Universität des Saarlandes
Naturwissenschaftlich-Technische Fakultät I
Fachrichtung Informatik
Bachelor-Studiengang Computer- und Kommunikationstechnik

Bachelorarbeit

Voice-Modeling Based on a Given F0-Track

vorgelegt von
Stefan Densow
am 10. Februar 2009

angefertigt unter der Leitung von
Prof. Dr.-Ing. Thorsten Herfet
betreut von
Dipl.-Ing. Eric Haschke

begutachtet von
Prof. Dr.-Ing. Thorsten Herfet
Prof. Dr. Dietrich Klakow

Erklärung
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und alle verwendeten Quellen angegeben habe.

Saarbrücken, den 10. Februar 2009

_____

Stefan Densow

Einverständniserklärung
Hiermit erkläre ich mich damit einverstanden, dass meine Arbeit in den Bestand der Bibliothek der Fachrichtung Informatik aufgenommen wird.

Saarbrücken, den 10. Februar 2009

_____

Stefan Densow

Zur Trennung einer Mixtur von Sprachsignalen ist die Kenntnis von Orthogonalitäten zwischen den Sprechern nötig. Die in dieser Arbeit vorausgesetzte Orthogonalität der Grundfrequenz, ermöglicht bei ihrer Kenntnis eine Trennung der Sprachsignale in einer Zeit-Frequenz-Darstellung. Dabei werden jedem Sprecher Zeit-Frequenz-Punkte der Sprachmixtur zugeordnet, die die Energien an seinen Harmonischen repräsentieren.

Die Aufgabe dieser Arbeit ist die Erstellung eines Verfahrens, mit dem nur aus dem Verlauf der Grundfrequenz und den orthogonal ausgedünnten Energieverläufen der Harmonischen eines Sprechers dessen ursprüngliches Sprachsignal rekonstruiert werden kann. Dazu wird zunächst die Gewinnung der Energieverläufe der Harmonischen diskutiert. Zudem wird die Entstehung von Sprachschall erläutert und daraus ein theoretisches Modell abgeleitet, welches später als Basis für ein Rekonstruktionsverfahren im Zeitbereich dient.

Separation of a mixture of speech signals requires knownledge of orthogonalities between speakers. In this work, orthogonal fundamental frequencies are assumed, which enables the separation of speech signals based on a time-frequency representation. Every speaker is assigned some time-frequency points, which represent the energy at the harmonics.

The task of this work is the development of a method that allows for the reconstruction of the original speech signal from the knownledge of orthogonally sparse energy tracks and the fundamental frequency track. We first consider the calculation of the harmonic energy tracks. Moreover, we discuss the process of speech production and derive a theoretical model which will serve as a base for a reconstruction method in the time domain.

Inhaltsverzeichnis

1 Einleitung

Die kognitiven Fähigkeiten, die dem Menschen zur Verfügung stehen, erlauben es ihm, seine Umgebung auf vielfältige Weise wahrzunehmen. Das Gehör ist ein besonders beeindruckendes Beispiel, da es nicht nur das bloße Bemerken von Schallereignissen erlaubt, sondern auch die detaillierte Wahrnehmung eines breiten Frequenzspektrums sowie die ungefähre Lokalisation von Schallquellen ermöglicht. Darüber hinaus kann es sich selektiv auf eine Schallquelle konzentrieren und unerwünschte Störgeräusche unterdrücken, sodass sie wesentlich abgeschwächt wahrgenommen werden. Z.B. Sprache auch in Gegenwart ähnlich lauter Interferenzen rein akustisch verstanden werden[1].

Wissenschaftlich erforscht wurde das Sprachverständnis bei Interferenz durch andere Sprecher oder Störgeräusche von Cherry [Aro92], der für die Fähigkeit zur selektiven Aufmerksamkeit des Gehörs den Begriff ,,Cocktail-Party Effekt" prägte.

Die maschinelle Nachbildung dieser Fähigkeiten ist für die Wissenschaft aufgrund der sich öffnenden Möglichkeiten ein sehr interessantes Gebiet. Im Windschatten dieser Bemühungen erweitern sich Verständnis sowie Wertschätzung der kognitiven Fähigkeiten des Menschen.

1.1 Die Motivation: Sprechertrennung

Das Problem der selektiven Konzentration auf einen Sprecher ist im Grunde ein Spezialfall des etwas anspruchsvolleren Problems der Trennung aller Sprecher einer Sprachmixtur. Im zweiten Fall müssen die Signalanteile allen vorhandenen Sprechern zugeordnet werden während dies im ersten Fall nur für einen Sprecher zu gefordert ist. Aus diesem Grund wird im Weiteren nur die Trennung aller Sprecher diskutiert.

Die Trennung einer Mixtur von Sprachsignalen bzw. von Signalen im Allgemeinen erfordert zunächst die Kenntnis von Signaleigenschaften, in denen sich die einzelnen Ausgangssignale voneinander unterscheiden. Es muss also Charakteristika geben, die orthogonal zueinander sind und somit eindeutig getrennt und einem einzelnen Signal bzw. Sprecher zugeordnet werden können. Orthogonalitäten hängen allerdings stark von der Repräsentation des Signals ab, sodass eine äquivalente Darstellung eines Signals in einer anderen Domäne notwendig sein kann, um orthogonale Charakteristika zu finden. So ist bei Trennung einer Mixtur von Sprachsignalen aufgrund ihrer harmonischen Struktur eine Frequenzbereichsrepräsentation gegenüber einer Zeitbereichsrepräsentation zumeist

[1] Bei Kombination mit visuellen Information (z.B. Lippenbewegung des Sprechers) kann das Sprachverständnis noch deutlich gesteigert werden.

vorzuziehen (siehe Abschnitt 3 über die menschliche Stimme), es sei denn jeder Sprecher lässt die anderen ausreden[2].

In [YR04] wird empirisch gezeigt, dass zwei (oder mehrere) Sprachsignale aufgrund ihrer harmonischen Struktur (siehe Abschnitt über menschliche Stimme) zumeist als hinreichend orthogonal im Frequenzbereich angenommen werden können[3], was auch im Weiteren so vorausgesetzt wird. Sollten sich doch einmal die Energien von Harmonischen zweier Sprecher kreuzen, ist dies in der Regel nur von kurzer Dauer und beeinflusst das Trennungsergebnis hoffentlich nur minimal.

Wie später in Abschnitt 3 detaillierter dargestellt, konzentriert sich die Energie eines Sprachsignals im Spektrum zum größten Teil[4] an den Harmonischen der Grundfrequenz f_0. Die erste Harmonische wird allgemein mit der Grundfrequenz gleichgesetzt [TM04]. Zur einfachen Handhabung der Notation sei der Frequenzindex h als Index der bezeichneten h-ten Harmonischen f_h definiert:

$$f_h = h \cdot f_0 \ , \quad h\text{-te Harmonische} \tag{1}$$

Ist die Grundfrequenz bekannt, kann die ungefähre Energie der Harmonischen grob im Frequenzbereich bestimmt und dem entsprechenden Sprecher zugeordnet werden (siehe Abschnitt 2). Dieses Verfahren entspricht dem Erstellen einer binären Zeit-Frequenz-Maske, welche aus einem Spektrogramm diejenigen Zeit-Frequenz-Punkte auswählt, die den Harmonischen eines Sprechers entsprechen. Eventuell auftretende Konflikte[5] können dabei auf unterschiedliche Weise gelöst werden. Für jeden erkannten Sprecher wird so ein harmonisch ausgedünntes Spektrogramm berechnet, das nur noch einen Bruchteil der Zeit-Frequenz-Punkte des Mixtursignals enthält. Anschließend wird die Energie eines jeden Zeit-Frequenz-Punktes bestimmt, sodass sich Zeitverläufe der Energien der Harmonischen (im Folgenden *Energietracks* genannt) ergeben. Im letzten Schritt muss nun aus den verbliebenen sprecherorthogonalen Daten (Grundfrequenztrack und Energietracks) das ursprüngliche Sprachsignal des einzelnen Sprechers rekonstruiert werden.

Um zunächst grundlegende Erfahrungen bei der Rekonstruktion eines Sprachsignals aus orthogonal beschränkten Daten zu gewinnen, werden in dieser Arbeit nur Sprachaufnahmen betrachtet, in denen ein einzelner Sprecher aktiv ist.

[2]Dann wären die Sprecher zeitorthogonal

[3]Dazu wurde berechnet, zu welchem Grad sich die Energieverteilungen mehrerer Sprachaufnahmen im Spektrogramm überlappen. Bei einer Mixtur aus zwei Aufnahmen wurde festgestellt, dass im Mittel ca. 90% der Energien jedes Sprachsignals vom anderen fast unberührt blieben, also orthogonal waren.

[4]Unter der Annahme, dass die Stimmbänder aktiv sind. Die Energie stimmloser Sprachlaute hingegen ist oft spektral sehr breit verteilt.

[5]Z.B. zwei Sprecher haben harmonische Energie am gleichen Zeit-Frequenz-Punkt.

In dieser Arbeit sollen zunächst grundlegende Erfahrungen bei der Rekonstruktion von Sprachsignalen aus orthogonal beschränkten Daten gewonnen werden. Daher werden nur Sprachaufnahmen eines einzelnen Sprechers und ohne Interferenzen betrachtet.

Die verwendeten Aufnahmen stammen aus der CMU Arctic Datenbank [KB03], die kurze, phonetisch reiche Sätze mehrerer einzelner Sprecher zur Verfügung stellt. Parallel zu den akustischen Aufnahmen sind auch Elektroglottogramme bereitgestellt, welche die Aktivität der Stimmbänder bei der Phonation reflektieren (siehe Abschnitt 3).

Das in dieser Arbeit verwendete Verfahren zur Grundfrequenzbestimmung wurde in einer vorangegangenen Arbeit [Krä08] erstellt.

1.2 Ziele dieser Arbeit

Das wesentliche Ziel dieser Arbeit ist die Rekonstruktion des ursprünglichen Sprachsignals eines einzelnen Sprechers aus der harmonisch bzw. orthogonal ausgedünnten Zeit-Frequenz-Darstellung des unverfälschten Signals. Das einzige Vorwissen besteht aus den Verläufen der Grundfrequenz und der Energien an den Harmonischen. Das wichtigste Kriterium an die Rekonstruktion ist die Verständlichkeit des wiederhergestellten Sprachsignals, sodass alle inhaltlichen Informationen erhalten bleiben und von Mensch und Maschine (z.B. ASR-System) gleichermaßen gut verstanden werden können.

Eine Einschränkung muss allerdings gemacht werden. Da die Orthogonalität von Sprachsignalen einzig von der Grundfrequenz f_0 abhängt, beschränkt sich die Rekonstruktionsmöglichkeit auf Sprachabschnitte, an denen die sie bekannt ist. Kann kein f_0 gemessen werden, weil vielleicht das entsprechende Verfahren fehlerhaft arbeitet oder die Stimmbänder schlicht nicht aktiv sind, existieren keine verwertbaren Orthogonalitäten, sodass kein Signal rekonstruiert werden kann. Um diesem Fall gerecht zu werden, müssten die verfügbaren Ausgangsdaten erweitert werden.

1.3 Aufbau

In Abschnitt 2 wird zunächst allgemein die Bestimmung der Energie eines Signals diskutiert und im Anschluss die Berechnung der harmonischen Energietracks auf Basis einer gegebenen Grundfrequenz erläutert. Darauffolgend wird eine Übersicht über Verfahren zur Bestimmung der Grundfrequenz von Sprache im Zeit- und Frequenzbereich gegeben und die Funktionsweise der hier verwendeten Methode dargestellt. Im Anschluss wird eine Methode zur Bestimmung von Amplituden- und Frequenzmodulationsanteilen in Signalen vorgestellt, welche in Abschnitt 3 zur Anwendung kommt. Dort werden die

Funktionsweisen von Stimmbändern und Vokaltrakt näher beleuchtet, um die Entstehung von Sprachschall zu verstehen. Daraus werden sich wichtige Erkenntnisse ergeben, die in Abschnitt 4 in die Rekonstruktion einfließen. Dort werden zwei Rekonstruktionsverfahren betrachtet, eines im Frequenzbereich und eines im Zeitbereich, und ihre Ergebnisse bewertet. In Abschnitt 5 wird die gesamte Arbeit zusammengefasst und ein Ausblick über mögliche Schritte zur Verbesserung der genannten Verfahren gegeben.

2 Signalverarbeitungswerkzeuge

Bevor wir uns Sprachsignalen widmen, seinen zunächst noch einige Diskussionen von benutzten Verfahren zur Analyse von Signalen vorangestellt. Besonders im Falle der präzisen Bestimmung der Energie im Frequenzbereich werden sich daraus wichtige Erkenntnisse ergeben.

Weiterhin werden eine kurze Zusammenfassung von Verfahren zur Bestimmung der Grundfrequenz eines Signals gegeben sowie eine Möglichkeit zur Abschätzung von Amplituden- und Frequenzmodulationsanteilen vorgestellt.

2.1 Diskrete Signalenergie

Die Energie eines zeitkontinuierlichen Signals $x(t)$ ist in der Signalverarbeitung in Anlehnung an die Physik definiert als

$$E_s := \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df \,, \tag{2}$$

wobei $x(t) \circ\!\!-\!\!\bullet X(f)$. Der zweite Schritt in der Gleichung entstammt dem Parseval-Theorem und entspricht der Energieerhaltung der Fourier-Transformation. Für den diskreten Fall gilt

$$E_s = \sum_{n=1}^{N} |x[n]|^2 = \frac{1}{N} \sum_{n=1}^{N} |X[n]|^2 \,, \tag{3}$$

wobei $x[n] \circ\!\!-\!\!\bullet X[n]$. Die Normalisierung $\frac{1}{N}$ hat ihre Ursache in den orthogonalen (aber nicht orthonormalen) Basisvektoren der diskreten Fourier-Transformation (DFT). Die genaue Formulierung ist allerdings eine Frage der Definition der DFT.

Die Energie eines Signals kann demnach im Zeit- und Frequenzbereich berechnet werden. Diese Eigenschaft wird von Nutzen sein, da eine Repräsentation der Energie wünschenswert ist, die die Zerlegung und separate Bestimmung der Energie an jeder im Signal darstellbaren Frequenz gestattet. Dafür bietet sich die Nutzung der Frequenzbereichs an.

Für gewöhnlich werden Signale betrachtet, die sich über der Zeit in ihren Charakteristika (Amplitude, Frequenz, etc.) ändern. Die Fourier-Transformation ist aber zunächst eine globale Methode, die alle Punkte des betrachteten Zeitsignals in die Berechnung von dessen Spektrum mit einbezieht. Frequenzen, die nur am Anfang oder Ende des Signals präsent sind, tauchen gleichermaßen im Spektrum auf, können aber nicht mehr ohne Weiteres zeitlich lokalisiert werden[6]. Diese Informationen sind zwar im Phasengang des

[6]Man kann auch sagen, das Signal ist im Frequenzbereich maximal spektral und minimal zeitlich

Spektrums enthalten, gehen aber bei der Energiebildung verloren. Plötzliche Änderungen (z.B. Attack-Transienten beim Anschlagen einer Gitarrensaite) sind somit zeitlich delokalisiert und einzelne kurze Pulse werden in ihrer Wirkung auf den gesamten Signalausschnitt verteilt. Deswegen lässt sich immer nur einen kleiner Ausschnitt des Signals unter der Annahmen betrachten, dass dessen Charakteristika dort hinreichend (abhängig von Signalart und Verwendung des Spektrums) konstant sind.

In der Praxis wird dies umgesetzt, indem ein zeitdiskretes Signal (diskretisiert mit der Samplingrate f_s) in Ausschnitte (*Frames*) bestehend aus N Samples zerlegt wird, die sich zu einem gewissen Grad überlappen können. Die überlappenden Frames der zeitlichen Länge $T_f = N/f_s$ erlauben die Berechnung eines Spektrums nicht nur im Intervall aufeinanderfolgender, nicht-überlappender Frames, sondern auch zu Zeitpunkten dazwischen und stellen so eine Art Zeitinterpolation dar. Dadurch kann der zeitlich glättende bzw. delokalisierende Effekt der Fourier-Transformation zwar nicht verhindert, aber dennoch etwas verringert werden. Übliche Überlappungsgrade sind 50% oder 75%. Alternativ kann die Überlappung auch als die Verschiebung benachbarter Frames um K Samples angegeben werden. Eine Verschiebung von $K = N/4$ entspricht einem Überlappungsgrad von $\frac{N-K}{N} = \frac{3}{4} \hat{=} 75\%$.

Nach dem Zerlegen des Signals in Frames der Länge N errechnet die DFT aus jedem Frame N in der Regel komplexe Koeffizienten [AO99], die diesen Signalauschnitt im Frequenzbereich repräsentieren[7]. Damit die Rücktransformation ein reelles Signal produziert, müssen die komplexen Koeffizienten Hermitesche Symmetrie[8] aufweisen, was implizit das Auftreten von Koeffizienten bei (mathematisch) positiven und negativen Frequenzwerten fordert.

Die im diskreten Signal korrekt repräsentierbaren Frequenzen sind laut dem Nyquist-Shannon Sampling Theorem allerdings auf Werte bis $f_s/2$ beschränkt. Damit wird klar, dass die N komplexen Koeffizienten im Frequenzbereich die Bandbreite $-f_s/2 \ldots f_s/2$ beschreiben. Aufgrund der Hermiteschen Symmetrie sind für uns aber nur die Hälfte der Koeffizienten interessant. Unabhängig von dieser Redundanz lässt sich die spektrale Breite eines komplexen Koeffizienten bzw. die Frequenzauflösung *df* angeben:

$$df = \frac{f_s}{N}$$

lokalisiert. Im Zeitbereich ist es umgekehrt. Ein Mittelweg wird z.B. von Wavelet-Transformationen beschritten, welcher hier aber nicht unmittelbar von Vorteil ist.

[7] Genaugenommen werden $2N$ Werte (N Realteile und N Imaginärteile) aus dem Zeitbereich in $2N$ Werte im Frequenzbereich transformiert. Von denen sind jedoch die Hälfte redundant, da hier nur reelle Zeitsignale betrachtet werden.

[8] $X(-f) = X^*(f)$ im kontinuierlichen Fall und analog im diskreten Fall.

Die Bandbreite f_s kann demnach in N Abschnitte ($Bins$)[9] unterteilt werden, jeder beschrieben durch einen komplexen Koeffizienten. Das Spektrum eines Frames beschreibt das Signal für die Dauer des Frames T_f, sodass man jedem Bin des Spektrums dieses Frames zusätzlich zur spektralen Ausdehnung df auch eine zeitliche Ausdehnung T_f zusprechen kann. Ein Bin hat dann die Zeit-Frequenz-Fläche

$$df \cdot T_f = df \cdot N \cdot T_s = df \cdot \frac{N}{f_s} = \frac{f_s}{N} \cdot \frac{N}{f_s} = 1\,. \tag{4}$$

Daher bleibt nur ein Freiheitsgrad übrig, entweder die Framelänge T_f oder die gewünschte Frequenzauflösung df. Die Framelänge unterliegt dabei den eingangs erwähnten Beschränkungen bzgl. der hinreichenden Periodizität des Signals. Praktisch sind der Frequenzauflösung damit Grenzen durch die Änderungsraten der Charakteristika des betrachteten Zeitsignals gesetzt. Bei Sprachsignalen werden in der Praxis oft Werte von $T_f = 20 \ldots 50$ ms verwendet, woraus sich eine Frequenzauflösung von $df = 50 \ldots 20$ Hz errechnet. In Abschnitt 4 wird empirisch ein Wert ab $T_f < 30$ ms als ausreichend empfunden, um die zeitliche Schärfe von plötzlichen Transienten (z.B. Plosive, siehe Abschnitt 3) weitgehend zu erhalten.

Weiterhin ist aus (Gl. 4) ersichtlich, dass die einzige Möglichkeit zur Erhöhung der Frequenzauflösung die Betrachtung eines zeitlich längeren Signalausschnitts ist.

2.1.1 Bestimmung harmonischer Energietracks

Zur Bestimmung der Energien an den Harmonischen ist zunächst die Kenntnis der Grundfrequenz erforderlich. Dazu wird in dieser Arbeit auf das in [Krä08] implementierte Verfahren zurückgegriffen (siehe Abschnitt 2, Grundfrequenzbestimmung), sodass zu jedem Zeitpunkt die Grundfrequenz als bekannt vorausgesetzt werden kann[10]. Das vorliegende Sprachsignal wird zunächst in Segmente mit durchgängigen f_0-Tracks zerlegt und anschließend wie beschrieben in Frames unterteilt und per DFT transformiert. Durch die Verwendung überlappender Frames kann die Zeitauflösung um deren Überlappungsgrad interpoliert werden, sodass die Spektren der Frames im Intervall $\frac{K}{N} \cdot T_f$ nun in Form eines Spektrogramms zur Verfügung stehen. An diesen Zeitpunkten $t_f = n \cdot \frac{K}{N} \cdot T_f$ werden auch die Frequenzen der Harmonischen[11] $f_h(t_f) = h f_0(t_f)$ aus der aktuellen Grundfrequenz

[9] Wenn im Folgenden vom Bin einer Frequenz gesprochen wird, sind die Bins an der positiven und negativen Frequenz im Spektrum gemeint. Die Energie eines Bins bzw. einer Frequenz ist dann als die Summe der Energien beider Bins zu verstehen. Aufgrund der Hermiteschen Symmetrie entspricht diese der doppelten Energie eines der beiden Bins.

[10] Falls keine Grundfrequenz vom gefunden wurde, wird sie zu 0 Hz angenommen.

[11] Die Anzahl der Harmonischen wird so gewählt, dass sie jederzeit unterhalb von $f_s/2$ liegen.

$f_0(t_f)$ bestimmt und die DFT-Koeffizienten aus den entsprechenden Bins des Spektrogramms ausgelesen. Aus ihnen wird dann nach (Gl. 3) die Energie des Bins berechnet und als Energie der entsprechenden Harmonischen gewertet:

$$E_h(t_f) = \frac{2}{N}|X_{t_f}[k_h]|^2 \tag{5}$$

Der Faktor 2 begründet sich aus der Hermiteschen Symmetrie der DFT-Koeffizienten[12] und der Bin k_h enthält die Frequenz der h-ten Harmonischen f_h.

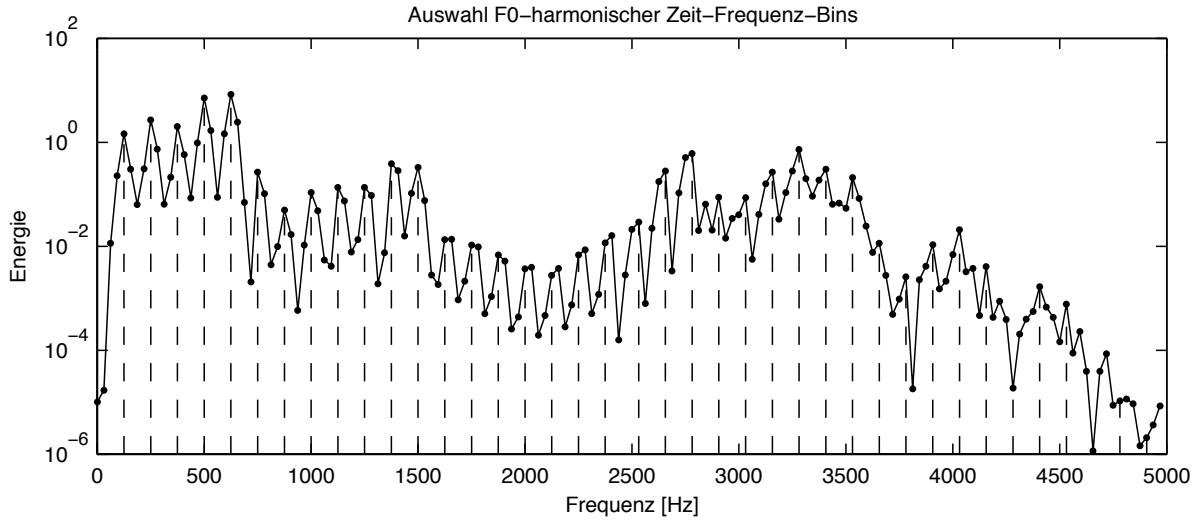

Abbildung 1: Auswahl von Zeit-Frequenz-Bins zur Bestimmung der Energie an den Harmonischen. Die gestrichelten Linien geben die Position der harmonischen Bins an. Das Spektrum stammt aus Frame 90 der Aufnahme, die in Abbildung 2 benutzt wurde. *Parameter:* $T_f = 32$ ms, $df = 31.25$ Hz, $N = 1024$, $K = 128$. Es wurde ein Tukey-Fenster mit $r = 0.8$ verwendet.

In Abbildung 1 ist beispielhaft zu sehen, welche Bins eines Frames als harmonisch erachtet und ausgelesen werden. Über die Länge des Gesamtsignals entstehen so Energietracks, die den Energieverlauf der Harmonischen wiedergeben und die Basis für die Rekonstruktion bilden. In Abbildung 2 sind die generierten Tracks über Zeit, Frequenz und Energie anhand eines längeren, stimmhaften Sprachsignals dargestellt. Darüber hinaus ist das Verhältnis der Energie aller ausgelesenen, harmonischen Bins eines Frames gegenüber der Gesamtenergie des Framespektrums[13] angegeben. Die Basislinie bei ca. 22%

[12]Die Energie einer Frequenz ist in diesem Fall gleichermaßen auf die beiden Koeffizienten bei positiver und negativer Frequenz verteilt. Es wird angenommen, dass nur die Bins der positiven Frequenzen genutzt werden, was die Notation des zweiten Bin-Index erspart.

[13]Durch die Fensterung eines Frames mit einer nicht-rechteckigen Fensterfunktion mit Amplitude 1

gibt das durchschnittliche Energieverhältnis bei zufälliger Wahl der extrahierten Bins an. Es ergibt sich direkt aus dem Verhältnis der ausgelesenen Bins[14] zur Gesamtzahl an Bins im Spektrum.

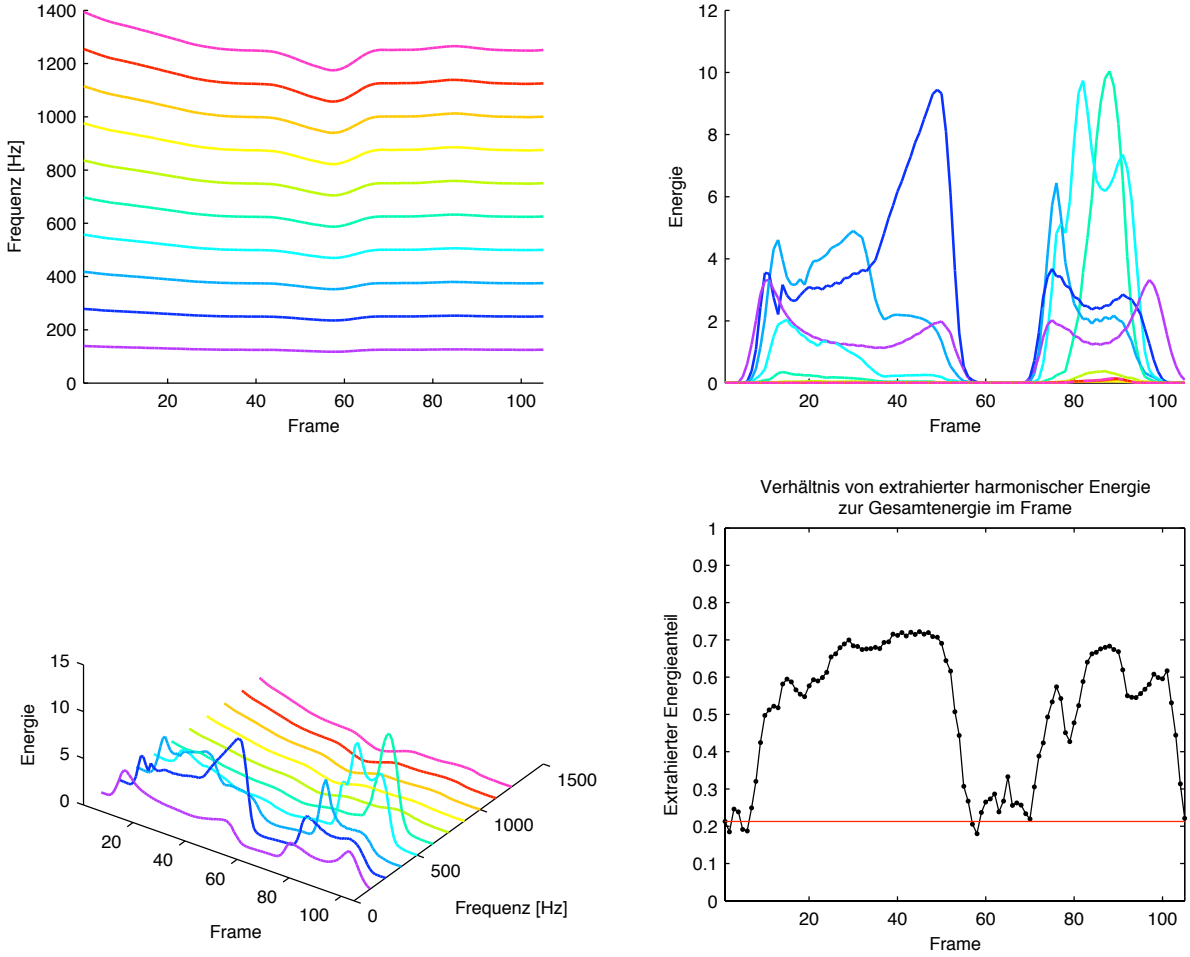

Abbildung 2: Energietracks einer 450 ms langen Aufnahme einer männlichen Stimme. Dargestellt sind die ersten 10 Harmonischen. *Parameter:* $T_f = 32$ ms, $df = 31.25$ Hz, $N = 1024$, $K = 128$. Es wurde ein Tukey-Fenster mit $r = 0.8$ verwendet.

Wir hatten angenommen, dass die meiste Energie eines Sprachsignals an den Harmonischen liegt und daher erwartet, dass der entnommene Energieanteil meist relativ hoch ist. Wie dargestellt, ist das aber keineswegs der Fall.

Die Energiebestimmung ist auf eine korrekte Grundfrequenzmessung angewiesen und kleine Fehler von 1-2 Hz können bei höheren Harmonischen zum Auslesen der falschen

geht Energie im Zeitsignal verloren, was im Spektrum reflektiert wird. Deswegen wurde die Gesamtspektralenergie als Referenz gewählt.

[14]Entspricht der Anzahl der Harmonischen.

Bins führen, da der Fehler durch die Berechnung der Harmonischen Frequenzen $f_h = h \cdot f_0$ linear mit der Frequenz wächst. In der Tat wurden solche Fehler mehrfach beobachtet. Mitunter wurden nicht die Bins von Harmonischen ausgelesen, sondern genau die Bins dazwischen. Solche Fehlabschätzungen scheinen zumindest teilweise dem evtl. zu starken Glätten der Grundfrequenztracks in der Postprocessingphase des F0-Trackers geschuldet zu sein, denn manuelle Messungen mit demselben Verfahren lieferten den korrekten f_0-Wert. Sie können vermutlich minimiert werden, indem z.B. Grundfrequenzbestimmung und Energiebestimmung auf denselben Datenstrukturen bzw. Frames rechnen und nicht lediglich hintereinander aufgerufen auf der selben Sprachaufnahme arbeiten.

Fehlerhafte Energiemessungen aufgrund unpräziser f_0-Werte liefern zum niedrigen Verhältnis der ausgelesenen Energie aber nur einen geringen Beitrag, da sie bei den niedrigen, energiereichen Harmonischen selten zum Auslesen des falschen Bins führen. Und die Bins höherer Frequenzen tragen oft eine so geringe Energie, dass dort ein Fehler wenig bewirkt.

Der wesentliche Grund ist der spektrale Leck-Effekt[15]. Dieser Effekt bewirkt, dass im Allgemeinen selbst eine konstante einzelne Frequenz in einem Frame nicht nur in ihrem Bin für einen Beitrag sorgt, sondern auch in vielen benachbarten Bins, wenn auch mit abnehmender Wirkung über der Frequenz. Die Ursache liegt in der Fourier-Transformation von zeitlich endlichen Signalen, welche im Allgemeinen nicht periodisch fortgesetzt werden können, ohne Nahtstellen (Diskontinuitäten in Amplitude oder Phasen) zu hinterlassen. Man kann ein zeitlich endliches, harmonisches Signal $x(t)$ auch als Multiplikation einer unendlichen harmonischen Schwingung $a(t)$ mit einer Fensterfunktion $w(t)$ betrachten. Die Fensterfunktion ist außerhalb eines endlichen Bereiches Null und schneidet somit ein endliches Signalstück $x(t) = a(t) \cdot w(t)$ der Länge T_f aus der Schwingung heraus. Im Spektrum stellt sich die Multiplikation als Faltung $X(f) = A(f) * W(f)$ dar, wodurch das Spektrum der Fensterfunktion $W(f)$ an jede Frequenz (bzw. an jeden Dirac-Puls) in $A(f)$ platziert wird. Das bislang scharfe Spektrum von $a(t)$ ist nun durch eine Überlagerung von Kopien der spektral sehr breiten Fensterfunktion dargestellt. Aus den Korrespondenzen der Fourier-Transformation ist bekannt, dass ein zeitlich gestrecktes Signal ein schmaleres Spektrum besitzt. Es wird so in Einklang mit (Gl. 4) klar, dass ein zeitlich längeres Fenster zu höherer spektraler Auflösung führt.

Die gebräuchlichste Fensterfunktion ist die Rechtecktfunktion

$$\prod\left(\frac{t}{T_f}\right) \circ\!\!-\!\!\bullet\; T_f \operatorname{sinc}\left(T_f \cdot f\right) = \frac{\sin\left(\pi f T_f\right)}{\pi f} \tag{6}$$

[15] engl. *spectral leakage*

mit der Amplitude 1. Sie lässt die Amplitude des herausgeschnittenen Signalstücks unverändert, hat einen schmalen Peak im Spektrum aber fällt über der Frequenz nur langsam ab. Durch Verwendung eines Rechteckfensters wird die Energie im Bereich des Peaks zwar minimal gestreut, dafür werden aber auch weiter entfernte Frequenzen einen relativ großen Energiebeitrag erhalten, der als Interferenz zu werten ist. Das Rechteckfenster wird immer implizit verwendet, wenn ein endliches Signalstück direkt Fourier-transformiert wird.

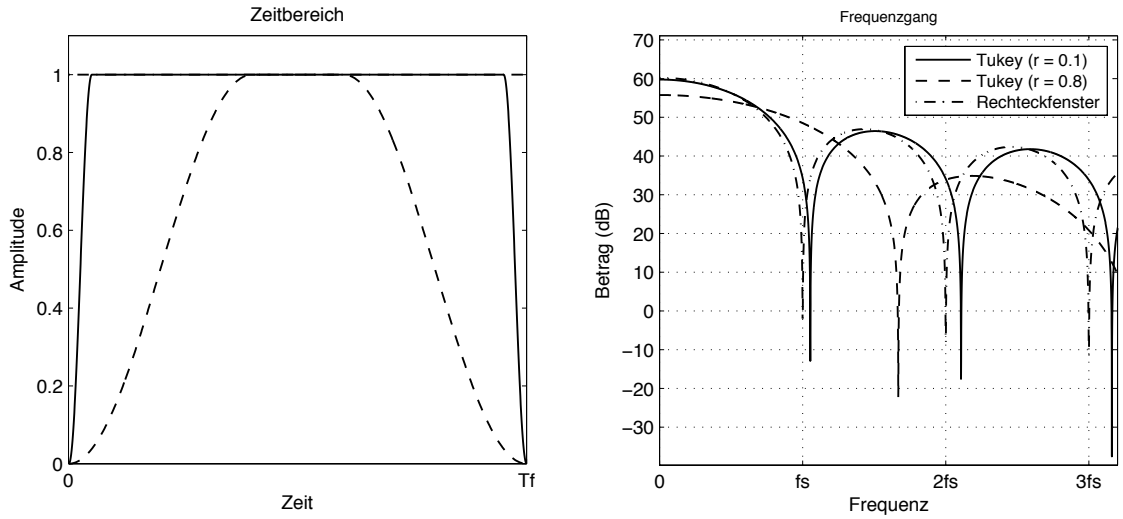

Abbildung 3: Vergleich zweier Tukey-Fenster ($r = 0.1$ und $r = 0.8$) mit einem Rechteckfenster. Die Nullstellen des Rechteckfensters im Frequenzgang fallen auf Vielfache der Samplingfrequenz f_s.

Alternativ können beliebige Fensterformen gewählt werden, welche die gewünschten spektralen Eigenschaften aufweisen. Glattere Fenster fallen in der Regel im Spektrum schneller über der Frequenz ab, haben aber auch einen breiteren Peak. In der Praxis verwendete Fenster sind neben $\sqcap(t)$ z.B. auch das Hann-Fenster (Raised-Cosine), Bartlett-Fenster (Dreiecksfunktion) sowie konfigurierbare Funktionen wie das Gauss-Fenster (Gauss-Glocke), Kaiser-Fenster oder Tukey-Fenster. In dieser Arbeit wird das Tukey-Fenster verwendet, da es sich anhand eines Parameters $r = 0 \ldots 1$ zwischen einem Rechteckfenster und einem Hann-Fenster variieren lässt.

Der Schritt in die Praxis führt nun über Zeit- und Frequenzbereichsabtastung zur DFT. Das bislang kontinuierliche Spektrum $X(f)$ wird nun im Intervall f_s/N diskretisiert [AO99]. An (Gl. 6) ist ersichtlich, dass der Frequenzgang des Rechteckfensters

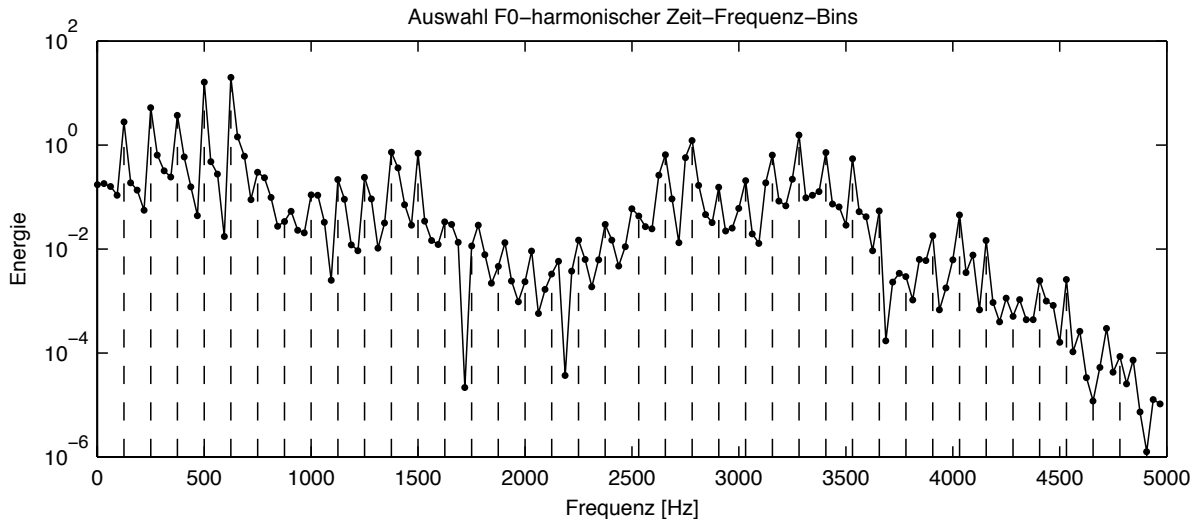

Abbildung 4: Auswahl von Zeit-Frequenz-Bins zur Bestimmung der Energie an den Harmonischen. Die gestrichelten Linien geben die Position der harmonischen Bins an. Das Spektrum stammt aus Frame 90 der Aufnahme, die in Abbildung 5 benutzt wurde. *Parameter:* $T_f = 32$ ms, $df = 31.25$ Hz, $N = 1024$, $K = 128$. Es wurde ein Tukey-Fenster mit $r = 0.1$ verwendet.

(abgesehen von der Stelle $f = 0$) Nullstellen in einem konstanten Intervall besitzt:

$$\frac{\sin(\pi f T_f)}{\pi f} \stackrel{!}{=} 0 \quad \Leftrightarrow \quad fT_f = n \quad , n \in \mathbb{Z}\backslash\{0\} \tag{7}$$

$$\Leftrightarrow \quad f\frac{N}{f_s} = n \quad \Leftrightarrow \quad f = n\frac{f_s}{N} \tag{8}$$

Sitzt eine Frequenz in $A(f)$ zufällig auf einem der Frequenzpunkte f_s/N, fallen die Nullstellen genau auf die gesampleten Frequenzpunkte, sodass in diesem Spezialfall der Leck-Effekt im diskreten Spektrum unsichtbar ist. Beliebige andere Frequenzlagen führen allerdings zum Samplen von $W(f)$ an den entsprechenden Punkten, wodurch sich der Leck-Effekt ins DFT-Spektrum überträgt.

In Abbildung 3 sind zwei Tukey-Fenster mit dem Rechteckfenster im Zeit- und Frequenzbereich verglichen. Deutlich sichtbar ist der höhere und schmalere Peak der Rechteckfunktion.

Für unsere Energieberechnungen bedeutet das nun, dass zunächst immer ein gewisser Teil der harmonischen Energie in benachbarten Bins endet. Zu welchem Grad das geschieht, hängt stark von der verwendeten Fensterfunktion ab. Zum Vergleich sind in Abbildung 5 das Spektrum eines Frames dargestellt, das mit einem Tukey-Fenster und $r = 0.1$ berechnet wurde. Deutlich sichtbar sind die spektral schärferen Harmonischen,

welche ihre Energie auf weniger Bins verteilen als in Abbildung 2. Dadurch steigt der Anteil der entnommen Energie verglichen zur gesamten spektralen Energie wesentlich. Allerdings fällt sind die Energietracks wesentlich unstetiger über der Zeit um den glatten Verlauf in Abbildung 2. Bei der Rekonstruktion wird sich zeigen, dass die stetigen Tracks sauberere Rekonstruktionen zulassen.

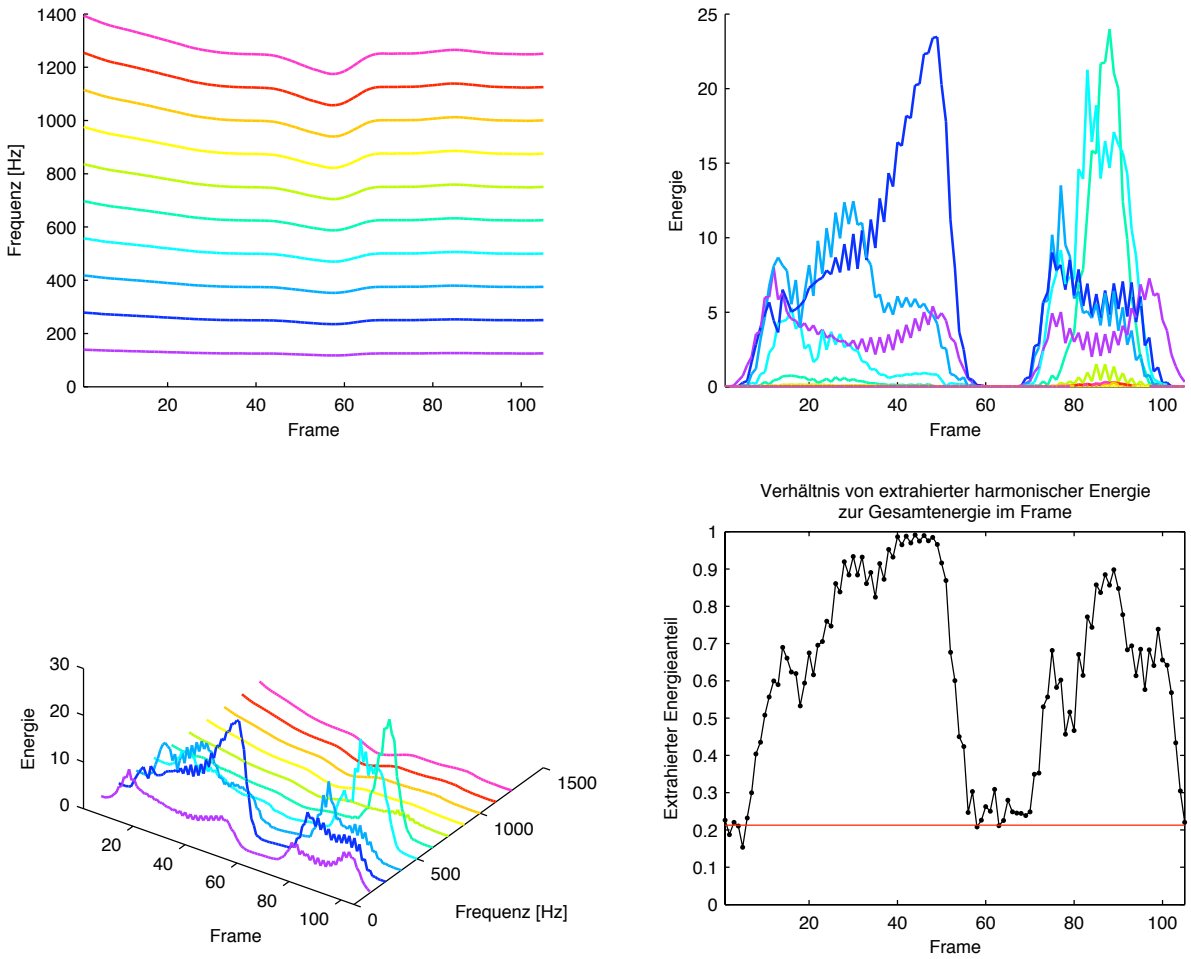

Abbildung 5: Energietracks einer 450 ms langen Aufnahme einer männlichen Stimme. Dargestellt sind die ersten 10 Harmonischen. *Parameter:* $T_f = 32$ ms, $df = 31.25$ Hz, $N = 1024$, $K = 128$. Es wurde ein Tukey-Fenster mit $r = 0.1$ verwendet.

Zusammenfassend ist die DFT zur Bestimmung der Energien von bestimmten Frequenzen zwar ein gangbarer Weg. Die bei Sprachsignalen erreichbare Frequenzauflösung ist gemessen an den vorhandenen spektralen Details (siehe Abschnitt 3) jedoch gerade ausreichend und wird zusätzlich durch den Leck-Effekt reduziert. Bei einer Mixtur von Sprachsignalen muss deshalb von Interferenz durch Harmonische fremder Sprecher ausgegangen werden.

2.2 Grundfrequenzbestimmung

Generell lassen sich die meisten Verfahren zur Bestimmung der Grundfrequenz eines Signals in zwei Kategorien einteilen. Zeitbereichsverfahren rechnen direkt mit dem Zeitsignal während Frequenzbereichsverfahren das Signal zunächst in eine entsprechende Repräsentation überführen (z.B. per DFT, wie in Abschnitt 3, Energiebestimmung). Daneben werden manch andere Methoden auch als Mischform betrachtet. Bei genauerem Blick weisen sie jedoch oft starke Übereinstimmung mit einer der beiden großen Kategorien auf und lassen sich mit ähnlichen Argumenten behandeln. Der folgende Überblick beschränkt sich deswegen auf Verfahren aus den beiden wesentlichen Kategorien und fasst im Wesentlichen die Erkenntnisse der detaillierteren Diskussion möglicher Verfahren in [Krä08] zusammen.

Im Allgemeinen können harmonische Signale aus verschiedenen Konfigurationen einer Grundfrequenz f_0 und h-ten Harmonischen der Frequenz $f_h = h \cdot f_0$ bestehen. Hier seien der Einfachheit halber multiple Grundfrequenzen ausgeschlossen, sodass alle auftretenden Harmonischen auf derselben Grundfrequenz basieren. Darüber hinaus seien zunächst nur rein periodische Signale betrachtet. Es ist möglich, dass in einem Signal nur die Grundfrequenz existiert oder nicht alle rechnerisch möglichen Harmonischen im Signal präsent sind. Genauso kann aber auch die Grundfrequenz fehlen und nur am spektralen Abstand der verbliebenen Harmonischen — direkt oder indirekt[16] — erkennbar sein. Die meisten der im Folgenden kurz angesprochenen Ansätze können die Grundfrequenz in zumindest einem der obigen Fälle nicht erkennen, nur wenige Verfahren stellen sich diesbezüglich als theoretisch robust heraus.

Wie im Abschnitt 3 erläutert wird, sind in unverfälschten Sprachaufnahmen immer die Grundfrequenz sowie mehrere Harmonische enthalten, sofern die Stimmbänder aktiv sind.

2.2.1 Zeitbereichsverfahren

Verfahren im Zeitbereich haben zu allererst den Vorteil des direkten Zugangs zum Signal, da dieses üblicherweise als Zeitsignal verfügbar ist.

Zu den einfachsten Methoden gehört das Auffinden von Peaks bzw. Nullstellen, um

[16] Falls sich zwei rechnerisch benachbarte Harmonische im Signal finden, kann die Grundfrequenz im Spektrum direkt aus deren Frequenzdifferenz bestimmt werden. Ist zwischen präsenten Harmonischen aber immer mindestens eine harmonische Frequenz nicht präsent, muss aus den spektralen Abständen von verfügbaren Harmonischen die Grundfrequenz indirekt bestimmt werden. Dies kann z.B. mit einem Schroeder-Histogramm geschehen [Krä08].

die Grundfrequenz f_0 bzw. Grundperiode $T_0 = f_0^{-1}$ aus deren zeitlichen Abständen zu bestimmen. Allerdings versagt dieser Ansatz wenn das Signal eine komplexere Konfiguration aus Obertönen besitzt. Dann können sehr leicht mehrdeutige Peaks und Nulldurchgänge auftreten, die mit der gesuchten Grundperiode nichts mehr zu tun haben sondern von der Intensität der Obertöne abhängen. Außerdem ist die Methode besonders bei Signalen mit geringer Amplitude sehr empfindlich gegenüber Rauschen, also ungeeignet für niedrige eine SNR[17].

Um dem Problem der zweideutigen Peaks und Nulldurchgänge aus dem Weg zu gehen, kann auch per Tiefpass der Frequenzbereich herausgefiltert werden, in dem die Grundfrequenz vermutet wird. Dies erfordert allerdings eine ungefähre Kenntnis ihres Wertes und kann nur dann zum Erfolg führen, wenn die Grundfrequenz auch im Signal enthalten ist. Alternativ kann versucht werden, die Amplitudenhüllkurve des Signals zu bestimmen, welche oftmals die Periode der Grundfrequenz besitzt. Aber auch hier finden sich Obertonkonfigurationen, bei denen dieser Ansatz zu falschen Ergebnissen führt.

Selbstähnlichkeitsmethoden Eine weitere Möglichkeit zur Bestimmung der Grundfrequenz im Zeitbereich besteht in der Untersuchung der Selbstähnlichkeit eines Signals. Dazu wird ein Ausschnitt des Signals rechnerisch mit einem anderen Ausschnitt desselben Signals verglichen, z.B. per punktweiser Multiplikation oder Subtraktion. Die über der Zeit aufsummierten Produkte bzw. Differenzen geben in Relation zu den Werten anderer Auschnittkombinationen Aufschluss über deren Ähnlichkeit.

Enthält ein Signal nun periodische Komponenten, werden Ausschnitte, die eine oder mehrere Grundperiodenlängen verschoben liegen, eine relativ hohe Ähnlichkeit zueinander aufweisen und einen Peak an der Position der entsprechenden zeitlichen Verschiebung generieren, welche direkt der gesuchten Grundperiode T_0 entspricht. Dieser Ansatz erweist sich in [Krä08] als zuverlässige Methode, die Grundfrequenz von den bisher betrachteten rein harmonischen Signalkonfigurationen zu bestimmen.

Eine Möglichkeit die Selbstähnlichkeit zu messen, beschreibt die Autokorrelationsfunktion (AKF), hier allgemein für den kontinuierlichen Fall sowie konkret mit einer Fensterlänge W für den diskreten Fall angegeben.

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x^*(t+\tau)dt \tag{9}$$

$$R_{xx}[k] = \frac{1}{W}\sum_{n=1}^{W} x[n]x^*[n+k] \tag{10}$$

[17] $\mathrm{SNR} = 10 \cdot \log\left(P_s/\sigma_n^2\right)$ ist das Verhältnis von Signalleistung P_s zu Rauschleistung σ_n^2 und intuitiv ein Maß für die Deutlichkeit eines Signals in Gegenwart von unerwünschten Interferenzen.

Der Parameter τ (bzw. analog k im diskreten Fall) gibt dabei die relative Verschiebung der verglichenen Ausschnitte zueinander an und wird auch als *Lag* bezeichnet. Bei rein harmonischen Signalen werden sich lokale Maxima gleicher Höhe an den Vielfachen der Grundperiode T_0 finden. Das ist bei realen Sprachsignalen allerdings nicht gegeben. Zwei Abschnitte werden in der Regel mit zunehmender Entfernung einander immer unähnlicher, da sich Signalcharakteristika wie z.B. die Grundfrequenz ändern. Dieser Verlust an Ähnlichkeit sollte auch von der AKF reflektiert werden. Wie in [Krä08] gezeigt, kann das im Falle von z.B. linear zunehmender Signalamplitude jedoch nicht gewährleistet werden. Die AKF errechnet dann Peaks an Vielfachen der Grundperiode $\tau = n \cdot T_0$, $(n > 1)$, die über dem Peak des korrekten Werts der Grundperiode bei $\tau = T_0$ liegen, wodurch eine zu große Grundperiode gemessen wird. Dieses Problem wirkt sich besonders bei Sprachsignalen kritisch aus, da Amplitudenänderungen die Regel sind.

Eine Alternative zur AKF besteht in der Bestimmung der Selbstverschiedenheit eines Signals. Dies kann u.a. mit der Squared Difference Function (SDF) geschehen. Dazu wird die Differenz zweier Signalwerte quadriert und über der Zeit aufsummiert:

$$S_{xx}(\tau) = \int_{-\infty}^{\infty} (x(t) - x(t+\tau))^2 dt \tag{11}$$

$$S_{xx}[k] = \frac{1}{W} \sum_{n=1}^{W} (x[n] - x[n+k])^2 \tag{12}$$

Entsprechend dem invertierten Ansatz, der Messung der Verschiedenheit zweier Signalausschnitte, sind bei der SDF nun die Dips von besonderem Interesse, da ihre Position auf große Ähnlichkeit der Signalausschnitte und damit die Grundperiode T_0 hinweist. Bei Betrachtung von überwiegend harmonischen Signalen mit zunehmender/abnehmender Amplitude zeigt sich, dass die Dips der SDF bei höheren Vielfachen der Grundperiode über dem Dip der Grundperiode liegen. Somit kann der niedrigste Dip (unter Vernachlässigung des Dips bei $\tau = 0$) als robuste Abschätzung für die Grundperiode der im Signal enthaltenen Harmonischen betrachtet werden. Der Kehrwert entspricht dann direkt der gesuchten Grundfrequenz.

Damit ist die SDF das einzige der betrachteten Zeitbereichsverfahren, das für alle bislang erwähnten harmonischen Signalformen korrekte Messungen der Grundfrequenz liefert.

2.2.2 Frequenzbereichsverfahren

Die Aufgabe der Bestimmung einer Frequenz legt eine Betrachtung des Problems im Frequenzbereich nahe. Dies geschieht in der Praxis häufig mit derselben Methode, die be-

reits bei der Energiebestimmung diskutiert wurde, und zwar durch Auswertung des per DFT erzeugten Spektrogramms. Auch hier greifen deshalb dieselben fundamentalen Einschränkungen der Frequenzauflösung, die in diesem Kontext der präzisen Frequenzmessung allerdings noch kritischer sind. Wie im Abschnitt 2.1 besprochen sind schon Fehler von $1 \ldots 2$ Hz geeignet, um die Energien der Harmonischen grob fehlerhaft zu bestimmen. Praktisch erreichbare Bin-Größen liegen aber gerade einmal bei maximal $df = 20 \ldots 30$ Hz Breite. Und auch das ist nur mit relativ großen Framelängen $T_f = 50 \ldots 33$ ms zu schaffen, was einer schlechteren Zeitauflösung zu einer führt.

Zusammenfassend muss deshalb schon vor Betrachtung möglicher Frequenzbereichsverfahren geschlussfolgert werden, dass sie selbst bei korrektem Auffinden der Grundfrequenz (bzw. des Bins, in dem sie liegt) deren Wert nicht in ausreichender Präzision bestimmen könnten.

Fazit Aus den genannten Gründen wurde für die Implementierung der Grundfrequenzbestimmung in [Krä08] die SDF-Methode gewählt, da sie robust gegenüber den verschiedensten harmonischen Signalkonfigurationen ist, zeitlich scharf lokalisierte[18], präzise Messung der Grundperiode erlaubt und relativ einfach zu handhaben ist, sowohl in der theoretischen Behandlung als auch in der praktischen Implementierung.

2.3 Modulationsbestimmung

Ein Sprachsignal ist das Resultat des Zusammenwirkens vieler physikalischer Effekte, welche den Luftstrom durch den Vokaltrakt auf unterschiedliche Weise modulieren und prägen. Um in Abschnitt 3 einen Einblick in die Wirkung dieser Modulationen zu erlangen, wird hier zunächst ein Verfahren vorgestellt, welches die effiziente Bestimmung von Amplituden- und Frequenzmodulation in einem bandbegrenzten Signal erlaubt. Aus der Kenntnis dieser Modulationsanteile ergeben sich wichtige Schlüsse über die Zusammensetzung eines Sprachsignals sowie über Möglichkeiten, es naturgetreu zu rekonstruieren.

Viele andere Methoden möglich. siehe [Kve03]

2.3.1 Energy Separation Algorithm

Die Definition der Energie eines Signals kann im Allgemeinen auf vielerlei Weise geschehen. Die Definition in der Signalverarbeitung (Gl. 2) ist an die Elektrotechnik angelehnt und kann als zeitlich aufintegrierte Leistung einer Spannung $x(t)$ an einem Widerstand von 1 Ω verstanden werden.

[18] Ohne Interpolation bis max. zur Samplingrate f_s.

Teager Energy Operator Eine alternative Betrachtung interpretiert ein harmonisches Signal $x(t) = A\cos(\omega t + \theta)$ als Auslenkung einer Masse m eines mechanischen harmonischen Oszillators. Die Frequenz ω ergibt sich unter Annahme einer Federkonstanten k zu $\omega = \sqrt{\frac{k}{m}}$ [TM04]. Aus kinetischer und potenzieller Energie addiert sich die Gesamtenergie des Oszillators zu:

$$E_{\text{Mech}} = \frac{m\dot{x}^2 + kx^2}{2} = \frac{m}{2}A^2\omega^2 \tag{13}$$

Für uns nur der Teil $A^2\omega^2$ der Energie interessant, der proportional zur Energie ist und deshalb im Folgenden als ein Maß der Energie einer harmonischen Schwingung betrachtet wird. Dieses Maß unterscheidet sich zentral von dem in der Signalverarbeitung definierten, da es zum einen instantan[19] und zum anderen frequenzabhängig ist. Eine Schwingung höherer Frequenz erfordert höhere Energie. Die Signalenergie in (Gl. 2) hingegen bewertet Energie unabhängig von der Lage im Spektrum. Beide Energien berücksichtigen die Amplitude einer Schwingung gleichermaßen quadratisch. Gleichung 13 kann somit als ein Maß der Energie verstanden werden, das den mechanischen (und intuitiven) Aufwand zur Erzeugung einer Schwingung reflektiert. Bis auf die explizit erwähnten Ausnahmen bezieht sich die Bezeichnung *Energie* in diesem Abschnitt auf (Gl. 13).

In [Kai90] wird für zeitdiskrete Signale eine Berechnungsvorschrift zur Approximation der instantanen Energie einer harmonischen Schwingung hergeleitet und dem Urheber Herbert M. Teager nach als *Teager Energy Operator* (TEO) $\Psi[\cdot]$ bezeichnet[20]. Alternativ wird der Operator auch *Teager-Kaiser Energy Function* (TKEF) genannt. Der TEO ist ein nicht-lineare Operator und verknüpft ein Signal mit seinen Zeitableitungen.

$$\Psi[x[n]] = x^2[n] - x[n-1]x[n+1] \tag{14}$$

Eine detaillierte Diskussion des kontinuierlichen TEO findet in [MKQ93] statt, während sich [Kai93] allgemeiner mit dessen wesentlichen mathematischen Eigenschaften beschäftigt.

$$\Psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t) \tag{15}$$

Entsprechend der Energie beschreibt auch der TEO eine instantane bzw. im diskreten Fall quasi-instantane Rechenvorschrift. Wie [Kai90] gezeigt, steigt die Präzision der Energieabschätzung bei zunehmender Differenz zwischen Oszillatorfrequenz ω und Samplingfrequenz. Eine Erhöhung der Samplingrate wird hier demnach die Genauigkeit erhöhen. Kriterien für eine ausreichend präzise Energieabschätzung werden im nächsten Abschnitt angegeben.

[19] Die Energie in (Gl. 2) erfordert ein nicht-verschwindendes Zeitintervall zur Bestimmung der Energie während (Gl. 13) instantan auf Änderungen von k oder m reagiert und somit eine Funktion der Zeit ist.

[20] $\dot{x}(t) = \frac{d}{dt}x(t)$

Da die Energieabschätzung durch eine Differenz erfolgt, ist eine Diskussion der Fälle nötig, in denen diese negativ ausfällt. In [BM94] werden der kontinuierliche und diskrete Fall besprochen und ein notwendiges Kriterium an ein Signal $x(t)$ hergeleitet, das die Positivität von $\Psi[x(t)]$ garantiert. Die hier als *Konkavitätskriterium* bezeichnete Bedingung fordert, dass zwischen zwei Nulldurchgängen der Amplitude des Signals eine positive Halbwelle konkav und eine negative Halbwelle konvex verlaufen muss. Mit anderen Worten, der Betrag des Signals muss zwischen zwei Nullstellen konkav sein. Andernfalls wird der TEO an nichtkonkaven Zeitstellen negativ. Für monochromatische Schwingungen, Exponentialfunktionen $x(t) = e^{-at}$, lineare Trends $x(t) = at + c$ oder Produkte dieser Funktionen ist das Konkavitätskriterium immer erfüllt [MKQ93]. Für allgemeine harmonische Signale[21] gilt es aber nicht, da abhängig von den Amplituden der Harmonischen sehr leicht Nichtkonkavitäten entstehen können. Aber auch wenn das nicht der Fall ist, entsteht aufgrund der Nichtlinearität des TEO eine fehlerhafte Energieabschätzung. Die Energie, die der TEO aus der Summe zweier Signale bestimmt, besteht im Allgemeinen aus den Energien der Einzelsignale sowie einem Kreuzterm [Kai93]. Die Summe zweier harmonischer Schwingungen besitzt einen entsprechenden Term, der mit der Differenzfrequenz beider Einzelschwingungen oszilliert [Kai90]. Ist man nur an einer der Energien interessiert, ist also eine Unterdrückung des anderen Signals erforderlich, etwa durch einen Bandpass zentriert um den interessanten Frequenzbereich.

$$x[n] = A_1 \sin(\Omega_1 n + \phi_1) + A_2 \sin(\Omega_2 n + \phi_2) \tag{16}$$

$$E[n] \approx A_1^2\Omega_1^2 + A_2^2\Omega_2^2 + A_1A_2(1 - \cos(\Omega_1 + \Omega_2)) \cos[(\Omega_1 - \Omega_2)n + (\phi_1 - \phi_2)] \tag{17}$$

Die diskreten Frequenzen[22] Ω_i haben die Einheit Radianten/Sample. Im kontinuierlichen Falle gelten die erwähnten Beziehungen analog. Intuitiv kann man dieses Verhalten des TEO verstehen, wenn man bedenkt dass vom Operator eigentlich nicht die Energie eines Signals sondern die eines einfachen harmonisches Oszillators bestimmt, der dieses Signal produziert. Ein solches System ist allerdings ein unzureichendes Modell für Signale, die aus der Überlagerung von eigentlich mehreren Schwingungen bestehen. In der Praxis werden Signale deswegen z.B. zuvor durch einen Bandpass gefiltert [Kve03], um Interferenzen abseits des interessanten Frequenzbereichs auszublenden.

Der Algorithmus Die Klasse an Signalen, welchen eine Energie wie in (Gl.13) zugeordnet werden kann, lässt sich auf amplituden- und frequenzmodulierte Signale verallge-

[21] Bestehend aus einer beliebigen Konfiguration einer Grundfrequenz f_0 und darauf basierenden Harmonischen f_h.

[22] $\Omega = \omega/T_s = \omega f_s$

meinern.

$$x(t) = a(t) \cdot \cos(\phi(t)) \,, \tag{18}$$

$$\text{mit} \quad \phi(t) = \omega_c t + \omega_m \int_0^t q(\tau) d\tau + \theta \,,\ |q(t)| \leq 1 \tag{19}$$

Dabei gilt die Einschränkung $a(t) = 1 + \kappa b(t)$, $0 \leq \kappa \leq 1$ und $|b(t)| \leq 1$, wobei κ der Modulationsfaktor und $b(t)$ das Basisbandsignal für die Amplitudenmodulation (AM) ist. Die instantane Frequenz ergibt sich zu:

$$\omega_i(t) \triangleq \frac{d}{dt}\phi(t) = \omega_c + \omega_m q(t) \,. \tag{20}$$

Dabei ist θ ein beliebiger Phasenoffset, ω_m die maximale Abweichung der instantanen Frequenz von ω_c und $q(t)$ das Basisbandsignal für die Frequenzmodulation (FM).

In [MKQ93] wird gezeigt, wie sich $\omega_i(t)$ und $a(t)$ unter Annahme einiger Bedingungen $x(t)$ aus der Kenntnis von $\Psi[x(t)]$ und $\Psi[\dot{x}(t)]$ recht genau abschätzen lassen. Im Grunde darf sich ein Signal nicht zu schnell bzw. zu stark ändern, um zuverlässig in seiner Energie und damit in Amplitude und Instantanfrequenz abgeschätzt werden zu können. Der präsentierte Algorithmus lautet:

$$\omega_i(t) \approx \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \tag{21}$$

$$|a(t)| \approx \frac{\Psi[x(t)]}{\Psi[\dot{x}(t)]} \tag{22}$$

Intuitiv ist die Idee, die Energie in ihre Anteile von der Amplitude A und der Frequenz ω aufzuspalten, um sie getrennt bestimmen zu können. Daher wird dieses Verfahren allgemein als *Energy Separation Algorithm* (ESA) und im kontinuierlichen Fall als *Continuous Energy Separation Algorithm* (CESA) bezeichnet. Daneben werden in [MKQ93] detaillierte Fehlerabschätzungen der ESA für verschiedene Signalformen durchgeführt und bei generellen AM-FM-Signalen folgende Kriterien hergeleitet:

- AM-Anteil: $\omega_a \ll \omega_c$ und $\kappa \ll 1$
- FM-Anteil: $\omega_f \ll \omega_c$ und $\lambda = \frac{\omega_m}{\omega_c} \ll 1$

Dabei entsprechen ω_a bzw. ω_f den maximale Frequenzen der Basisbandsignale $a(t)$ bzw. $b(t)$, also den maximalen Änderungsraten von Amplitudenumschlag bzw. instantaner Frequenz. Intuitiv darf das Signal über der Zeit also nicht zu schnell oder zu stark variieren. Zusammengefasst sind diese Anforderungen weiche Kriterien, welche zu einem höheren

Fehler bei der Abschätzung von $|a(t)|$ und $\omega_i(t)$ führen, je mehr der verfügbare Spielraum ausgeschöpft wird.

Mit Hilfe diskreten TEO wird in [MKQ93] auch eine diskrete Version des ESA, der *Discrete Energy Separation Algorithm* (DESA) hergeleitet. Abhängig von der gewählten Approximation der Zeitableitung durch einen Differenzenquotienten, ergeben sich verschiedene Versionen, die sich in ihren Eigenschaften allerdings nur minimal unterscheiden. Im Folgenden wird DESA1[23] betrachtet, welcher den Mittelwert aus Vorwärtsdifferenz und Rückwärtsdifferenz zweier Samples bildet.

Betrachten wir nun ein diskretes AM-FM-Signal:

$$x[n] = a[t] \cdot \cos(\phi[n])\,, \tag{23}$$

$$\text{mit} \quad \phi[n] = \Omega_c n + \Omega_m \int_0^n q[k]dk + \Theta \tag{24}$$

Die instantane Frequenz berechnet sich zu

$$\Omega_i[n] \triangleq \frac{d}{dn}\phi[n] = \Omega_c + \Omega_m q[n]\,. \tag{25}$$

Integration und Ableitung sind dabei symbolisch zu verstehn. Alle diskreten Frequenzen Ω sind kleiner als π.

Die DESA1-Methode lautet dann:

$$y[n] = x[n] - x[n-1] \tag{26}$$

$$\Omega_i[n] \approx \arccos\left(1 - \frac{\Psi[y[n]] + \Psi[y[n+1]]}{4\Psi[x[n]]}\right) \tag{27}$$

$$|a[n]| \approx \sqrt{\frac{\Psi[x[n]]}{1 - \left(\frac{\Psi[y[n]]+\Psi[y[n+1]]}{4\Psi[x[n]]}\right)^2}} \tag{28}$$

Die im kontinuierlichen Fall erwähnten Anforderungen an ein Signal $x[n]$ gelten weiterhin.

In Abbildung 6 sind die Abschätzungen von Ψ, dem Amplitudenumschlag, der instantanen Frequenz sowie die entsprechende Fehler für zwei Beispielfunktionen dargestellt. Gut zu erkennen ist der relativ gleichmäßig verteilte Fehler.

In [MKQ93] wurde die Präzision der DESA-Abschätzungen bei Musterfunktionen empirisch über verschiedene Kombinationen von AM (bis zu 50%) und FM (bis zu 25%) berechnet. Im Mittel lagen Amplituden- und Frequenzfehler immer unter 1%. Bei zusätzlichem weißen Rauschen und 30dB SNR lagen die Fehler mit 5-Punkt Median Smoother

[23] Die „1" weist auf die Verwendung eines Differenzenquotienten hin, der nur ein direkt benachbartes Sample betrachtet und so Frequenzen bis $f_s/2$ registrieren kann. DESA2 verwendet dagegen den linken und rechten Nachbarn (Distanz 2), was vergleichbare Genauigkeit liefert, aber nur Frequenzen bis $f_s/4$ wahrnehmen kann. Das kann aber durch Erhöhung der Samplerate korrigiert werden.

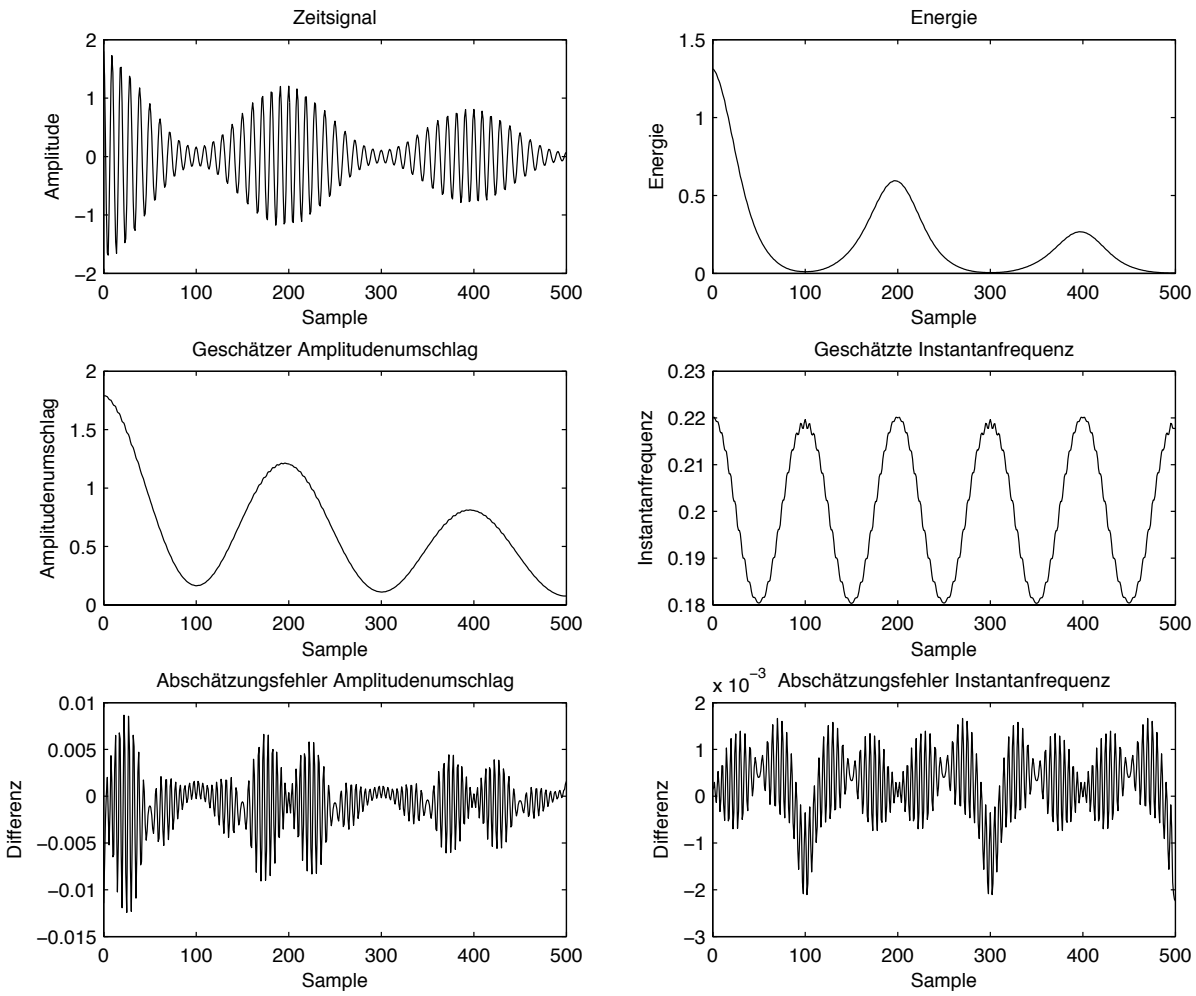

Abbildung 6: Energie und DESA1-Abschätzung des Amplitudenumschlag und der Instantanfrequenz einer exponentiell abfallenen AM-FM-Schwingung mit 80% AM und 10% FM: $x[n] = 0.998^n(1 + 0.8\cos(n \cdot \pi/100))\cos[n \cdot \pi/5 + \sin(n \cdot \pi/100)]$.

gut unter 10%. Die Verwendung eines Median Filters begründet sich in der Beobachtung, dass Abschätzungsfehler üblicherweise als große Amplitudenausschläge gegenüber den direkten Nachbarn auftreten. Fehler dieser Art lassen sich am elegantesten mit einem Median Filter beheben, wohingegen ein Mean Filter den hohen Ausschlag anteilig auf die Nachbarn verteilen würde. Aufgrund der Differenzbildung über wenige Samples ist die Genauigkeit des TEO und damit auch des DESA sehr empfindlich gegenüber Rauschen. Durch geeignete Filter oder Spline-Interpolation [DM01] kann dies allerdings wesentlich reduziert werden.

Allgemein kann für den Fall $\Psi = 0$ die Abschätzung des Amplitudenumschlags an diesem Punkt nullgesetzt und die instantane Frequenz von einem Nachbarwert interpoliert

werden, da angenommen wird, dass sich $\omega_i(t)$ langsamer ändert als $x(t)$.

Vergleich mit Hilbert-Transformation Im vorangegangenen Abschnitt wurde ein Verfahren vorgestellt, mit dem der Amplitudenumschlag und die instantane Frequenz eines in der Regel bandlimitierten Signals bestimmen lässt. Diese Problemstellung aus der Signalverarbeitung bekannt und kann auch mit der Hilbert-Transformation angegangen werden.

[PM94] vergleicht beide Verfahren ausführlich miteinander und findet als wesentlichen Unterschied die höhere Komplexität der Hilbert-Transformation. Sie besitzt allgemein quadratische Komplexität[24] ($\mathcal{O}(N^2)$) und wird zudem über Ausschnitte eines Signals berechnet. DESA ist hingegen in niedriger linearer Komplexität $\mathcal{O}(N)$ berechenbar, da es maximal 5 Signalsamples benötigt (DESA1 und DESA2) und diese mit nur wenigen Rechnoperationen verknüpft [MKQ93].

Bei Sprachsignalen[25] ist der Fehler der Hilbert-Transformation etwas geringer, wobei dies stark von der Länge des FIR-Filters abhängt, mit dem die Hilbert-Transformation approximiert wird. Bei Verhältnissen von ω_a/ω_c bzw. ω_f/ω_c von 100 oder höher (wie z.B. in Kommunikationssignalen) sinkt der DESA-Fehler unter den der Hilbert-Transformation.

Bandpass-Filterung Damit die Anforderungen der ESA an die Bandlimitierung eines Signals in der Praxis gewährleistet werden können, muss es zuvor in der Regel durch einen entsprechend angelegten Bandpass gefiltert werden. Die Eigenschaften des Bandpasses spielen zudem eine Schlüsselrolle bei der Gewährleistung des Konkavitätskriteriums.

Bei der Anwendung der DESA auf Sprachsignale hat sich gezeigt, dass Gabor-Filter diesbezüglich gute Eigenschaften besitzen, während z.B. ein -60dB Equiripple-Bandpass wesentlich mehr Nichtkonkavitäten lieferte. Dies steht vermutlich mit dem exponentiellen Abfall des Frequenzgangs des Gabor-Filters in Zusammenhang, wodurch höhere Frequenzen, die zu solchen Nichtkonkavitäten beitragen, wirkungsvoller unterdrückt werden.

$$h(t) = \frac{\alpha}{\sqrt{\pi}} \exp\left(-\alpha^2 t^2\right) \cdot \cos\left(2\pi f_c t\right) \tag{29}$$

$$H(f) = \frac{1}{2} \exp\left(-\frac{\pi^2 (f - f_c)^2}{\alpha^2}\right) + \frac{1}{2} \exp\left(-\frac{\pi^2 (f + f_c)^2}{\alpha^2}\right) \tag{30}$$

[24] Dies kann auf Kosten der Genauigkeit durch eine gröbere Approximation der Hilbert-Transformation auf Kosten der Genauigkeit verringert werden.

[25] Bzw. allgemein bei Signalen, deren Basisbandbreiten ω_a und ω_f ungefähr ein Zehntel der Trägerfrequenz ω_c betragen.

3 Die menschliche Stimme

In diesem Abschnitt sollen Entstehung und Eigenschaften von Sprachschall diskutiert werden, da die erlangten Kenntnisse für die Rekonstruktion von Sprachsignalen von zentraler Bedeutung sind.

Die Produktion von Sprachschall lässt sich in drei Bereiche unterteilen. Die Lunge stellt einen konstanten Luftdruck bereit, welcher die Stimmlippen (ugs. auch Stimmbänder genannt) im Kehlkopf zu Schwingungen anregt. Der entstandene Schall wird dann im Vokaltrakt (Rachen, Mund und Nasenhöhle) auf vielfache Weise moduliert und durch Mund und Nase abgestrahlt. In der Literatur werden Stimmlippen und Vokaltrakt auch oft als Quelle bzw. Filter des Sprachapparats betrachtet [Krö07]. Die Beteiligung der Stimmlippen bei der Lautbildung ist aber nicht unbedingt erforderlich. Luft kann auch durch den verengten Vokaltrakt gepresst werden, um stimmlose Laute zu erzeugen.

3.1 Phonetisches Vokabular

Zunächst sollen kurz einige linguistische Grundbegriffe eingeführt werden, um die spätere Diskussion zu erleichtern. Das kleinste Sprachschallsegment, welches noch als eigenständig wahrgenommen werden kann, wird als *Phon* bezeichnet. Einzelne Phone werden anhand ihrer akustischen Eigenschaften unterschieden. Falls sie sich unterschiedlich anhören aber sprachliche dieselbe Bedeutung[26] haben, werden sie als zum selben *Phonem* zugehörig kategorisiert. Phone können demnach als Realisierungen von Phonemen betrachtet werden.

Bei der Art der Phoneme lassen sich grundlegend zwei Kategorien unterscheiden, stimmhafte Laute, sog. *Resonanten* und stimmlose Laute, sog. *Obstruenten*. Resonanten werden unter Beteiligung der Stimmbänder gebildet, während diese bei Obstruenten passiv sind. Resonanten können ihrerseits in *Vokale* (z.B. /a/, /e/, ..) und *Nasale* (z.B. /m/, /n/) unterteilt werden. Bei Vokalen lässt der Mund den Luftstrom aus der Lunge frei entweichen, während er bei Nasalen geschlossen ist und der Strom über die Nasenhöhle durch die Nase entweicht (siehe auch Abbildung 7). Bei Obstruenten unterscheidet man im Wesentlichen zwischen *Frikativen* (z.B. /s/, /f/, /v/, /z/) und *Plosiven* (z.B. /b/, /g/, /k/, /p/). Frikative entstehen durch Pressen des Luftstroms durch eine Engstelle im Vokaltrakt, welche an verschiedenen Stellen im Mund über eine bestimme Anordnung der Artikulatoren (z.B. Zunge, Zähne oder Lippen) erzeugt werden kann. Bei Plosiven

[26] Zum Beispiel wird der Buchstabe „r" im Deutschen in verschiedenen Dialekten unterschiedlich ausgesprochen, ohne eine andere Bedeutung zu bekommen.

hingegen wird der Vokaltrakt durch die Artikulatoren an einer Stelle ganz verschlossen und plötzlich wieder geöffnet [Wag].

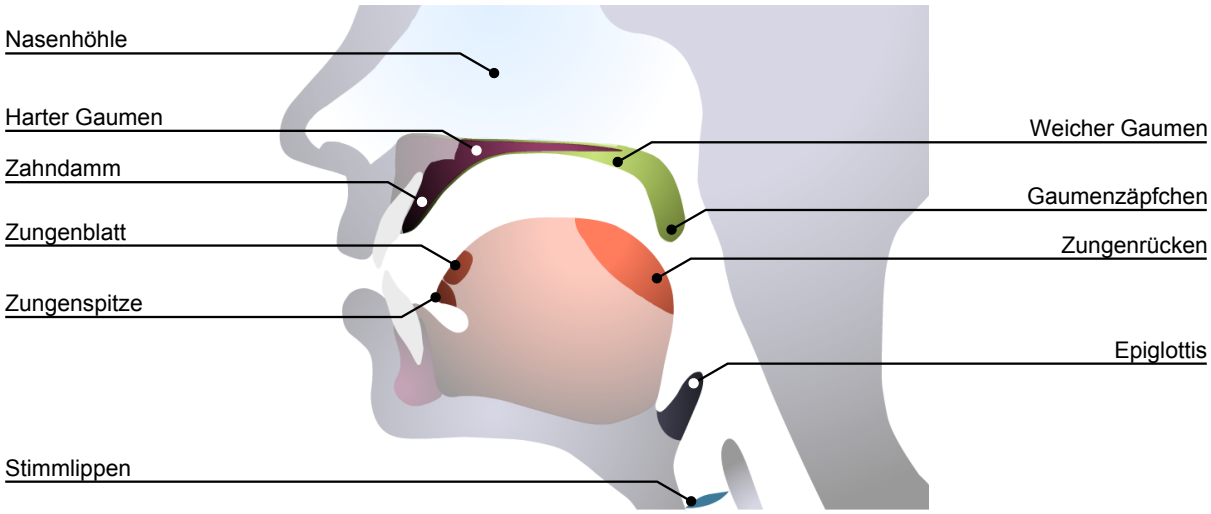

Abbildung 7: Schematischer Aufbau des Vokaltrakts. (Quelle: [Ph0])

3.2 Stimmlippen

Die Stimmlippen sind eine lippenförmige, schwingungsfähige Gewebestruktur, welche im Kehlkopf über der Luftröhre sitzt (siehe Abbildung 7). Sie bestehen im Inneren aus Muskelgewebe (*Stimmmuskel*, siehe auch Abbildung 8), welches die Stimmlippen anspannen kann, um sie zur Schwingung zu befähigen. Über dem Muskelgewebe befindet sich eine elastische Gewebeschicht (das eigentliche *Stimmband*), welche abschließend von einer Schleimhaut bedeckt ist. Jede der Schichten besitzt andere elastische Eigenschaften und kann sich bis zu einem gewissen Grad unabhängig von den Anderen bewegen [Ree05]. Die Stimmlippen laufen zur Wirbelsäule hin auseinander, wo die Enden mit Stellknorpeln verbunden sind. Diese werden über die Kehlkopfmuskulator gesteuert und erlauben die präzise Ausrichtung der Stimmlippen sowie den Verschluss des Spaltes zwischen ihnen (der *Glottis* bzw. auch *Stimmritze*).

Wenn die Stimmlippen zur Produktion eines Lautes eingesetzt werden (*Phonation*), geschieht dies in der Regel nach einem festen Muster. Zunächst wird die Glottis durch Anspannung der Stimmmuskel und Ausrichtung der Stellknorpel locker verschlossen und die Lunge baut einen subglottalen Überdruck auf. Übersteigt der Druck einen kritischen Punkt, werden die Stimmlippen auseinander gepresst, sodass Luft schnell entweicht.

Durch den subglottalen Druckverlust schließt sich die Glottis direkt danach wieder[27] und der Zyklus beginnt von Neuem [Krö07]. Physikalisch gesehen wird aus einem konstanten subglottalen Überdruck ein stoßartiger Luftstrom erzeugt, der aufgrund seiner periodischen Charakteristik in erster Näherung als Schwingung betrachtet werden kann. Der Luftdruckverlauf hinter den Stimmbändern repräsentiert direkt den nachfolgend vom Vokaltrakt modulierten Schall und ist schematisch in Abbildung 9 dargestellt [dVSVV02].

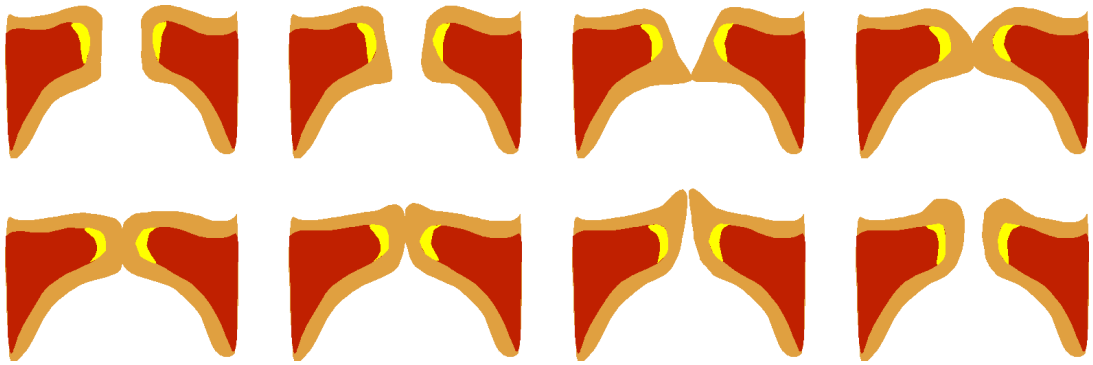

Abbildung 8: Bewegungsablauf der Stimmlippen bei Phontation. Dargestellt sind Stimmmuskel (Dunkelrot), das elastische Stimmband (Gelb) und die Schleimhaut (Orange). (Quelle: wikipedia.de)

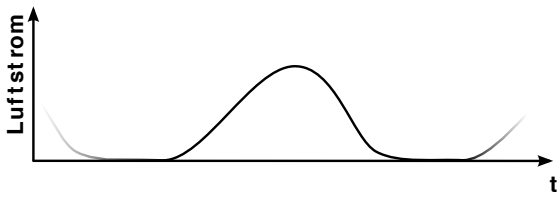

Abbildung 9: Schematische Darstellung der Stärke des Luftstroms direkt hinter den Stimmbändern. Angelehnt an [dVSVV02].

Es handelt sich beim Verlauf des glottalen Luftstroms nicht um eine monochromatische Schwingung, dennoch aber um eine periodische Signalform, welche sich als Superposition von Harmonischen der Grundperiode T_0 (Fourier-Reihe) darstellen lässt. Es muss sich also um ein Signal der Struktur $\sum c_n \exp(jn\omega_0 t)$ handeln, wobei $\omega_0 = 2\pi f_0 = 2\pi/T_0$ die Frequenz der Grundperiode, des zeitlichen Abstandes zwischen zwei Luftströßen, repräsentiert. Werden dem Verlauf entsprechende Phasen ϕ_n angenommen, lässt sich die

[27] Bernoulli-Effekt.

Struktur zu

$$\sum a_n \cos(n\omega_0 t + \phi_n) \tag{31}$$

vereinfachen. Der von den Stimmlippen produzierte Schall enthält also im Allgemeinen aus physikalischen Gründen höhere Harmonische der Grundfrequenz.

Das Verhältnis von Öffnungs- zu Verschlussdauer steht bei der Phonation in direktem Verhältnis zum Obertonreichtum des Lautes. Ein sehr kleines Verhältnis mit kleiner Öffnungsdauer entspricht einem zeitlich relativ scharfen Impuls, welcher inheränt ein reicheres Spektrum als bspw. eine Sinus-Halbwelle besitzt[28]. Im Mittel wird allerdings angenommen, dass die Harmonischen mit grob -12 dB/oct gegenüber der Grundfrequenz abfallen [DM08], was von eigenen Messungen bestätigt wurde.

Während eine interferenzfreie akustische Messung Stimmlippenschalls schwierig ist, erlaubt es ein Elektroglottogram (EGG) zumindest den Öffnungsgrad der Glottis zu bestimmen [Mic99]. Dazu werden zwei Elektroden an den Seiten des Kehlkopfes angebracht, durch die anschließend ein kleiner Strom in der Größenordnung 10mA mit einer Frequenz von 300 kHz bis einigen MHz geleitet wird. Während des Sprechens können dann Leitwerkschwankungen von 1-2 % gemessen werden, welche in direktem Zusammenhang mit dem Öffnungsgrad der Glottis stehen. Auf diese Weise kann die Aktivität der Stimmbänder entkoppelt von Effekten des Vokaltrakts aufgezeichnet werden.

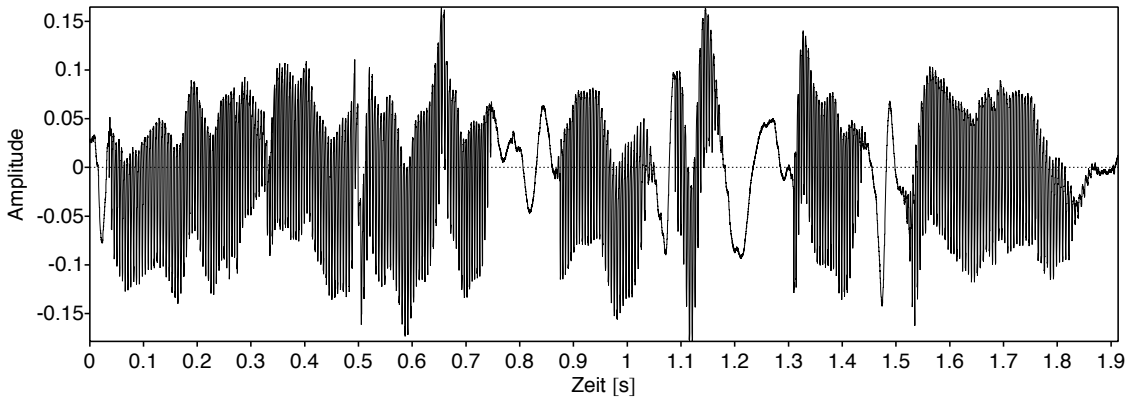

Abbildung 10: EGG-Aufnahme der Äußerung ,,Gad, your letter came just in time" einer weiblichen Stimme. Deutlich sichtbar sind die Phonationspausen.

[28]Bei Falsett-Gesang werden die Stimmlippen stark aneinandergepresst, sodass die Öffnungsdauer sehr kurz gegenüber der Verschlussdauer ist. Bei eigenen Aufnahmen mit einem Amateursänger konnten so 27 Harmonische basierend auf einer Grundfrequenz von ca. 810 Hz identifiziert werden. Höhere Frequenzen konnten die Geräte nicht aufgezeichnen.

In Abbildung 10 und 11 sind die EGG-Aufnahme und deren Spektrogramm einer weiblichen Stimme dargestellt. Deutlich erkennbar ist die kammartige harmonische Struktur und die Abnahme der Intensität der Harmonischen zu höheren Frequenzen hin.

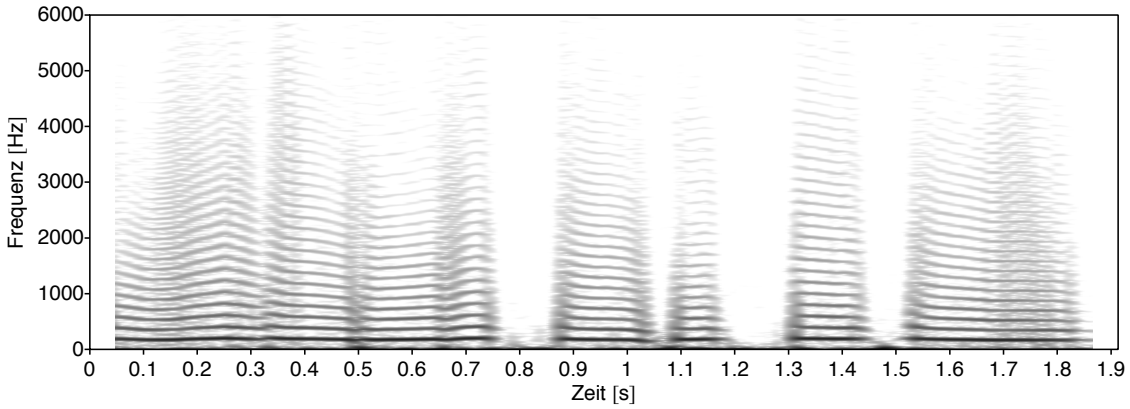

Abbildung 11: Spektrum der EGG-Aufnahme aus Abbildung 10 mit einer Fensterlänge von $T_f = 50$ms. Es sind alle Zeit-Frequenz-Punkte dargestellt, die sich weniger als 60 dB vom Maximalwert unterscheiden.

3.3 Vokaltrakt

Wie eingangs erwähnt, wird der von den Stimmlippen produzierte Schall im Vokaltrakt stark moduliert. Akustisch wirken Rachen, Mund und Nasenhöhle als Hohlraumresonatoren, die in Regel drei bis vier relevante Resonanzfrequenzen (*Formanten*) im Bereich von einigen Hundert oder wenigen Tausend Hertz ausbilden. Zusätzlich können durch die Nasenhöhle auch Dämpfungen auftreten, sofern der akustische Weg nicht durch den Weichen Gaumen und das Gaumenzäpfchen verschlossen ist [DM08] (siehe Abbildung 7).

Die genauen Frequenzen der Formanten hängen in komplexer Weise von der Geometrie des Vokaltraktes ab, welche über vielfältige Weise von den Artikulatoren (Zunge, Zähne, Lippen) verändert werden kann und sind überwiegend unabhängig voneinander. So lässt sich z.B. die Position des ersten Formanten durch die Mundöffnung steuern, während der zweite Formant von der Lage des Zungenrückens beeinflusst wird [Krö07]. Die Resonanzen des Vokaltraktes werden dem Schall der Stimmlippen aufgeprägt, d.h. Harmonische, die spektral in den Formanten liegen, werden entsprechend verstärkt und andere gedämpft. Dieser Effekt ist in Abbildung 12 dargestellt, welche das Spektrogramm der zu Abbildung 10 gehörenden akustischen Aufnahme zeigt. Während die Energie in der EGG-Aufnahme konstant über der Frequenz abnahm, sind hier starke Unterschiede über der Frequenz

und der Zeit sichtbar, welche durch Resonanzen und die vielfätigen Artikulationsarten entstehen.

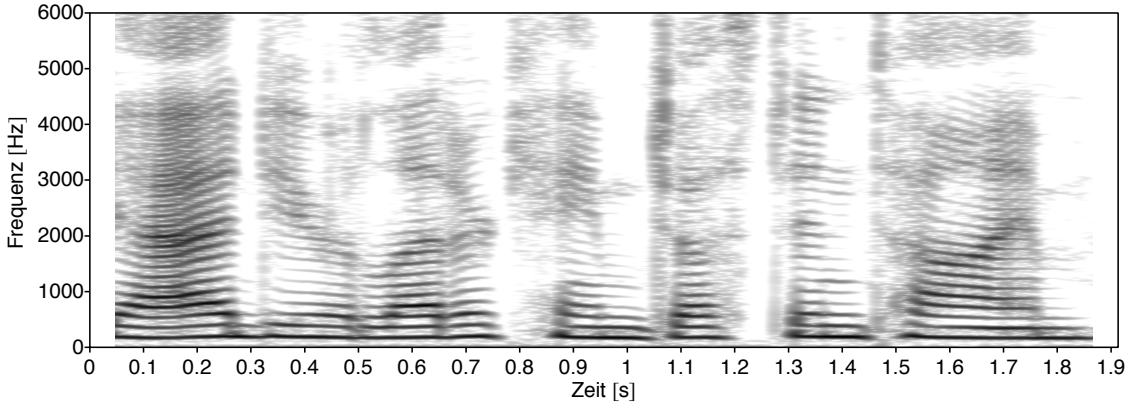

Abbildung 12: Spektrum der akustischen Aufnahme zur EGG-Aufnahme aus Abbildung 10 mit einer Fensterlänge von $T_f = 50$ms. Deutlich erkennbar sind die streifenartigen Verläufe der Harmonischen und die unterschiedlichen Ausprägungen verschiedener Frequenzbereiche über der Zeit.

In Abbildung 13 ist dieselbe Aufnahme dargestellt, allerdings wurde das Spektrum mit einer wesentlich kleineren Framelänge T_f berechnet, was nach (Gl. 4) zu einer geringen Frequenzauflösung führt. Die Harmonischen sind so nicht mehr aufzulösen, dafür werden aber die spektral wesentlich breiteren Formanten deutlicher sichtbar.

3.3.1 Modulation des Phonationsstroms

Neben der Verstärkung/Dämpfung von Harmonischen gibt es aber noch andere wesentliche Modulationseffekte, die aerodynamische Gründe haben. In [MKQ93] wird auf theoretische und experimentelle Hinweise auf starke Modulation des Luftstroms im Vokaltrakt hingewiesen, die Teager in seiner Arbeit [TT90] gefunden hat.

Laut Teagers Erkenntnissen ist der Luftstrom über den Querschnitt des Vokaltrakts hochgradig inhomogen und instabil. Er kann sich an Oberflächen konzentrieren, sich ablösen und andernorts wieder anheften sowie zwischen Oberflächen oszillieren. Dadurch ändern sich die effektive Querschnittsfläche und die Luftmassen, was zu einer Verschiebung der Resonanzfrequenz führt. Zusätzlich können sich um den Luftstrom auch Wirbel bilden, welche den Strom modulieren [MKQ93].

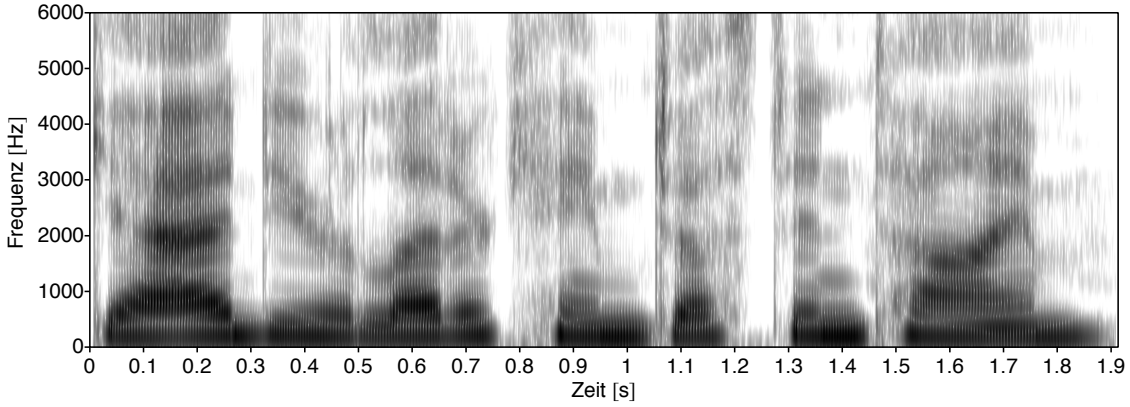

Abbildung 13: Spektrum der akustischen Aufnahme zur EGG-Aufnahme aus Abbildung 10 mit einer Fensterlänge von $T_f = 5$ms. Erkennbar sind 3 bis 4 Formanten als dunkle und spektral relativ breite Verläufe.

Klassisches Sprachmodell Bevor einige Messungen an Sprachsignalen erläutert werden, muss kurz auf das in [MKQ93] Bezug genommene, sehr verbreitete lineare Modell der Sprachproduktion eingegangen werden. Dort wird die Sprachproduktion ein eine Quelle (entsprechend den Stimmlippen) und einen Filter (entsprechend dem Vokaltrakt) aufgeteilt und getrennt modelliert. Die Quelle besteht aus einer einfachen Impulsfolge mit definiertem Intervall T_0. Im Spektrum entspricht dies ebenfalls einer Impulsfolge [GRS05] im Intervall $f_0 = T_0^{-1}$, welches der Grundfrequenz entspricht. Auf diese Weise werden sehr einfach die Harmonischen erzeugt, welche allerdings alle noch die gleiche Energie (im Sinne von (Gl. 3)) besitzen.

Die Wirkung des Vokaltrakts wird in diesem Modell auf einen einfachen All-Pole-IIR-Filter reduziert, welcher die selektive Verstärkung von Frequenzbändern durch die Positionierung von entsprechenden Polen in seinem Frequenzgang erlaubt. So wird die Wirkung der Formanten des Vokaltrakts nachgebildet. Weiterhin wird eine gewisse lokale Stationarität angenommen, d.h. dass der Vokaltrakt über Zeitschritte von 10 bis 30 ms als konstant betrachtet wird. Bei Anwendung des Filters auf die Impulsfolge wird dessen gleichförmiges Spektrum mit dem Frequenzgang des Filters multipliziert und somit ein synthetisches Sprachsignal generiert [LD95].

Praktische Verwendung findet dieses Model z.B. bei der Sprachcodierung als Basis für *Linear Predictive Coding* (LPC). Dort werden die Filterkoeffizienten per Linear Prediction aus dem Sprachsignal abgeschätzt und repräsentieren dann die aktuelle Formantenstruktur. Aus diesen Filterparametern und der Grundperiode wird anschließend versucht, mit

dem beschriebenen Verfahren das Ursprungssprachsignal zu rekonstruieren.

Jeder Pol[29] des Filters (bzw. jeder Formant) entspricht einem Oszillator 2. Ordnung mit einer exponentiell abklingenden Impulsantwort [GRS05]. Liegt der Pol auf einer Frequenz ω, entspricht die Impulsantwort einem exponentiell gedämpften Kosinus:

$$h(t) = Ae^{-\sigma t}\cos(\omega t + \theta) \tag{32}$$

Gemäß dem klassischen Modell wird der Filter jede Grundperiode T_0 durch einen Dirac-Puls angestoßen und erzeugt in der Folge die Superposition aller Impulsantworten aller Formanten, welche dann exponentiell abklingt, bis im Abstand von T_0 die Superposition der nächsten Impulsantworten am Ausgang erscheint. Zwischen zwei Impulsen liegt am Ausgang demnach eine exponentiell abklingende Schwingung bestehend aus den Formantfrequenzen.

Laut [MKQ93] hat Teager mit Hilfe des TEO Sprachaufnahmen auf den exponentiell abklingenden Verlauf zwischen zwei Pulsen[30] untersucht. Der TEO reagiert auf einen exponentiell abklingenden Kosinus mit [Kai90]:

$$x[n] = Ae^{-an}\cos(\Omega n) \tag{33}$$

$$\Psi[x[n]] = A^2e^{-2an}\sin^2(\Omega) \approx A^2\Omega^2e^{-2an} \tag{34}$$

Die Näherung setzt voraus, dass die Samplingfrequenz f_s mindestens eine Größenordnung über der des Kosinus liegt. Bei realen Sprachsignalen sollte die vom TEO berechnete Energie entsprechend (Gl. 34) ebenfalls exponentiell abfallen. Damit die Energieabschätzung verlässlich funktioniert, kann aber nur einer der Formanten betrachtet werden. Wie in Abschnitt 2 erwähnt, kann hier ein Gabor-Bandpass verwendet werden, um einen Formant herauszufiltern.

In den Untersuchungen in [MKQ93] finden sich zwischen zwei Impulsen allerdings nicht nur exponentiell abklingende Regionen sondern mitunter 2-3 weitere Energiepulse, deren Ursprung nicht genau geklärt ist. Möglicherweise sind sie praktisches Zeugnis der im Vorfeld beschriebenen Modulationen des Luftstroms. Die zeitliche Größenordnung der genannten Modulationsarten liegt unter der einer Grundperiode, weswegen sie als *microtime scale phenomenon* bezeichnet werden [MKQ93].

Bei eigenen Untersuchungen von Sprachaufnahmen konnten solche Energiepulse allerdings nur sporadisch und in geringerer Ausprägung gefunden werden. Möglicherweise

[29] Bzw. jedes konjugiert komplexe Paar Pole.

[30] Die Position von Quellimpulsen im Sinne des klassischen Modells lässt sich in der Praxis durch Zeitableitung einer EGG-Aufnahme gewinnen.

ist der Effekt stark sprecherabhängig oder ein Artefakt aus einer Interferenz. Abbildung 14 zeigt die Energieabschätzung der zweiten Formanten eines Ausschnitts der auch in Abbildung 12 verwendeten Sprachaufnahme. Zudem sind die aus dem Energieverlauf per DESA gewonnenen Abschätzungen für den Amplitudenumschlag und die Instantanfrequenz dargestellt.

Auffällig ist die im Takt der Grundperiode sehr stark oszillierende Energie, deren fallende Flanken bei genauerer Betrachtung Ähnlichkeit mit dem erwarteten exponentiellen Abklingen haben. Der DESA-Abschätzung zufolge ist der größte Teil der Oszillation der starken Amplitudenmodulation zuzuschreiben, da sich die Frequenz in Relation zur Amplitude wesentlich weniger ändert. Die beschriebenen Energiepulse zwischen den höheren Pulsen der Grundperiode können nur vereinzelt ausgemacht werden. In Verbindung mit den beobachteten Amplituden- und Frequenzmodulationen bewegt diese Beobachtung [MKQ93] dazu, das klassische lineare Modell einer lokal stationären Resonanz, die ein Formant modelliert, zu erweitern. Sie stellen die Impulsantwort einer Sprachresonanz als exponentiell gedämpftes AM-FM-Signal dar:

$$R[n] = a[n]\cos(\phi[n]) \tag{35}$$

$$= r^n A[n]\cos\left(\Omega_c n + \Omega_m \int_0^n q[k]dk + \phi[0]\right)\ ,\ |q[n]| \leq 1\ ,\ r \in (0,\,1) \tag{36}$$

Dabei ist Ω_c die Mittenfrequenz und $\Omega_i[n] = \Omega_c + \Omega_m q[n]$ ist analog zu (Gl. 25) die instantane Frequenz des modellierten Formanten. Das klassische Modell ist für den Fall konstanter Amplitude $A[n]$ und konstanter Instantanfrequenz $\Omega_i[n]$ enthalten.

Allerdings ist festzuhalten, dass nicht alle Frequenzmodulationsanteile eines Formanten von dessen Modulation stammen. Wie in [MKQ93] anhand eines synthetischen Signals gezeigt, verursacht ein Grundperiodenimpuls in der Instantanfrequenz einen Impuls, obwohl nur die Amplitude des Eingangssignals Modulation aufweist. Durch die Bandpassfilterung bekommt der Impuls im Frequenztrack eine gewisse zeitliche Breite.

Ergänzend sind in Abbildung 15 die Abschätzungen der vierten Formanten des gleichen Sprachsegments dargestellt. Es bestätigt sich die in [MKQ93] gemachte Beobachtung, dass die Modulation in Amplitude und Frequenz bei höheren Formanten stärker ausgeprägt ist.

Zusammenfassend betrachtet zeigt sich durch die Auswertung der AM- und FM-Anteile von Formanten, dass beim Phonationsprozess starke Modulationen auftreten, deren Ursache die betrachteten Quellen in den Modulationen des Luftstroms im Vokaltrakt sehen. Die zeitliche Größenordnung der beobachteten Modulationen liegt oft unter der einer Grundperiode. Damit wird auch klar, warum Frequenzbereichsmethoden wie z.B.

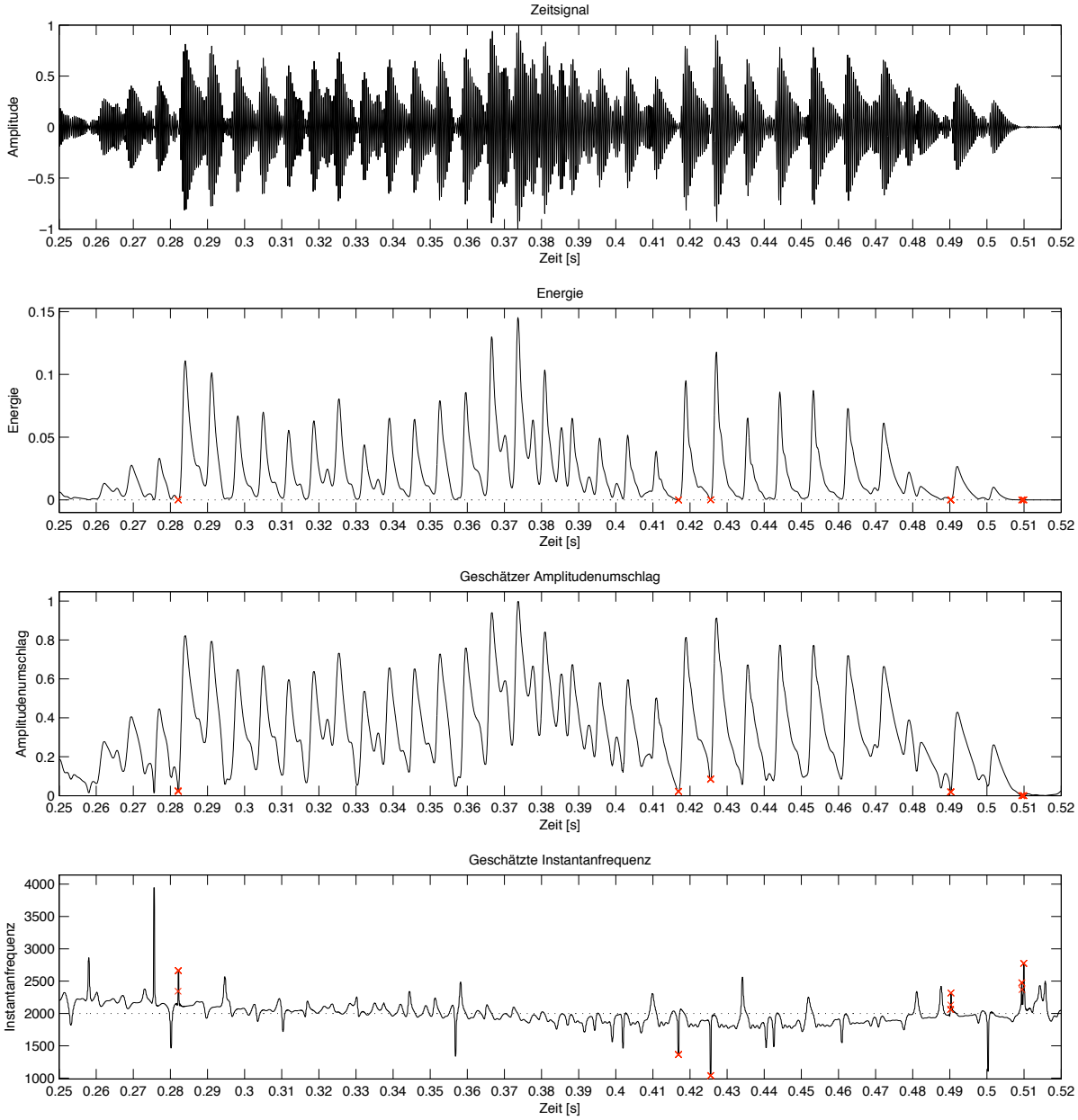

Abbildung 14: Abschätzungen von Energieverlauf (TEO), Amplitudenumschlag und instantaner Frequenz (ESA, beide Tracks per 11-Punkt Median-Filter geglättet) des zweiten Formanten im Ausschnitt „Gad" der Aufnahme aus Abbildung 12. Die Grundfrequenz sinkt im Laufe der Aufnahme von ca. 140 auf 110 Hz, was einer Grundperiode von ca. 7-9 ms entspricht. Die roten Kreuze markieren Fehler durch negative Werte des TEO. Parameter des Gabor-Filters: $f_c = 2000$ Hz, $\alpha = 1000$.

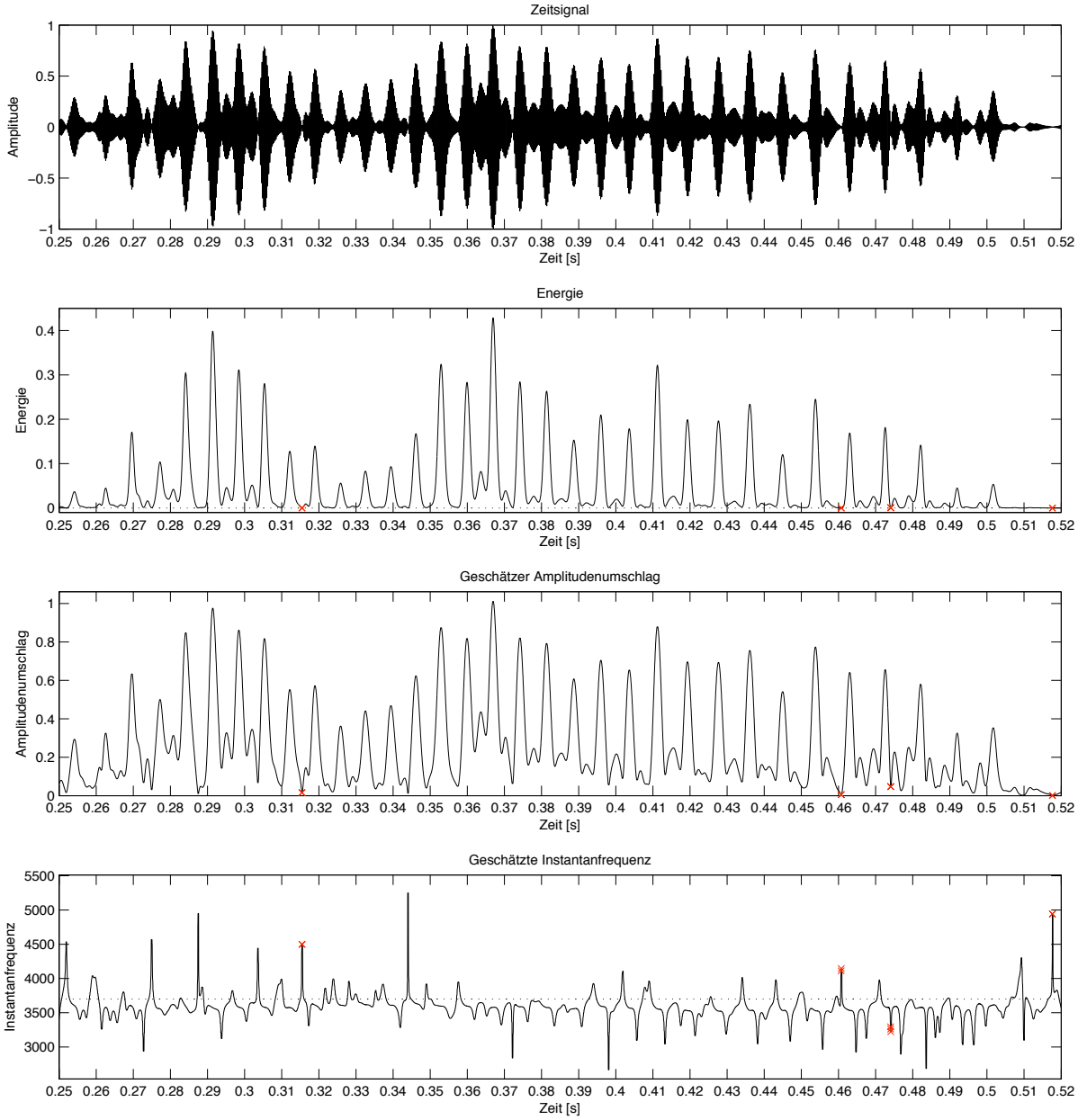

Abbildung 15: Abschätzungen von Energieverlauf (TEO), Amplitudenumschlag und instantaner Frequenz (ESA, beide Tracks per 11-Punkt Median-Filter geglättet) des vierten Formanten im Ausschnitt „Gad" der Aufnahme aus Abbildung 12. Die Grundfrequenz sinkt im Laufe der Aufnahme von ca. 140 auf 110 Hz, was einer Grundperiode von ca. 7-9 ms entspricht. Die roten Kreuze markieren Fehler durch negative Werte des TEO. Parameter des Gabor-Filters: $f_c = 3700$ Hz, $\alpha = 1000$.

die DFT in diesem Falle zu kurz greifen. Die in Abschnitt 2 diskutierten Beschränkungen der Frequenz- bzw. Zeitauflösung machen es sehr schwierig, derart feine und schnelle Modulationen zu registrieren. Werden nur die Magnituden eines Spektrogramms betrachtet, gehen die beobachten Mikromodulationen zwischen den Artefakten des Leck-Effektes unter.

Leider können im Rahmen dieser Arbeit nicht mehr die Auswirkungen der Formantenmodulation auf die Harmonischen untersucht werden. Es ist jedoch mindestens eine Amplitudenmodulation zu erwarten, da der Frequenzgang an der Position einer Harmonischen sowohl durch Frequenz- als auch durch Amplitudenmodulation der Formanten beeinflusst wird.

4 Rekonstruktion

In den vorangegangenen Abschnitten wurde auf die Entstehung von Sprachschall eingegangen und ein einfaches harmonisches Stimmmodell (Gl. 31) hergeleitet. Zusätzlich stehen der Grundfrequenztrack einer Sprachaufnahme und die daraus erstellten Energietracks zur Verfügung. Die Aufgabe besteht nun darin, aus diesen Informationen das ursprüngliche Sprachsignal möglichst verständlich zu rekonstruieren. Da die Ausgangsdaten zwar beschränkt sind aber direkt aus einem Sprachsignal stammen, wird die Problemstellung zunächst dem Gebiet der Sprachrekonstruktion und nicht der Sprachsynthese zugerechnet. Es wird als sinnvoll erachtet, dass eventuell synthetisierte Informationen erst in einem zweiten Schritt in das rekonstruierte Sprachsignal einfließen, wenn die Möglichkeiten der Sprachrekonstruktion ausgeschöpft sind und so die Qualität verbessert werden kann.

In diesem Abschnitt werden zwei verschiedene Ansätze zur Rekonstruktion des ursprünglichen Sprachsignals diskutiert. Zuerst werden die Möglichkeiten im Frequenzbereich ausgelotet und anschließend das hergeleitete Stimmmodell im Zeitbereich verwendet.

4.1 Spektrale Rekonstruktion

In Abschnitt 2 wurde bei der Berechnung der Energietracks gezeigt, dass sie sich sehr direkt aus dem Spektrogramm eines Sprachsignals ergeben. Daher ist es durch Invertieren von 3 möglich, die Beträge der harmonischen Bins[31] wiederherzustellen und so ein harmonisch ausgedünntes Spektrogramm zu erzeugen. Die Phase eines Bins kann allerdings nicht aus der Energie gewonnen werden, da sie in 3 verlorengeht. Nun wäre eine direkten Rücktransformation des Spektrogramms möglich, um das Zeitsignal zu gewinnen. Dabei müssen allerdings einige Punkte bedacht werden.

Falls das Spektrogramm aus zeitlich überlappenden Frames besteht, wird das nach der Rücktransformation wieder der Fall sein. Eine gebräuchliche Methode, überlappende Regionen zu verschmelzen beschreibt Overlapp-Add [Smi08]. Dabei muss allerdings auf die Verwendung einer geeigneten schiefsymmetrischen Fensterfunktion[32] korrekter Amplitude[33] geachtet werden, da das Verfahren sonst nicht mathematisch korrekt arbeitet und Artefakte erzeugt.

Durch das Fehlen vieler DFT-Koeffizienten liefern die entsprechenden Basisfunktio-

[31] Bei positiven und negative Frequenzen.

[32] Z.B. ein Dreieckfenster oder ein Tukey-Fenster mit beliebigem Parameter

[33] Die Fenster müssen sich an jedem Zeitpunkt zu 1 aufsummieren damit keine Amplitudenskalierung des Zeitsignals auftritt.

nen $\exp(j2\pi nk/N)$ keinen Beitrag zum Zeitsignal eines Frames. Dadurch treten bei Verwendung nicht-überlappender Frames Amplitudensprünge an den Grenzen benachbarter Frames auf. Akustisch sind sie als regelmäßiges Knacken mit der Periode der Framelänge T_f wahrnehmbar. Bei Verwendung überlappender Frames und Overlap-Add können solche Artefakte vermieden werden, da das Zeitsignal jedes Frame mit einer Fensterfunktion multipliziert wird, welche zu den Framegrenzen hin zu Null wird und jegliche Amplitudendifferenzen ein- bzw. ausblendet.

Schwerer als der Verlust nicht-harmonischer DFT-Koeffizienten wiegt der aber Verlust der Phasen der harmonischen Koeffizienten. Durch ihre insgesamt höhere Energie haben sie wesentlich größeren Einfluss auf den Verlauf des Zeitsignals als die inharmonischen Bins.

Für sich genommen hat die Phase eines DFT-Koeffizienten[34] $\angle(X[k])$ nur die Aussage einer Phasenverschiebung der Basisfunktion $s_k[n] = \exp(j2\pi nk/N)$ und gibt somit deren genaue Zeitposition wieder. Durch Superposition mit den skalierten und verschobenen Basisfunktionen anderer Bins können an den N Samplepunkten Werte jedes beliebigen kontinuierlichen Zeitsignals mit Frequenzen $f < f_s/2$ erzeugt werden. Die konkrete Form des diskreten Zeitsignals wird (abgesehen von den Bin-Beträgen) durch die Phasendifferenzen der Bins untereinander bestimmt. Eine Zeitverschiebung um n Samples korrespondiert zu einer linear mit der Frequenz wachsenden Phasenverschiebung, sodass der komplexe Koeffizient eines Bins $X[k]$ mit Frequenz f die Phasendifferenz $\Delta\phi = n \cdot 2\pi f/N$ erfährt. Da harmonische Frequenzen[35] in einem konstanten Intervall im Spektrum liegen, erfahren benachbarte Harmonische bei einer Zeitverschiebung die gleiche Änderung der Phasendifferenz zueinander.

Ohne Phaseninformationen ist es allerdings nicht möglich, an jedem Samplewert den korrekten Zeitsignalwert zu erzeugen, da die Basisfunktionen durch den Betrag eines Bins nur in ihrerer Amplitude skaliert werden können. Erschwerend kommt hinzu, dass zwar die Frequenz einer Harmonischen bekannt ist, aber durch das Spektrum nicht reflektiert wird. Wie in Abschnitt 3 besprochen, wird durch den Leckeffekt im Allgemeinen Energie einer Frequenz auf benachbarte Bins verteilt. Wenn die Frequenz f_h der Harmonischen im Zeitsignal etwas abseits der Frequenz ihres Bins liegt, wird das durch eine asymmetrische Energieverteilung in den benachbarten Bins deutlich, da das Spektrum der Fensterfunktion (siehe auch Abbildung 3) um f_h zentriert liegt. Da die Nachbarbins fehlen, wird das Zeitsignal des harmonisch ausgedünnten Spektrogramms ohne Phaseninformationen nur Frequenzen darstellen können, welche den Bin-Frequenzen $n \cdot df = n \cdot fs/N$ enspre-

[34] Hier wird wie üblich nur der Koeffizient der positiven Frequenz betrachtet.

[35] Genauso wie die Basisfrequenzen der Bins

chen. Bei einer Framelänge von $T_f = 30$ ms ergibt sich so eine spektrale Binbreite von $df \approx 33$ Hz. In dieser Form der Rekonstruktion wird sich die Stimme demnach sehr monoton anhören, solang die Grundfrequenz und die energiereicheren unteren Harmonischen (siehe Abbildung 2) im selben Bin bleiben (bzgl. der Frequenz). Sobald aber z.B. die Grundfrequenz in ein benachbartes Bin wechselt, wird das deutlich an der Änderung der Tonhöhe um df hörbar sein.

Da uns aber die exakte Frequenz f_h und ungefähre Energie einer Harmonischen bekannt ist, können die benachbarten Bins unter Annahme einer Fensterfunktion und des Spektrums einer Harmonischen abgeschätzt werden (siehe Erläuterung des Leck-Effektes, Abschnitt 2). Ohne weiteres Vorwissen können wir eine Harmonische entsprechend dem einfachen Stimmmodell in (Gl. 31) als Dirac-Puls im Frequenzbereich modellieren. Die Beträge und Phase der Nachbarbins können dann dem Frequenz- bzw. Phasengang der verwendeten Fensterfunktion an den entsprechenden Positionen entnommen werden. Mit diesem Ansatz wird im Optimalfall also eine Sinusschwingung der Frequenz f_h im Zeitbereich erzeugt. Der Aufwand ist allerdings vergleichsweise hoch. Im nachfolgenden Abschnitt werden wir eine wesentlich direktere Methode zum Erreichen desselben Ziels diskutieren, weshalb dieser Weg nicht weiter verfolgt wird.

Wie bereits erwähnt sind für den Verlauf des Zeitsignals eines Frames die Phasendifferenzen zwischen den Zeit-Frequenz-Bins des Frames verantwortlich. Ähnliches gilt bei überwiegend periodischen Signalen für die Phasendifferenzen von zeitlich benachbarten Bins derselben Frequenz.

Angenommen ein harmonisches Zeitsignal mit konstanter Grundfrequenz wird in Frames zerlegt (Überlappung ist auch möglich). Dann enthalten alle Frames phasenverschobene Kopien desselben Signals. Daher ändern sich die Phasendifferenzen zwischen den spektral benachbarten harmonischen[36] Bins von zeitlich benachbarten Frames um gerade den Phasenwert, der abhängig von der Frequenz nötig ist, um eine Zeitverschiebung von $\Delta t = \frac{K}{N} T_f$ zu beschreiben. Die Phasendifferenzen zeitlich benachbarter harmonischer Bins derselben Frequenz sind somit konstant.

Ändert sich die Grundfrequenz, führt das zu einer zeitlichen Stauchung/Streckung, wodurch sich die Position der Grundperioden mit zunehmender Zeit relativ zum starren Frameintervall $\frac{K}{N} T_f$ immer mehr verschiebt. Bei einer Zerlegung in Frames ändern sich dadurch effektiv die Phasendifferenzen zwischen benachbarten Frames linear mit der Änderung der Grundfrequenz. Die Phasendifferenzen zeitlich benachbarter harmoni-

[36] Exakte Vielfache von f_0 sind eigentlich erforderlich. Im diskreten Fall ist das nur näherungsweise gegeben, da die Abstände harmonischer Bins leicht variieren können.

scher Bins derselben Frequenz sind damit nicht mehr konstant sondern stehen in direkter Beziehung zur Grundfrequenz. Das bedeutet im Umkehrschluss, dass mit der Kenntnis der Grundfrequenz die Phasendifferenzen zeitlich benachbarter Bins berechnet werden können.

Unbekannt bleiben dann noch die absoluten Phase der zeitlich ersten harmonischen Bins, welche zeitliche Position der entsprechenden Basisfrequenzen im Zeitsignal steuern. Die Literatur [WL82] geht davon aus, dass relative Phasenunterschiede von Harmonischen in der Regel nicht wahrgenommen werden können, einige Spezialfälle von maximaler Superposition oder Teilauslöschung der beteiligten Harmonischen ausgenommen. Um diesen Fällen aus dem Weg zu gehen, werden die Initialphasen zufällig gewählt. Die Phasen der Harmonischen in nachfolgenden Frames werden über der Zeit mit der beschriebenen Methoden aus dem Grundfrequenztrack ermittelt.

Bei der eigenen Implementierung wurde allerdings nicht alle beschrieben Möglichkeiten verwirklicht. Durch diese Art der Rekonstruktion wird versucht, im Frequenzbereich Annahmen umzusetzen, die im Zeitbereich einfacher zu handhaben sind, wie z.B. das einfache Modell sinusförmige Modell einer Harmonischen oder die Definition von Initialphasen. Selbst Optimalfall werden durch diesen Ansatz nur sinusförmige amplituden- und frequenzmodulierte Harmonische erzeugt. Im nächsten Abschnitt wird effektiv dasselbe angestrebt und mit einem Modell im Zeitbereich umgesetzt.

4.2 Zeitbereichsrekonstruktion

In Abschnitt 3 wurde aus physikalischen Argumenten ein einfaches Stimmmodell hergeleitet, welches sich in ähnlicher Form auch in Bereich der Sprachkodierung wiederfindet [MQ86]. Dort werden allerdings nicht die unbedingt die Harmonischen modelliert, sondern nur die wesentlichen spektralen Peaks im Zeitbereich rekonstruiert. Um eine möglichst vollständige Rekonstruktion zu erlauben, werden hier sämtliche Harmonische, deren Energietrack bekannt ist, angesetzt. Ein Vorteil des Modells ist zudem die Trennbarkeit des Modells in Quelle und Filter, um bei Bedarf beide getrennt behandeln zu können. In diesem Kontext erscheint dies aber nicht direkt von Wert, kann aber möglicherweise bei der Weiterentwicklung des dargestellten Verfahrens hilfreich sein.

Das angesetzte Modell beschreibt ein Sprachsignal als Superposition von Harmonischen

$$s(t) = \sum_{h=1}^{H} A_h(t) \cos(\phi_h(t)) \ , \tag{37}$$

wobei

$$A_h(t) = a_h(t)M(\omega_h(t);t) \tag{38}$$

$$\phi_h(t) = \Phi(\omega_h(t),t) + \int_0^t \omega_h(\tau)d\tau + \theta_h \tag{39}$$

den Amplituden- bzw. Phasenverlauf der h-ten Harmonischen über der Zeit beschreiben. Die Übertragungsfunktion $H(\omega;t)$ des Vokaltraktes ist dabei in Frequenz- und Phasengang aufgespalten:

$$H(\omega;t) = M(\omega;t)\exp(j\Phi(\omega;t)) \tag{40}$$

Hier soll aber zunächst nicht zwischen Quelle und Filter unterschieden werden. Zudem basiert die Berechnung der Harmonischen auf dem Grundfrequenztrack, der an jedem Samplepunkt ausgelesen wird[37]. Damit lässt sich das Modell vereinfachen und wird fortan als *Harmonische Rekonstruktion* bezeichnet:

$$s(t) = \sum_{h=1}^{H} a_h(t)\cos(\phi_h(t)) \ , \tag{41}$$

wobei

$$\phi_h(t) = \int_0^t h\omega_0(\tau)d\tau + \theta_h \tag{42}$$

$$H = \left\lfloor \frac{f_s}{2\max(f_0)} \right\rfloor \tag{43}$$

Der Amplitudenverlauf $a_h(t)$ errechnet sich nach (Gl. 5) direkt aus dem Energietrack der h-ten Harmonischen[38]. Aus den gleichen Gründen wie im vorigen Abschnitt werden hier zudem zufällige Initialphasen θ_h verwendet. Damit die Rekonstruktion nicht stumpf klingt, ist ein ausreichender Anteil hoher Frequenzen[39] erforderlich. Da in diesem Fall keine Interferenzen durch andere Schallquellen zu erwarten sind, wurde der maximal möglich Wert gewählt, der bei durchgehender Nutzung aller Harmonischen möglich ist.

Direkt ersichtlich sind einige wesentliche Vorteile gegenüber der Rekonstruktion eines Sprachsignals aus einem harmonisch ausgedünnten Spektrogramm:

- Das Modell arbeitet kontinuierlich und erfordert zur Rekonstruktion keine Berücksichtigung von Frames bzw. von Methoden, diese zu verschmelzen.

[37] Dazu wurde der f_0-Track auf die Zeitauflösung der Samplingfrequenz hoch interpoliert.

[38] welcher ebenfalls auf die Zeitauflösung der Samplingrate hochinterpoliert wurde.

[39] durchaus über 10 kHz

- Jede Harmonische ist individuell dargestellt und unabhängig in Amplitude und Winkel modulierbar. Diese Eigenschaft stellt besonders im Hinblick auf die Erweiterung des Modells um evtl. synthetisierte Informationen einen großen Vorteil dar.

- Bei hinreichend glatten Amplitudenumschlägen $a_h(t)$ und Phasenverläufen $\phi_h(t)$ erzeugt es ein phasenkontinuierliches Signal.

- Bei hinreichender Anzahl Harmonischer lassen sich unter Annahme einer Dummy-Grundfrequenz auch stimmlose Laute (z.B. Obstruenten) modellieren, da dem Spektrum sehr gleichmäßig Energie entnommen wird. In der Tat wurde diese Methode gewählt, um Segmente ohne Grundfrequenztrack zu rekonstruieren und somit die Rekonstruktion einer kompletten Äußerung zu erreichen. In Gegenwart von Interferenzen ist die natürlich nicht zulässig, da die spektrale Orthogonalitätsbedingung ohne Grundfrequenz pauschal verletzt ist.

Ein so rekonstruiertes Sprachsignal hört sich angesichts des einfachen Modells überraschend verständlich und natürlich an. Bei ausreichend vielen Harmonischen ist die Rekonstruktion dem Original sehr ähnlich. Die Verständlichkeit ist nahe am Original, es sind nur geringe Abstriche zu machen. In Abbildung 16 und 17 sind die Spektrogramme zweier rekonstruierter Sprachsegmente dargestellt. Die spektrale Energieverteilung ist größtenteils gut reproduziert worden. Bei der Rekonstruktion weist aber eine etwas ausgeprägtere harmonische Struktur auf, was aufgrund der harmonisch ausgedünnten Energien aber zu erwarten ist.

An dieser Stelle wurde auch empirisch ein grober Wert für die Framelänge ermittelt. Es zeigte sich, dass ab einer Framelänge von 30 ms (mit einer Frameüberlappung von 87,5%) Plosive ihre zeitliche Schärfe hörbar verloren und zu verlaufen begannen.

In den Rekonstruktionen treten allerdings besonders bei männlichen Stimmen markante Artefakte auf, wodurch sich die Rekonstruktion etwas blechern aber dennoch harmonisch anhört. Ein zusätzliches Glätten der Energietracks verstärkt diesen Effekt. Dies wird als Indiz gewertet, dass zumindest im Energietrack (und damit in den Amplitude) Modulationen gegenüber dem Original verlorengegangen sind. Die Ursache dafür wird in der Zeitintegration der DFT bei der Energiebestimmung gesehen.

Die Vermutung, fehlende Modulation ist für die produzierten Artefakte verantwortlich, motivierte letztlich die Verwendung von ESA, um die AM- und FM-Anteile in einem Sprachsignal genauer zu studieren. In Abschnitt 3 wurde deshalb DESA auf die Formante eines Sprachsignals angewandt, um Rückschlüsse auf die Modulation der Harmonischen ziehen zu können.

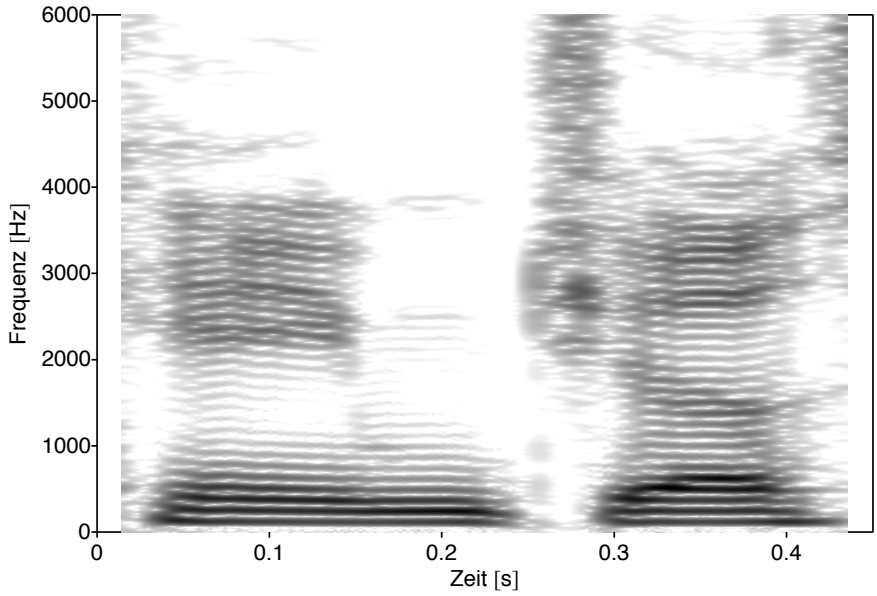

Abbildung 16: Spektrogramm einer 450 ms langen Aufnahme einer männlichen Stimme. Dargestellt sind alle Frequenzen bis zu 50dB unterhalb der Maximalenergie. Diesem Segment liegen Abb. 5 und 2 zugrunde. Es wurde ein Tukey-Fenster mit $r = 0.8$ verwendet.

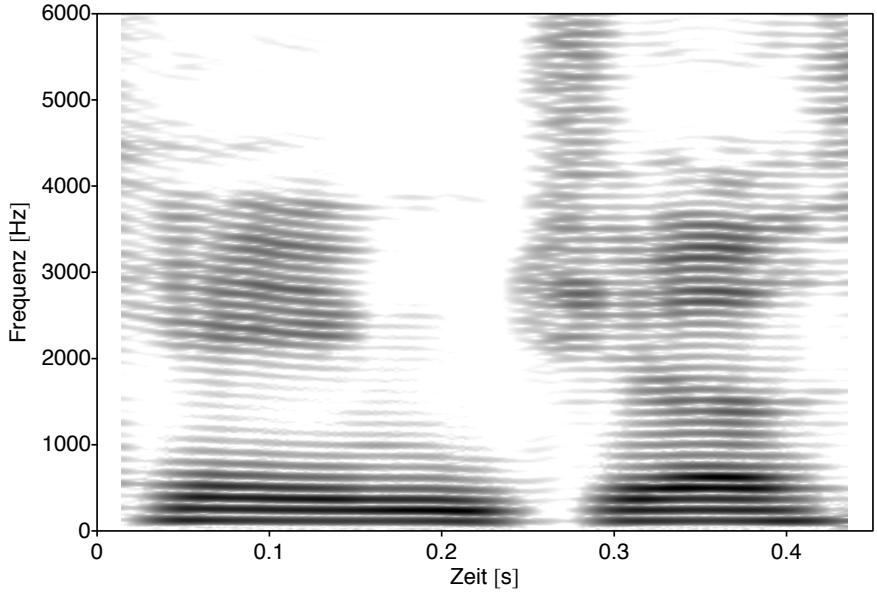

Abbildung 17: Spektrogramm der harmonischen Rekonstruktion einer 450 ms langen Aufnahme einer männlichen Stimme. Dargestellt sind alle Frequenzen bis zu 50dB unterhalb der Maximalenergie. Dieses Segment ist auch in Abb. 5 und 2 dargestellt. Es wurde ein Tukey-Fenster mit $r = 0.8$ verwendet.

Hier soll zunächst DESA auf die harmonische Rekonstruktion desselben Sprachsegments angewandt werden, um die Abschätzungen von Energie (im Sinne des TEO), Amplitudenumschlag und instantaner Frequenz eines Signals zu gewinnen, dessen Parameter vollständig bekannt sind. Es soll sozusagen die Koheränz der abgeschätzten Tracks zu den bekannten Daten überprüft werden. Dazu sind in Abbildung 18 und 19 die DESA-Ergebnisse des zweiten bzw. vierten Formanten einer Rekonstruktion dargestellt. Das verwendete Sprachsegment ist dasselbe, wie in Abbildung 14. Da am Ende jedoch keine Grundfrequenz mehr gemessen werden kann, ist es etwas kürzer als in der originalen Sprachaufnahme.

Die Energietracks, aus denen das Segment rekonstruiert wurde, sind in Abbildung 2 dargestellt und zeigen relativ glatte Verläufe. Auch der Grundfrequenztrack ist sehr glatt. Im Resultat sollte das rekonstruierte Sprachsegment verglichen mit dem Originalsignal wesentlich geringere Modulation aufweisen. Dazu scheinen aber die stark und mitunter sehr unregelmäßig oszillierenden AM- und FM-Tracks der zweiten und vierten Formanten im Widerspruch zu stehen. Besonders der Energietrack in Abbildung 18 zeigt viele Pulse, die im Gegensatz zu Abbildung 14 nicht nur im Takt der Grundperiode auftreten. Allerdings ist festzuhalten, dass die Rekonstruktionsqualität hochgradig von einer korrekten Bestimmung der Energietracks abhängt, welche wiederrum auf eine präzise Grundfrequenzabschätzung angewiesen ist.

Im Rahmen dieser Arbeit ist es aber leider nicht mehr möglich, auf erwähnten scheinbaren Widersprüche zwischen Signalparametern und DESA-Abschätzungen einzugehen.

4.3 Zusammenfassung und Ausblick

Anhand der geführten Diskussion hat sich eine Rekonstruktion im Zeitbereich gegenüber der im Frequenzbereich in vielen Belangen als flexibler und einfacher zu handhaben herausgestellt. Die erreichte Verständlichkeit der Rekonstruktion ist gut, es sind nur geringe Abstriche gegenüber dem Original zu machen. Die Natürlichkeit der Rekonstruktion wird allerdings durch charakteristische Artefakte beeinträchtigt. Die Ursache dafür wird trotz möglicherweise fehlleitender ESA-Abschätzungen nach wie vor in einem Mangel an Modulation in den Amplitudenverläufen und evtl. den Frequenzverläufen der harmonischen Rekonstruktion vermutet.

Künftige Untersuchungen können sich beispielsweise mit dem weiteren Studium der ESA-Methode beschäftigen. Anhand synthetischer Formanten könnten z.B. die Abhängigkeit von Abschätzungen durch Interferenzen von individuellen Harmonischen oder benachbarten Formanten näher beleuchtet werden. Ein weiterer Ansatz besteht in der Un-

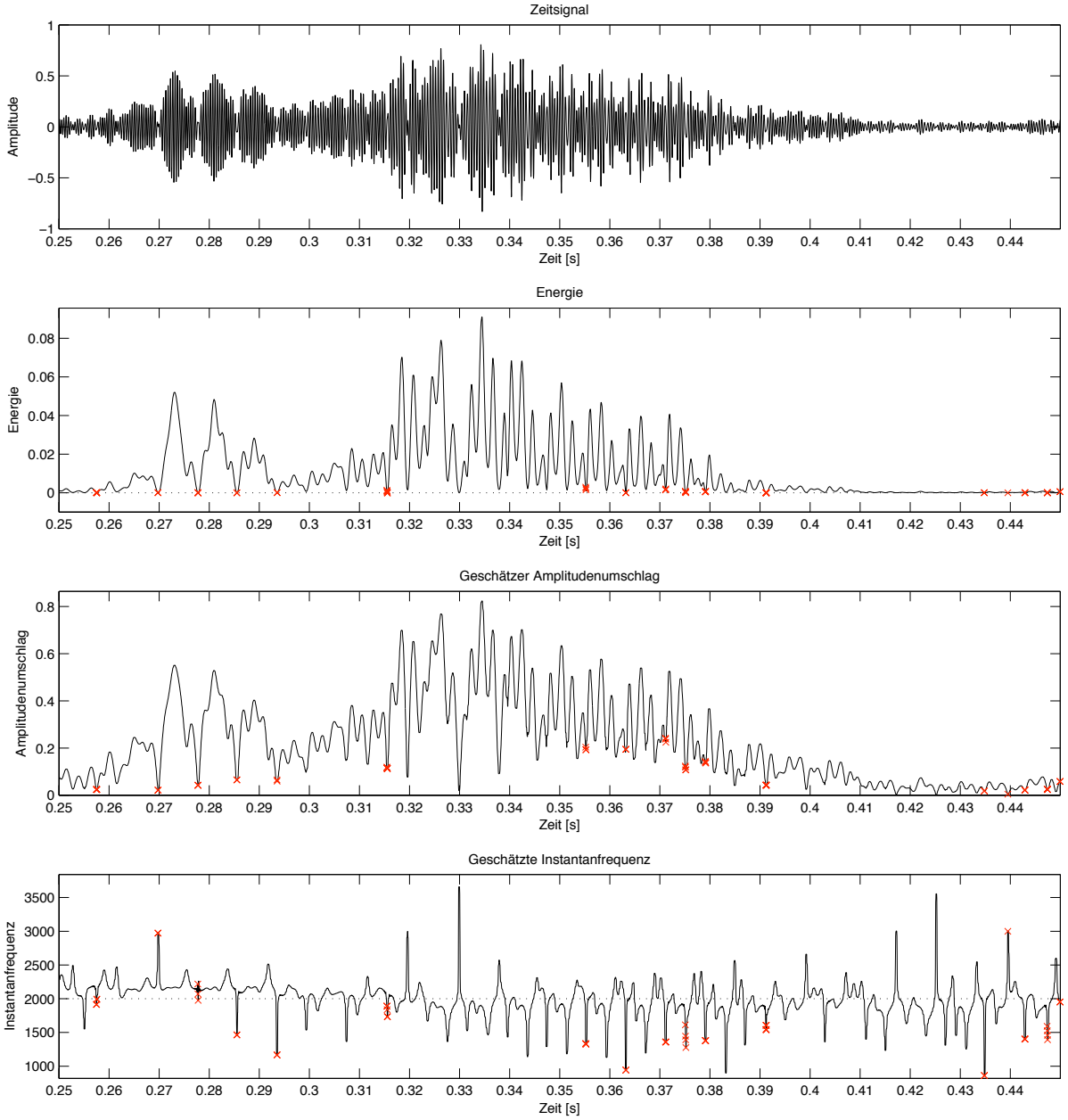

Abbildung 18: Abschätzungen von Energieverlauf (TEO), Amplitudenumschlag und instantaner Frequenz (ESA, beide Tracks per 11-Punkt Median-Filter geglättet) des zweiten Formanten der harmonischen Rekonstruktion des in Abbildung 14 benutzten Sprachsegments. Die roten Kreuze markieren Fehler durch negative Werte des TEO. Parameter des Gabor-Filters: $f_c = 2000$ Hz, $\alpha = 1000$.

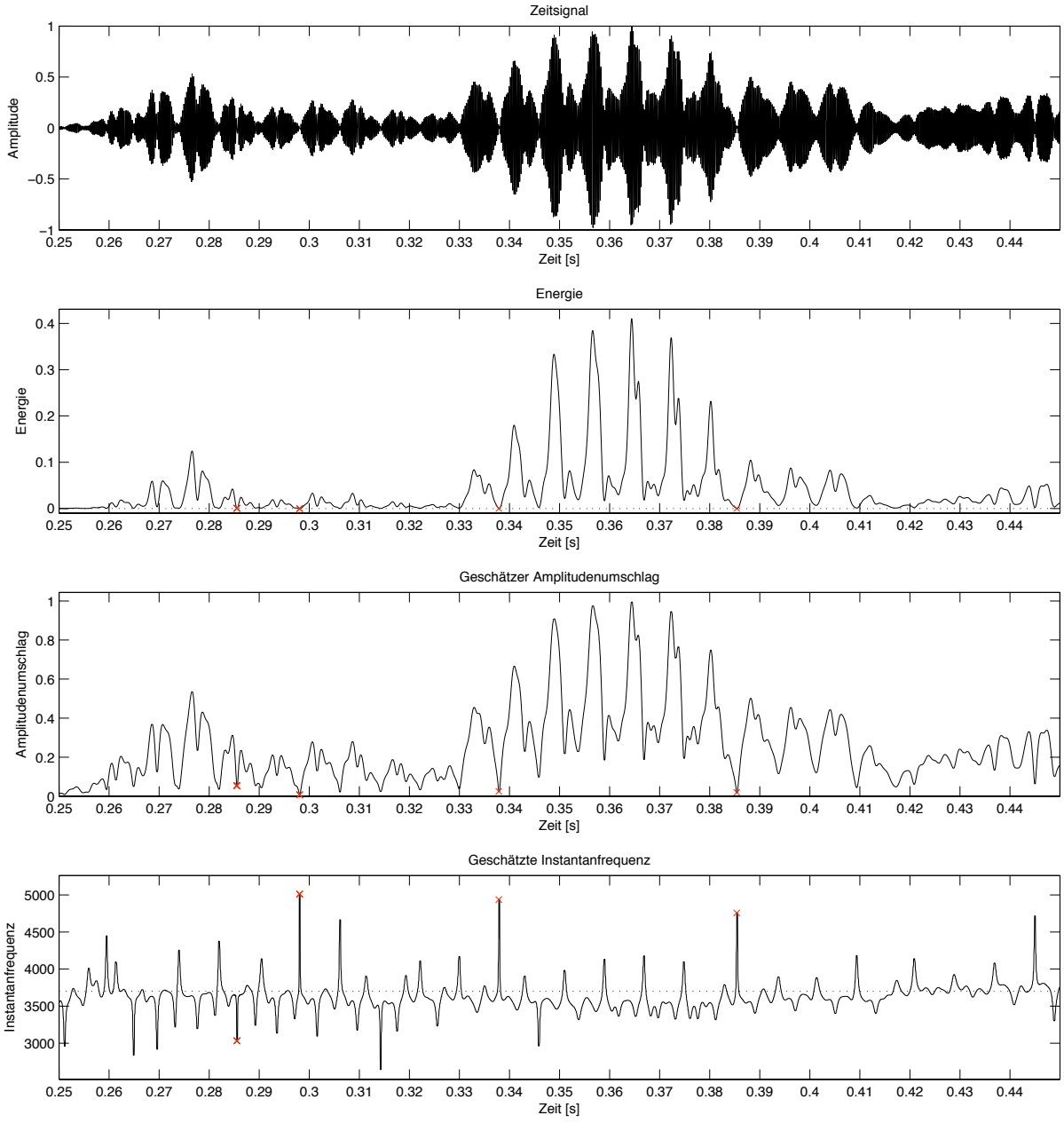

Abbildung 19: Abschätzungen von Energieverlauf (TEO), Amplitudenumschlag und instantaner Frequenz (ESA, beide Tracks per 11-Punkt Median-Filter geglättet) des vierten Formanten der harmonischen Rekonstruktion des in Abbildung 15 benutzten Sprachsegments. Die roten Kreuze markieren Fehler durch negative Werte des TEO. Parameter des Gabor-Filters: $f_c = 3700$ Hz, $\alpha = 1000$.

tersuchung der Modulation individueller Harmonischer durch eine amplituden- und frequenzmodulierte Formante (Gl. 36).

Literatur

[AO99] Alan Oppenheim, Roland Schafer, John Buck: *Discrete time signal processing*. Prentice Hall, 2nd Auflage, 1999.

[Aro92] Arons, Barry: *A Review of The Cocktail Party Effect*. Journal of the American Voice I/O Society, 12:35–50, 1992.

[BM94] Bovik, Alan C. und Petros Maragos: *Conditions for Positivity of an Energy Operator*. IEEE transactions on signal processing, 42(2):469–471, February 1994.

[DM01] Dimitriadis, Dimitrios und Petros Maragos: *An improved energy demodulation algorithm using splines*. In: *ICAASSP-2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 6, 2001.

[DM08] Deutsch, Werner A. und Sylvia Moosmüller. *Akustisches Modell der Sprachproduktion* [online]. 2008. URL: `http://www.kfs.oeaw.ac.at/content/view/329/482/` [Zugriff am 29.01.2008].

[dVSVV02] Vries, M.P. de, H.K. Schutte, A.E.P. Veldman und G.J. Verkerke: *Glottal flow through a two-mass model: Comparison of Navier-Stokes solutions with simplified models*. Journal of the Acoustical Society of America, 111(4):1847–1853, 2002.

[GRS05] Girod, Bernd, Rudolf Rabenstein und Alexander Stenger: *Einführung in die Systemtheorie*. Teubner, 3. Auflage Auflage, 2005.

[Kai90] Kaiser, James F.: *On a simple algorithm to calculate the 'energy' of a signal*. In: *ICASSP-90, International Conference on Acoustics, Speech, and Signal Processing*, 1990.

[Kai93] Kaiser, James F.: *Some useful properties of Teager's energy operators*. In: *ICASSP-93, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993.

[KB03] Kominek, John und Alan W. Black: *CMU ARCTIC: Databases For Speech Synthesis*, 2003. URL: `http://www.festvox.org/cmu_arctic/cmu_arctic_report.pdf`.

[Krä08] Krämer, Jochen: *Multiple Fundamental Frequency Estimation for Cognitive Source Separation.* Bachelorarbeit, Universität des Saarlandes, 2008.

[Krö07] Kröger, Bernd J. *Artikulatorische und akustische Phonetik - Ein Kurzüberblick* [online]. 2007. URL: `http://www.logopaedie.rwth-aachen.de/personen/dozenten/bkroeger/documents/Kroeger_PhonetikSkript_2007.pdf` [Zugriff am 29.01.2008].

[Kve03] Kvedalen, Eivind: *Signal processing using the Teager Energy Operator and other nonlinear operators.* Candidatus Scientiarum Abschlussarbeit, University of Oslo, 2003.

[LD95] Lu, Shan und Peter C. Doerschuck: *Nonlinear Modeling and Processing of Speech with Applications to Speech Coding.* Technical Report, School of Electrical and Computer Engineering, Purdue University, 1995.

[Mic99] Michaelis, Dirk: *Das Göttinger Heiserkeits-Diagramm - Entwicklung und Prüfung eines akustischen Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen.* Doktorarbeit, Georg-August-Universität zu Göttingen, 1999.

[MKQ93] Maragos, Petros, James F. Kaiser und Thomas F. Quatieri: *Energy Separation in Signal Modulations with Application to Speech Analysis.* IEEE Transactions on Signal Processing, 41(10):3024–3051, October 1993.

[MQ86] McAulay, Robert J. und Thomas F. Quatieri: *Speech Analysis/Synthesis Based on a Sinusoidal Representation.* IEEE Transactions on Acoustics, Speech and Signal Processing, 34(4):744–754, August 1986.

[Ph0] [online]URL: `http://www.uiowa.edu/~acadtech/phonetics/` [Zugriff am 29.01.2008].

[PM94] Potamianos, Alexandros und Petros Maragos: *A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation.* Signal Processing, 37(1):95–120, 1994.

[Ree05] Reeve, Matthew. *The Bernoulli Effect and Vocal Fold Vibration* [online]. 2005. URL: `http://www.voicesource.co.uk/article/151` [Zugriff am 29.01.2008].

[Smi08] Smith, Julius O. *Spectral Audio Signal Processing, October 2008 Draft* [online]. 2008. URL: http://ccrma.stanford.edu/~jos/sasp/ [Zugriff am 21.01.2009].

[TM04] Tipler, Paul A. und Gene Mosca: *Physik – Für Wissenschaftler und Ingenieure*. Elsevier, 2004.

[TT90] Teager, H. M. und S. M. Teager: *Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract*. Proceedings of the NATO Advanced Study Institute on Speech Production and Speech Modelling, Bonas, France, July 17-29, 1989, 1990. URL: http://books.google.com/books?id=oIs5xt9KvJwC.

[Wag] Wagner, Karl Heinz. *Phonetik und Phonologie* [online]. URL: http://www.fb10.uni-bremen.de/khwagner/phonetik/phonologie.htm [Zugriff am 03.02.2009].

[WL82] Wang, David L. und Jae S. Lim: *The Unimportance of Phase in Speech Enhancement*. IEEE Transactions on Acoustics, Speech and Signal Processing, 30(4):679–681, 1982.

[YR04] Yilmaz, Özgür und Scott Rickard: *Blind Separation of Speech Mixtures via Time-Frequency Masking*. IEEE Transactions on Signal Processing, 52(7):1830–1847, 2004.