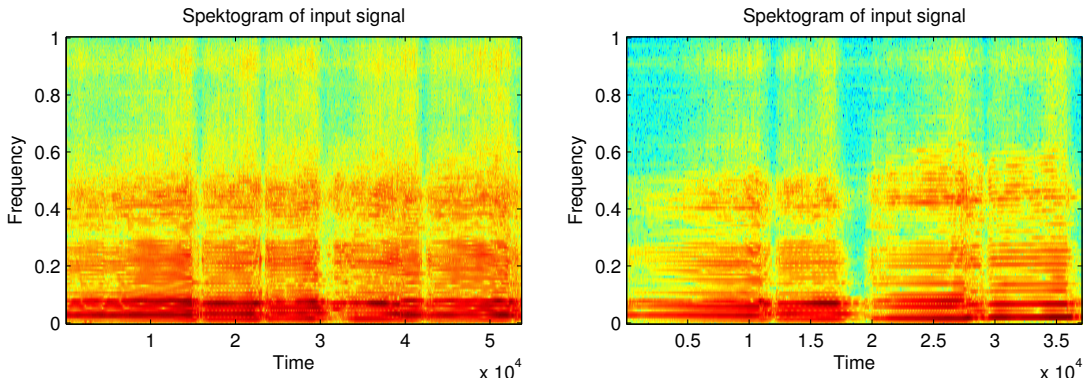
(b) Glockenspiel Synthese, Zeitdomäne

Abbildung 7: Parameter des Glockenspiel, Signalausschnitt des Glockenspiel. Hier fällt die gut gelungene Synthese auf. Diese Beispiel bekräftigt, dass ein fourierbasiertes Verfahren für reine Sinustöne gute Resultate liefert.



(a) Gepfiffenes Eingangsignal (b) Gepfiffenes Signal Synthese

Abbildung 8: Eingangsspektrum eines gepfiffenen Sweep und dessen Synthese. Wie bereits beim Glockenspiel erwähnt, werden in diesem Testfall sehr gute Ergebnisse erzielt.



(a) Instrument Erhu (b) Instrument Erhu

Abbildung 9: zwei Eingangsspektren des Instrument Erhu

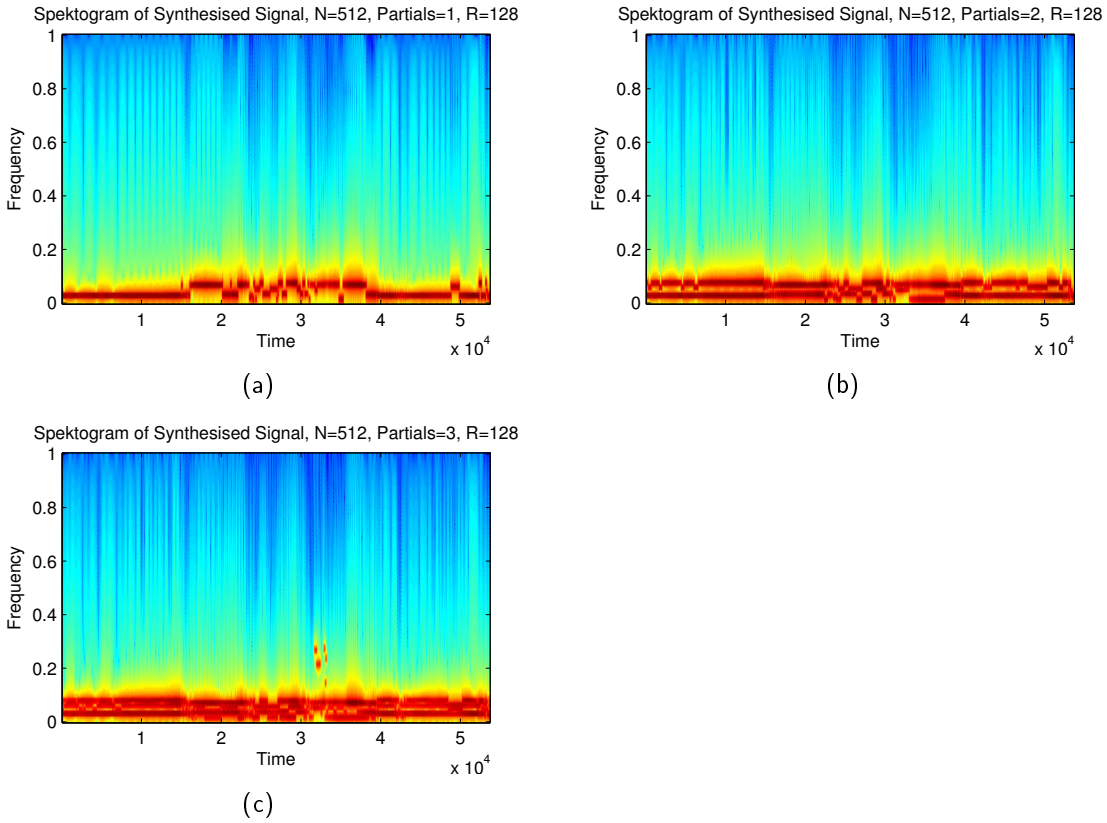


Abbildung 10: Synthesespektrum des Instrument Erhu. Mit steigender Anzahl an Teiltonspuren steigt auch die Audioqualität. Es kommt aber zunehmend zu Phasensprüngen, teils aufgrund der fehlenden Logik, teils durch ein vereinfachtes Modell begründet.

3.4 Diskussion

Bei näherem Studium der Desam-Toolbox hat sich der Autor als außerordentlicher Mitbeleger in ein neues Themengebiet vorgewagt und vertieft eingearbeitet. Dabei wurde ersichtlich, dass die Reassignment-Methode zusätzlich vektorisierte implementiert werden könnte. Dadurch könnten die ganzen Vorteile eines geschärften Spektrums ihre Verwendung finden. So wäre es sinnvoll erst nach dem Reassignment von Frequenz und Zeit die spektralen Scheitelpunkte zu suchen bzw. zu verfolgen.

Eine Erweiterung in der entstandenen Applikation sollte eine Logik hinter dem Peaktracking beinhalten.

Wie bereits erwähnt wäre ein erweitertes Modell wie von McAulay [MQ86] wünschenswert. McAulay schlägt ein Phasenmodell dritten Grades vor. Dies erscheint im Nachhinein als leicht realisierbar, da die Reassignmentfunktion bereits alle benötigten Parameter liefert. Damit könnte aus einer vollständig parametrisierten analytischen Funktion synthetisiert werden. Dies würde verschiedenste Filterungen vor der Synthese ohne qualitative Verluste ermöglichen, z.B: Zeitskalierungen. Die Änderung der Application würde die Rechenzeit kaum erhöhen.

Literatur

[FACM03] P. Flandrin, F. Auger, and E. Chassande-Mottin, *Time-frequency reassignment: From principles to algorithms, in Applications in Time-Frequency Signal Processing*, A. Papandreou-Suppappola, Ed. CRC Press, 2003.

[KGV78] K. Kodera, R. Gendrin, and C. Villedary, "Analysis of time-varying signals with small bt values," *IEEE J ASSP*, vol. 26, no. 1, pp. 64–76, 1978.

[LBD+10] M. Lagrange, R. Badeau, B. David, N. Bertin, J. Echeveste, O. Derrien, S. Marchand, and L. Daudet, "The desam toolbox: Spectral analysis of musical audio." in *Proceedings of the Digital Audio Effects (DAFx'10) Conference*, Graz, Austria, September 2010, pp. 254–261. [Online]. Available: http://dafx10.iem.at/proceedings/papers/LagrangeBadeauDerrienMarchandDaudetDavidBertinEcheveste_DAFx10_P59.pdf

[LM07] M. Lagrange and S. Marchand, "Estimating the instantaneous frequency of sinusoidal components using phase-based methods," *Audio Engineering Society*, vol. 55, pp. 385–399, June 2007.

[Mey93] Y. Meyer, *Wavelets: Algorithms & Applications*. Society for Industrial and Applied Mathematic, 1993.

[MQ86] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, 1986.

[Smi07] J. O. Smith, *Introduction to Digital Filters: with Audio Applications*. W3K Publishing, 2007.

[SS86] J. O. Smith and X. Serra, "Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," https://ccrma.stanford.edu/~jos/sasp/PARSHL_Program.html, 1986.